POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Action Density based Frame Sampling for Human Action Recognition in Videos

TESI DI LAUREA MAGISTRALE IN

COMPUTER SCIENCE AND ENGINEERING

INGEGNERIA INFORMATICA

Author: **Zekun Mu**

Student ID: 10743739
Advisor: Marco Marcon
Academic Year: 2021-2022

Abstract

Action recognition has been widely used to identify and monitor special activities in videos, and a proper frame sampling method can not only reduce redundant video information, but also improve the accuracy of action recognition. In this paper, action density based frame sampling methods are proposed to discard the redundant video information and select the rational frames for neural networks to achieve high accuracy on human action recognition in videos. In particular, action density is introduced in our methods to indicate the intensity of actions in videos, and a reinforcement learning based frame selection mechanism with considering the action density as the reward is proposed to select frames with the best action features. Then, a segmented frame sampling (SFS) method is proposed for multi-channel neural network and a non-isometric frame sampling (NFS) method is proposed for single-channel neural network, respectively, to simultaneously select a series of the rational frames (i.e., achieve the frame sampling in videos) based on the RLFD mechanism for action recognition. Via the evaluations with various neural networks and datasets, our results not only show the effectiveness of using action density as a metric in frame selection, but also show that the proposed SFS and NFS method can achieve great effectiveness and rationality in frame sampling and can assist in achieving better accuracy of action recognition, in comparison with existing methods.

Key-words: action recognition, frame sampling, reinforcement learning

Abstract in lingua italiana

Il riconoscimento di azioni è stato ampiamente utilizzato per identificare e monitorare attività specifiche nei video, e un metodo appropriato di campionamento fotogrammi può, non solo ridurre le informazioni video ridondanti, ma anche migliorare l'accuratezza del riconoscimento delle azioni. In questo documento, si propongono metodi di campionamento dei fotogrammi basati sulla densità di azione per scartare le informazioni video ridondanti e selezionare i fotogrammi utili per le reti neurali al fine di ottenere un elevata precisione sul riconoscimento dell'azione eseguita nei video. In particolare, la densità di azione è introdotta nei nostri metodi per indicare l'intensità delle azioni nei video, e un meccanismo di selezione dei fotogrammi basato sull'apprendimento rafforzato, considerando la densità di azione come parametro premiante nella selezione dei fotogrammi con le migliori caratteristiche di azione. Viene poi proposto un metodo di campionamento per fotogrammi segmentati (SFS) per una rete neurale multicanale e un metodo di campionamento per fotogramma non isometrico (NFS) per la rete neurale a un canale singolo. Ciò consente di selezionare simultaneamente una serie di fotogrammi affini basato sul meccanismo RFD per il riconoscimento delle azioni. Attraverso le valutazioni con varie reti neurali e set di dati, i nostri risultati non solo dimostrano l'efficacia dell'uso della densità di azione come metrica nella selezione dei fotogrammi, ma dimostrano anche che metodi SFS e NFS proposti risultano particolarmente efficaci nel campionamento dei fotogrammi e può contribuire a ottenere una maggiore accuratezza nel riconoscimento delle azioni rispetto ai metodi esistenti.

Parole chiave: Riconoscimento dell'azione, campionamento di fotogrammi, apprendimento con rinforzo

Contents

Introduction

Taking advantage of advanced communication, computation and smart devices techniques, action recognition emerged with machine learning and neural networks has been proposed to identify and monitor special human activities in videos [1, 10, 15]. Due to the automatic and intelligence in human actions monitoring, action recognition has been widely used in AI applications, such as Intelligent nursing, intelligent monitoring, video retrieval, etc..Recently, considerable efforts on building and developing neural networks to achieve high efficiency of action recognition have been developed, in which video frames are usually randomly or continuously selected from videos and served as the input of neural networks to identify the activities in videos [6, 7, 20–22]. However, in this scenario, because the action information in videos may not be evenly distributed in video frames, important action information included in the current video frames will be lost, leading to low accuracy of action recognition in videos. Hence, selecting rational video frames (i.e., frame sampling) as the input of neural networks to improve action recognition accuracy is also an important issue in action recognition.

To achieve the rational video frame selection for neural networks in action recognition, a number of efforts on frame sampling has been developed with the objective of achieving great completeness of action information in sampled frames and reducing the redundant information [13, 25]. For instance, S.N. Gowda et al. [8] proposed a SMART frame sampling method, which can use attention and relational model to select rational frames with high credibility. S. Yeung et al. [28] proposed a frame sampling method, namely FrameGlimpse, which can select frames based on the confidence degree predicted by RNN. However, most of these existing efforts selected each frame independently in videos and ignored the temporal continuity of sampled frames, which may cause the incomplete representation of actions in sampled frames and then achieve low accuracy of action recognition. Hence, this calls for designing a frame sampling method, which can reduce the redundant information and ensure the continuity of actions in sampled frames, thereby achieving the great accuracy of action recognition with low computational complexity.

To address these issues, in this paper, action density based frame sampling is proposed to select rational video frames for neural networks to achieve high accuracy on action recognition. Particularly, action density is introduced in our method to indicate the intensity of the actions in videos and used as a metric to assist in frame sampling, thereby achieving both the complete representation and the continuity of actions in sampled frames.

The contributions of this paper can be summarized as follows.

First, an action density determination method is proposed to determine the intensity of actions in videos, in which the motion information of actions is extracted by frame difference and background subtraction methods. Then, a reinforcement learning based frame selection (RLFS) mechanism with action densities as the reward is proposed to determine video frames that include the best action features in videos. Due to the motion information in the video is not evenly distributed in video frames, action density can be used as an effective indication to determine frames with significant features of actions. Considering the action density, the sampled frames are preferable suitable for the human visual experience.

Second, a segmented frame sampling (SFS) method is proposed to select frames based on RLFS mechanism with predefined sampling frequencies for multi-channel neural network, in which the frames with high rewards in the RLFS mechanism are selected. Our segmented frame sampling method can ensure the timing consistency of frames sampled for the upper layer and the lower layer of multi-channel neural network. That is, the sampled frames at a certain moment in the upper layer are bound to appear in the same time range of that in the lower layer. By doing this, the reliability of action recognition in the temporal space can be achieved in our SFS method.

Third, to mitigate the frame sampling aggregation on the temporal space in single-channel neural network, a non-isometric frame sampling (NSF) method is proposed select the frames with various sampling frequency for different video clips, in which the video is divided into several non-isometric clips based on the action densities, and the clips with different action densities are sampled with different frequencies. Similarly, in each clip, video frames with high rewards in RLFS mechanism are selected for single-channel neural network to achieve action recognition. In this way, the impact of frame sampling aggregation can be avoided and the integrity of actions in the sampled frames at the temporal space can be guaranteed.

Lastly, many evaluations have been conducted based on the HMDB51 and UCF101 datasets in both multi-channel neural networks (i.e.SlowFast) and single-channel neural networks (I3D, TSN, SlowOnly) to evaluate the effectiveness of using action density as a metric for frame selection, as well as the effectiveness of the proposed

SFS and NSF methods in comparison with existing schemes. The results show that action density can effectively reflect the features of different actions in videos and can be effectively used as a metric for the frame selection. In comparison with existing methods, the evaluation results also show that both the proposed SFS and NSF methods can more effectively select out the rational frames in a video and achieve better accuracy on action recognition. In addition, the results also show the effectiveness and rationality of SFS method for multi-channel neural networks and NSF method for single-channel neural networks on frame sampling.

The remainder of the paper is organized as follows: In Section 2, we conduct a literature review. In Section 3, we present our action density based frame sampling methods, including motion information based action density determination method, RLFS mechanism, and SFS and NSF method. Our performance evaluations are shown in Section 4. We conclude the paper in Section 5.

1 Related Works

Recently, a number of efforts on convolutional neural networks have been developed to improve the accuracy of action recognition [3, 9, 11, 20]. For example, X. Wang et al. [24] proposed a non-local neural network, which can achieve great action recognition efficiency based on the long-term time dependence among video frames. Some existing methods also focused on the optimizations and improvements of convolutional neural networks through decomposing convolutional kernels in various ways [16, 22, 26, 29]. In addition, the two-stream network involving both apparent flow and optical flow has been widely applied for action recognition as well [6, 7, 12, 18]. For instance, L.Wang et al. [23] proposed a temporal segment network (TSN), in which frames are sampled from the evenly divided video. However, most of these existing methods focused on the optimizations and improvements of convolutional neural networks and usually randomly or uniformly select frames to recognize actions, which may lose important action information because of the uneven motion information distribution in video frames, leading to the low efficiency on action recognition.

A number of efforts also has been developed to select the frames with significant features of actions as the input of neural network to achieve action recognition [8, 28, 30]. For example, Wu et al.[25] select the frames by LSTM, which can achieve great action recognition efficiency with fewer frames sampled. Korbar et al. [13] proposed a frame sampling method, namely SCSampler, which can select the frames that can assist in action classification in untrimmed videos. In addition, some existing methods achieve the frame sampling through reinforcement learning [4], in which a frame can be sampled if and only if the corresponding reward can be larger than the predefined threshold. However, most of these frame sampling methods focused on untrimmed videos and achieved low action recognition accuracy for short videos. Additionally, these methods select each frame independently, which may ignore the continuity of frames in the temporal space and lead to the incomplete representation of the whole coherent motion in sampled frames, thereby achieving low accuracy of action recognition.

Different from existing methods, in this paper, the action density is introduced as a metric to determine the frame with best motion features in videos by reinforcement learning method. Then, with considering the difference of neural networks used in action recognition, two action density based frame sampling methods, namely SFS and NSF, are proposed for multi-channel neural networks and single-channel neural networks, respectively, to simultaneously select a series of video frames on basis of guaranteeing the integrity and continuity of actions in sampled frames, thereby assisting in achieving great efficiency of action recognition in neural networks. In addition, our action density based frame sampling methods can be applied to both untrimmed and trimmed videos with various lengths.

2 Action Density based Frame Sampling

To achieve excellent accuracy and low complexity on action recognition in videos, in our paper, action density based frame sampling is proposed for human action recognition. Firstly, a motion information based action density determination method is proposed to determine the in tensity of actions in videos, and a reinforcement learning based frame selection (RLFD) mechanism is proposed to determine frames with the best action features. Then, a segmented frame sampling (SFS) method is proposed for multi-channel neural network and a non-isometric frame sampling (NFS) method is proposed for a single-channel neural network to select frames based on RLFD mechanism with the objective of ensuring the integrity of actions in the sampled frame and achieving great accuracy and reliability on action recognition.

2.1 Action Density Determination

An action density determination method is proposed, in which the motion information of actions is extracted by jointly Frame-Difference method and Background Subtraction method, which can achieve greater time efficiency in comparison with optical flow methods [17].

The Frame-Difference method can effectively recognize the moving objects by comparing the consecutive frames, while the Background-Subtraction method can effectively recognize the static objects by subtracting background images. Hence, the effective motion information of actions can be extracted by jointly using Frame-Difference method and Background-Subtraction method. In particular, in our paper 3-frames difference is applied for motion information extraction, due to 2-frames difference cannot achieve great efficiency for extracting the motion information of actions with high moving speed. The extracted motion information by 3-frame difference can be represented as

$$I_{fout}^{n}(x,y) = | f_{n+1}(x,y) - f_n(x,y) | \cap | f_n(x,y) - f_{n-1}(x,y) | \tag{2.1}$$

where $f_n(x,y)$ represent the pixel value in the position (x, y) of current frame $f_n(x,y)$ f_n. And $\cap$ means logical and, which means when (x, y) in two differing frames is not zero, we set it as 1.

With the Background-Subtraction method, the differential image can be obtained by subtracting a background image with the current frame, and the extracted motion information by background subtraction can be represented as

$$\begin{gathered} I_{bout}^{n}(x,y) = | f_n(x,y) - B | \\ B_n(x,y) = \frac{\sum_{i=n-k}^{n+k} f_i(x,y)}{2k+1} \end{gathered} \tag{2.2}$$

where $f_i(x,y)$ is the value of pixel (x, y) in current frame f_i. $B_n(x,y)$ is the value of background image in the position (x,y) of current frame. k is the parameter which controls the computation scale of the background images.

It is worth mentioning that although we have defined a scale parameter k to calculate the background model, the effect of background subtraction method is still unsatisfactory in the scene of camera movement. If the video with camera moving is involved in the actual application, the video needs to be cropped in advance.

Through integrating the motion information obtained by 3-frame differing and background subtraction, the action density in video frames can be determined. Due to the Background-Subtraction method can better extract the main objects in the video while Frame-Difference method can better extract the motion information of actions generated overtime, in our method, the action density is determined mainly by the Frame-Difference method and supplemented by the Background-Subtraction, which can be represented as

$$\begin{gathered} D_n = \sqrt{\sum | E(x,y) + I_{fout}(x,y) |^2} \\ E(x,y) = \begin{cases} \alpha \cdot I_{fout}(x,y), & (I_{bout}(x,y) > Avg(I_{bout})) \\ 0, & (Else) \end{cases} \end{gathered} \tag{2.3}$$

where D_n is the action density of frame f_n, α is the enhancement factor in the range of [0, 1] to enhance the impact of motion area covered by the body objects on action density determination. $Avg()$ function is aim to compute the average pixel value of the I_{bout}^{n}.

Note that, when the overlap is existed between the motion information extracted by 3-frames difference and background subtraction (i.e., $(I_{bout}(x,y) > Avg(I_{bout}))$), our method considered that the motion information extracted by 3-frames difference can better represent the action features and thus the action density can be determined by enhanced motion information extracted by 3-frames difference, i.e., $D_n = \sqrt{\sum |\alpha \cdot I_{fout}(x,y) + I_{fout}(x,y)|^2}$. Otherwise, the action density can be determined by normal motion information extracted by 3-frames difference, i.e., $D_n = \sqrt{\sum |I_{fout}(x,y)|^2}$ The introduced enhancement factor α can effectively mitigate the noise generated by lighting, environment, camera movement in action density determination.

2.2 Reinforcement Learning based Frame Selection (RLFD)

With the action density as a part of a reward function (i.e., introducing action density as a metric in frame selection), a reinforcement learning based frame selection (RLFD) mechanism is proposed in this section to select a number of frames that can best represent the action features through the one-step temporal difference method. In our RLFD mechanism, two operation actions exist for a frame: Accept and Abandon, which means to select the frame or not to extract action features. In addition, the reward of accepting a frame f_n (i.e., the operation action for the frame is accepted) can be represented as

$$R_n = \frac{D_n}{RF_n \cdot RM_n} \tag{2.4}$$

where D_n is the action density of frame f_n, R_n is the reward, RM_n represents the number of frames that still need to be selected, and RF_n is a rejection factor with the initial value of 1. In each frame selection, the parameter RM_n and RF_n can be updated as

$$\begin{cases} \begin{Bmatrix} RM_{n+1} = RM_n \\ RF_{n+1} = 1 \end{Bmatrix} & \text{if } f_n \text{ is abandoned} \\ \begin{Bmatrix} RM_{n+1} = RM_n - 1 \\ RF_{n+1} = RF_n + \delta \end{Bmatrix} & \text{if } f_n \text{ is accepted} \end{cases} \tag{2.5}$$

where δ is a penalty factor. As shown in Equation (2.5), abandoning a frame f_n will be not obtained any reward, but it will obtain a less rejection factor RF_n in comparison with accepting a frame f_n , which will lead to the increase of reward of accepting the following frames.

Similarly, when a frame is accepted, the rejection factor RF_n will be decreased by adding a penalty factor δ, resulting in the decrease of the reward of accepting the following frames. In addition, if the number of selected frames achieves the required number, the reward will be negative when additional frames are selected. By doing this, the long-term rewards are considered in our RLFD mechanism and the selection of consecutive frames can be avoided.

Based on the required number of frames that need to be selected, with the reward function (Equation (2.4)) and action function (Equation (2.5)), the required number of frames that can achieve maximum cumulative rewards can be selected out and used for action recognition.

Note that, the reason for introducing rejection factor RF_n in our RLFD mechanism is to ensure the rationality of temporal distribution of selected frames (i.e., accepted frames). For example, without the rejection factor RF_n , in a video of long-distance running, more frames that include the sprint action due to high action density in these frames and a few frames that include running action will be selected. In this scenario, the action in this video may be classified as a short-distance running with high probability, which is an unexpected action recognition result. Hence, via introducing action density D_n , the number of available candidate frame RM_n and rejection factor RF_n , frames selected by our RLFD mechanism can not only include the best action features, but also ensure the rationality of temporal distribution of actions in videos.

2.3 Segmented Frame Sampling (SFS) for Multi-Channel Neural Network

The multi-channel neural network in frame sampling is usually organized as a hierarchical pyramid recognition network, such as SlowFast [5], TPN [27], etc., in which the multiple groups of frames sampled with different predefined sampling frequencies are input into the multi-channels neural networks, and then through fusing the outputs of all multi-channels in neural networks, the action in videos can be recognized.The "multi-channel" in the multi-channel neural network here does not refer to the channel in the convolutional network, but usually means that these networks will split the input video stream into multiple input contents. At the same time, each input may have different sampling frequency and sampling method. In this section, a segmented frame sampling (SFS) is proposed to select frames based on RLFD mechanism for multi-channel neural networks to achieve effective action recognition.

In the multi-channel neural network, each channel (i.e.,each layer of hierarchical pyramid recognition network) has a predefined sampling frequency, denoted as $k_i \in \{k_1, k_2, ..., k_M\}$, where k_i is the predefined sampling frequency for the i^{th} channel and M is the total number of channels in multi-channel neural network.

In our SFS method, the total video frames are sequentially divided into several segments with an approximate number of frames in each segment. Based on the sampling frequency, denoted as k , each frame segment is also divided into several video clips, with each clip including k frames. Then, only one frame from each video clip that can achieve the highest cumulative reward in RLFD mechanism will be selected. For example, for the first channel of multi-channel neural network whose sampling frequency is k_1, the sampled frames can be represented as

$$f_1 = \arg\max \sum_{i=1}^{S_f} \sum_{j=1}^{\frac{N_f}{k_1 \cdot S_f}} \{R_{f_x^i} \mid \frac{N_f \cdot (i-1)}{S_f} + (j-1) \cdot k_1 \leq x \leq \frac{N_f \cdot (i-1)}{S_f} + j \cdot k_1, N_f\} \tag{2.6}$$

where $R_{f_x^i}$ is the reward of accepting frames f_x^i in RLFD mechanism and f_x^i is the symbol of the x^{th} frame in the i^{th} frame segment. N_f is the total number of frames in a video and S_f is the total number of frame segments divided. k_1 is the sampling frequency for the first channel of multi-channel neural network.

In fact, Equation(2.6) is a formalized representation of our frame selection method. For the multi-channel neural network, we divided them into several segments. In each segments, we process on RLFD mechanism and get the frame sequence which can get the max reward.

f_1 is the sampled frames set for the first channel of multi-channel neural network, which now can be represented as

$$\begin{aligned} f_1 = \{ & \{f_1^{1,1}, f_2^{1,1}, \ldots, f_{\frac{N_f}{k_1 \cdot S_f}}^{1,1}\}, \ldots, \\ & \{f_1^{m,1}, f_2^{m,1}, \ldots, f_{\frac{N_f}{k_1 \cdot S_f}}^{m,1}\}, \ldots, \\ & \{f_1^{S_f,1}, f_2^{S_f,1}, \ldots, f_{\frac{N_f}{k_1 \cdot S_f}}^{S_f,1}\}\} \end{aligned} \tag{2.7}$$

where $\{f_1^{m,1}, f_2^{m,1}, \ldots, f_{\frac{N_f}{k_1 \cdot S_f}}^{m,1}\}$ is the sampled frame set from the m^{th} frame segment for the first channel of multichannel neural network. Obviously, the sampled frames for the first channel of multi-channel neural network can also be organized as Sf sampled frame segments. Then, each sampled frame set $\{f_1^{m,1}, f_2^{m,1}, \ldots, f_{\frac{N_f}{k_1 \cdot S_f}}^{m,1}\}$ is divided as several video clips with each clip of k_2 frames,which is used to be selected by upper layer, and only one frame from each video clips that can achieve the highest cumulative reward in RLFD mechanism are selected as the sampled frames for second channel of multi-channel neural network, which is similar to frame selection for first channel of multi-channel neural network (i.e., Equation (2.6)).

Obviously, the sampled frames selected for the second channel of multi-channel neural network can also be organized as Sf frame segments and the frames in each segment come from the corresponding sampled frame segment for the first channel of multi-channel neural network. That is

$$f_2 = \{\{f_1^{1,2}, f_2^{1,2}, \dots, f^{1,2}_{\frac{N_f}{k_1 \cdot k_2 \cdot S_f}}\}, \dots, \\ \{f_1^{m,2}, f_2^{m,2}, \dots, f^{m,2}_{\frac{N_f}{k_1 \cdot k_2 \cdot S_f}}\}, \dots, \\ \{f_1^{S_f,2}, f_2^{S_f,2}, \dots, f^{S_f,2}_{\frac{N_f}{k_1 \cdot k_2 \cdot S_f}}\}\} \tag{2.8}$$

Similarly, the frames for the higher channels of multichannel neural network (i.e., higher layers of hierarchical pyramid recognition network) can be sampled in the same way, and the number of sampled frames for the m^{th} channel of multi-channel neural network are

$$n_m = \frac{N_f}{S_f \cdot \prod_{i=1}^{i=m} k_i} \tag{2.9}$$

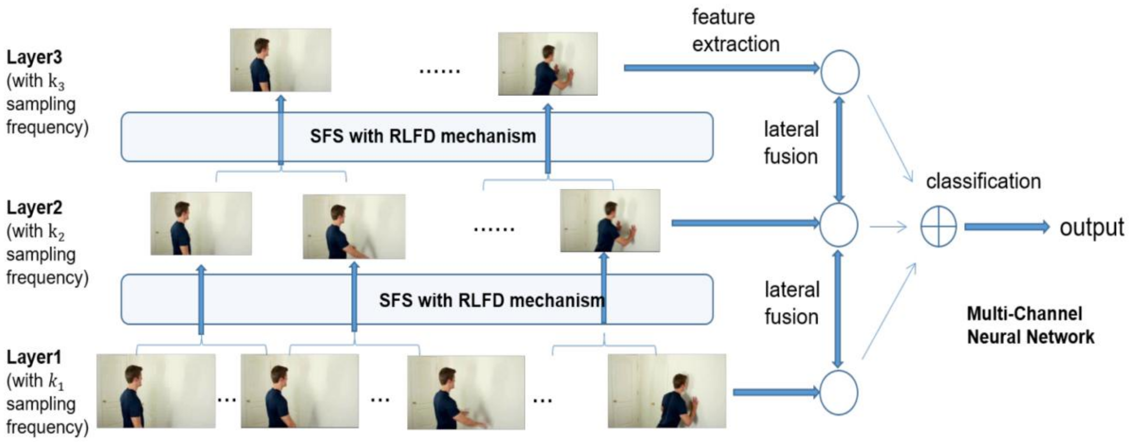

Figure 2.1:An example of segmented frame sampling for multi-channel neural network

Fig. 2.1 shows an example of segmented frame sampling for 3-channel neural network, in which the frames for the higher channel are sampled from the frames sampled for, the lower channel in the same frame segment based on RLFD mechanism. Obviously, the higher the channel, the less the number of frames sampled. Then, the frames sampled for each channel are input into the 3-channel neural network (i.e., 3-layer pyramid recognition network), and finally the outputs of all three channels are fused to recognize the actions in videos.

Hence, in our SFS method, the frames sampled for each channel of the multi-channel neural network include the frames from all original S_f frame segments of the video. By doing this, the timing consistency of frames sampled for the upper layer and the lower layer of multi-channel neural network can be guaranteed. In addition, through sampling frames from segments with fixed frames, the motion information in sampled frames is relatively uniform in the temporal space. That is, not only the frame with the highest intensity of actions, but also the motion information with less intensity of actions can be sampled for the neural network, which is beneficial to achieve highly reliable action recognition.

2.4 Non-isometric Frame Sampling (NFS) for Single-Channel Neural Network

In the field of video processing, neural networks are more single channel. Similarly, "single-channel" refers to the use of a single input in the processing of video streams. For the single-channel neural network, if the frames are sampled only according to the rewards of the RLFD mechanism, the frame sampling aggregation in the temporal space may be caused. For example, in the frame sampling for a video with a high-speed sprint, the frame including starting and ending motions may be abandoned, and frames with running actions will be concentrated sampled due to strenuous exercise in these frames. Obviously, in this scenario, the continuity and reliability integrity of the action in sampled frames will be violated, which may even affect the accuracy of action recognition. The reason is that the multi-channel network has a relatively clear segmented sampling process in the algorithm structure, so the use of SFS can better maintain the sequential structure between layers while single-channel networks have no such structure.

To this end, in this section, a non-isometric frame sampling (NFS) method is proposed for a single-channel neural network, in which the video is divided into several non-isometric video clips based on the action densities, and the frames are sampled with different sampling frequency in these non-isometric video clips. Especially, the video clips with higher action densities are considered as focused-clips and are assigned with higher sampling frequency, while the video clips with low action densities are considered as unfocused-clips are assigned with lower sampling frequency. That is, the higher the action densities in a video clip, the higher the sampling frequency for this video clip. By doing this, in our NFS method, both a mass of important motion information of actions with high action densities in focused-clips and a small number of motion information of actions with low action densities in unfocused-clips can be sampled, thereby avoiding the frame sampling aggregation and guaranteeing the integrity of actions in the sampled frames at the temporal space.

To determine the focused-clips in a video, the action density threshold should be determined in our NFS method, which can be defined as the average of the first $\beta \cdot N_f$ maximum action densities in all frames of this video, which can be represented as

$$D_{threshold} = \frac{\sum_{D_n \in Top\{\beta \cdot N_f\}\ in\ D} D_n}{\beta \cdot N_f} \tag{2.10}$$

where D_n is the action densities of frame f_n and D is the action density set of all frames. β is a scale factor with the range of (0, 1].

In our NFS method, if the action densities of a number of continuous video frames are all larger than the action density threshold, these continuous frames are organized as a focused-clip, which can be represented as

$$f_{focus} = \{f_n \mid \forall n \in [n_s, n_e], D_n \geq D_{threshold}\} \tag{2.11}$$

where f_{focus} represents the focused-clip. n_s and n_e represent the sequence number of the first frame and last frame of this focused-clip and have the constraint of $0 \leq n_s \leq n_e \leq N_f$. Similarly, the continuous video frames whose action densities are all lower than the action density threshold are organized as an unfocused clip, which can be represented as

$$f_{unfocus} = \{f_n \mid \forall n \in [n_{us}, n_{ue}], D_n \leq D_{threshold}\} \tag{2.12}$$

In this way, the video frames can be divided into several video clips alternated between focused-clips and unfocused-clips, and the frame sampling frequency in each clip (both the focused-clip and unfocused-clip) are defined based on the number of frames included in the clip, which can be represented as

$$\begin{cases} k_c = \dfrac{\omega_c \cdot N_f \cdot K}{\sum_{c=1}^{N_c} \omega_c \cdot (n_e^c - n_s^c)} \\ \sum_{c=1}^{N_c} \omega_c = 1 \end{cases} \tag{2.13}$$

where kc is the frame sampling frequency of video clip c, N_c is the total number of divided video clips, K is the total sampling frequency for the whole video, n_e^c and n_s^c represent the sequence number of the first frame and last frame of video clip c. ω_c is the weight of video clip c in total video clips, which is in the range of [0,1]. Finally, in each video clip, the frames can be sampled uniformly with the sampling frequency k_c based on the aforementioned RLFD mechanism. Through sampling frames with different frequencies determined in Equation (2.13) for video clips rather than directly sampling frames from whole video frames, the frame sampling aggregation can be avoided.

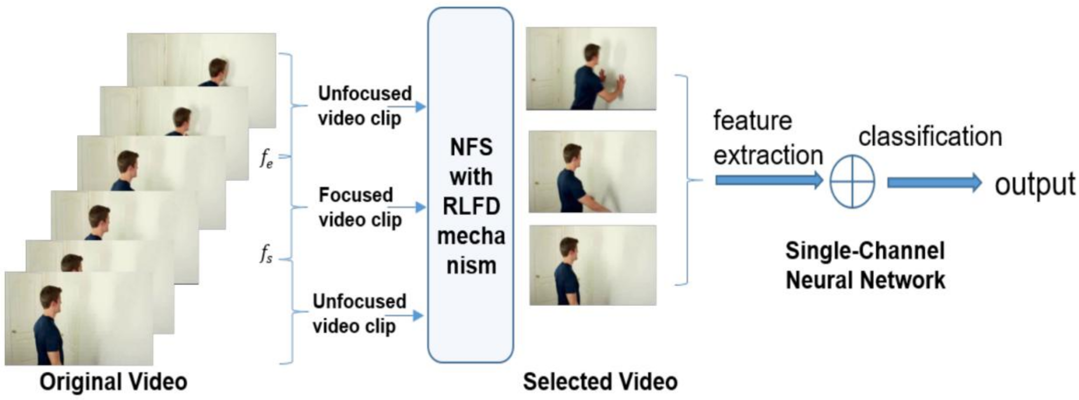

Figure 2.2:An Example of action density based non-isometric frame sampling for single-channel neural network

Fig. 2.2 shows an example of NFS for single-channel neural network, in which the original video is divided as two unfocused-clips and one focused-clip, and the focused-clip starts at frame f_s and end at frame . f_e The frames will be selected with different frequencies from these three video clips based on the RLFD mechanism for single-channel neural network to achieve action recognition.

Note that, the proposed SFS method in the last section can also be used for the single-channel neural network, in which all video frames are divided into several isometric frame segments, and in each frame segment, the frames are selected with the same frequency based on RLFD mechanism. The effectiveness of both the SFS method and NFS method for single-channel neural networks is evaluated in the next section, and the results show that NSF method achieves better effectiveness on frame sampling for single channel neural networks, which will be detailed mentioned in the following section.

3 Evaluations

In this section, the effectiveness of motion density based frame sampling (both SFS and NFS) is evaluated in terms of the accuracy of action recognition in comparison with discrete frame sampling (DFS), random frame sampling (RFS), and full-frame sampling (FFS). In particular, discrete frame sampling (DFS) can also be considered as uniform frame sampling, in which the frame a uniformly sampled from videos with a fixed sample frequency. Random frame sampling (RFS) is to randomly select frames from videos with a fixed sample frequency. The full-frame sampling (FFS) is to select all frames in the video for the neural network to achieve action recognition.

3.1 Datasets and Preparation

In this experiment, two data sets, HMDB51[14] and UCF101[19], which are relatively mainstream in the field of behavior recognition, were selected.

Brown University released HMDB51 datasets in 2011. Most of the videos in this datasets are from movies, and some of them are from public databases and online video libraries such as YouTube. The database contains 6849 samples,which divided into 51 categories, each category contains at least 101 samples. Hmdb51 datasets have small amount of data, convenient training, clean background and good characteristic difference between action classes, which is convenient for various experiments.

UCF101 is a series of databases published by the University of Central Florida(UCF) since 2012. The database samples come from a variety of sports samples collected from BBC/ESPN radio and television channels, as well as samples from YouTube. The sample consisted of 13,320 videos in categories such as makeup, music equipment and sports. UCF101 has the greatest diversity in action, and there are great differences in camera movement, object appearance and posture, object proportion, viewpoint, chaotic background, lighting conditions and other aspects. Videos from the same group may have some common characteristics, such as similar background, similar perspective and so on.

The depth learning framework used in these experiments is PyTorch. UCF101 and HMDB51 input frames are all 224x224 in size. For partial enhancement, the default enhancement coefficient α is 0.3, and for strategy iteration, the default penalty coefficient is 0.3 for parameter setting in the reward function.

Although the frame sampling method itself does not involve the construction of neural network. But in order to verify the effect of frame sampling, we still need to train on the neural network skeleton. We set the hyper parameters momentum=0.9, LR =0.1, weight_decay=0.0001 and min_LR =0. In these experiments, UCF101 and HMDB51 datasets were selected to train 250 epochs on SlowFast[5] and SlowOnly[5], and 100 epochs on I3D[2] network. Meanwhile, in the control experiment for sampling methods, HMDB51 was used to train 50 epochs on SlowOnly network and TSN[23] network respectively.

In the evaluation of action recognition accuracy, Top1−accuracy and Top5−accuracy are considered, in which Top1−accuracy means that the probability of real action is the top one recognized action, and Top5−accuracy means that the probability of the real action in the top five recognized actions. In addition,SlowFast [5] is used as the multi-channel neural network, and SlowOnly [5], I3D [2] and TSN [23] are used as the single-channel neural networks.

3.2 Effectiveness of RLFD mechanism

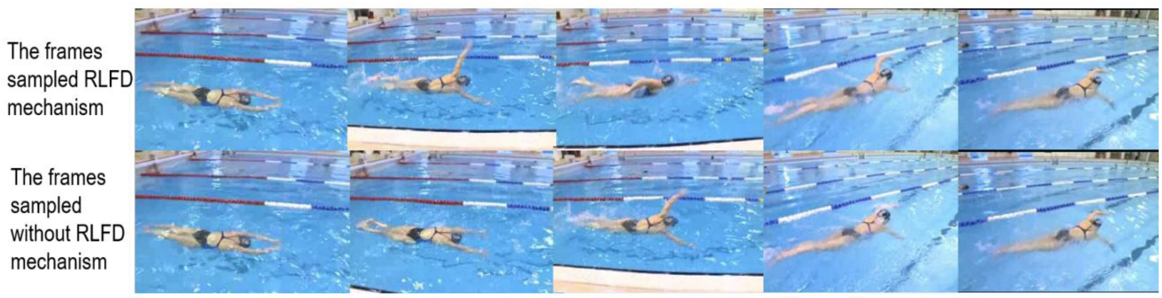

Figure 3.1:The frames sampled with and without RLFD mechanism

Fig. 3.1 shows the frames sampled with and without RLFD mechanism for single-channel neural networks, respectively. As is shown in Fig. 3, in comparison with the frames sampled without RLFD mechanism (i..e, the second row of Fig. 3), the frame sampled with RLFD mechanism (i..e, the first row of Fig. 3.1) can more directly show that the people in the video are swimming, which can demonstrate that the RLFD mechanism can effectively select out the frames with the best action features.

3.3 Effectiveness of SFS for Multi-Channel Neural Network

	Method	F=0.125	F=0.25
Baselines (Backbone: SlowFast)	Random	89.55	86.54
	Uniform	90.72	86.68
	AAS	91.60	86.92
Our Method (Backbone: SlowFast)	MDFS（our method）	91.55	87.10
	MDFS（our method） $\alpha = 0.3$	**91.83**	**87.46**

Table 3.1:The action recognition accuracy of SFS, DFS and RFS with SlowFast [5] and UCF10 [19]

Table 3.1 shows the Top5 − accuracy rate on action recognition in our segmented frame sampling (SFS), discrete frame sampling (DFS), and random frame sampling (RFS), and attention aware sampling (AAS) [4] with SlowFast neural network and UCF101 datasets, in which the discrete frame sampling (DFS) is evaluated with the same sampling frequency to our segmented frame sampling (SFS). As shown in Table 3.1, the Top5 − accuracy rate of our SFS method is larger than that of the RFS, DFS and AAS method. That means sampling frames with considering action density can select frames with more prominent motion features for neural networks, thereby achieving more greater accuracy on action recognition. In addition, the results also show the effectiveness and rationality of our segmented frame sampling method for multi-channel neural networks. Additionally, our SFS method with enhancement factor α (i.e.,α = 0.3) achieves better accuracy than that without enhancement factor α

(i.e.,α = 0), which shows show the effectiveness of introducing enhancement factor α in our action density method.

3.4 Effectiveness of NFS for Single-Channel Neural Network

	Method	F=0.25	F=0.5
Baselines (Backbone: SlowOnly)	Random	80.84	81.73
	Uniform	80.97	82.02
	All frames	83.74	83.74
	MDFS（our method）	81.40	83.08
Our Method (Backbone: SlowOnly)	MDFS（our method） $\alpha = 0.3$	**82.62**	**84.65**

Table 3.2:The action recognition accuracy of RFS, DFS, FFS and NFS with SlowOnly [5] and UCF101 [19]

Table 3.2 shows the Top5 − accuracy rate on action recognition in our non-isometric frame sampling (NFS), discrete frame sampling (DFS), random frame sampling (RFS) and full-frame sampling (FFS) with SlowOnly neural network and UCF101 datasets.

As shown in Table 3.2, in comparison with RSF, DFS and FFS, when the sampling frequency is 0.5, our NFS can achieve the greatest accuracy on action recognition. As the sampling frequency is reduced to 0.25, the accuracy of all methods is reduced, our NFS method achieves better accuracy than RFS and DFS. Although our NFS method achieves a little less accuracy than the FFS method, the frames sampled in our NFS method are much less than that in the FFS method. That means our NFS method can achieve great action recognition accuracy with low complexity, due to reducing a large amount of redundant video information. In addition, the results also show the effectiveness of our NFS method with enhancement factor α (i.e.,α = 0.3) for single-channel neural network on frame sampling.

3.5 Effectiveness of NFS and SFS for Single-Channel Neural Network

Method	Acc.Top1	Acc.Top5
I3D r50	34.88	64.31
I3D r50 （+ SFS）	36.07	65.65
I3D r50 （+NFS）	**38.03**	**67.02**

Table 3.3:The action recognition accuracy of NFS and SFS with I3D [2] and HMDB51 [14]

Admittedly, single-channel neural networks can also use SFS for sampling. However, as described earlier in this paper, we designed a new method for single-channel identification networks, NFS, which also explains the necessity and rationality of the proposed new method. Table 3.3 shows Top1−accuracy rate and Top5−accuracy rate on action recognition in NFS and SFS with single-channel neural network (i.e., I3D r50) and HMDB51 datasets.

As is shown in Table 3.3, both Top1−accuracy rate and Top5−accuracy rate in NFS are better than that in SFS. The results demonstrate that the proposed NSF is more suitable for the single-channel neural network in comparison with the SFS method. In addition, by comparing the improvement range of top1 accuracy and Top5 accuracy, it can be found that top1 accuracy has a relatively high accuracy improvement from 36.07 to 38.03, while top5 accuracy only increased from 65.65 to 67.02, which reflects the addition of sampling scheme to determine important regions. In more stringent identification requirements (that is, the application environment with higher accuracy requirements), it can reflect better classification characteristics.

3.6 Impact of penalty factor δ in RLFD mechanism on the accuracy of action recognition

	Acc.Top1	Acc.Top5
TSN-rgb (MDFS with $\delta = 0$)	47.45	77.24
TSN-rgb (MDFS with $\delta = 1$)	47.58	78.05
TSN-rgb (MDFS with $\delta = 0.3$)	**48.21**	**78.64**

Table 3.4:The action recognition accuracy of NFS with different penalty factor δ in RLFD mechanism in TSN [23] and HMDB51 [14]

Table 3.4 shows the Top5−accuracy action recognition accuracy of NFS when penalty factor δ in RLFD mechanism is set as 0, 0.3, 1, respectively. As shown in Table 4, when the penalty factor δ is 0.3, the greatest accuracy can be achieved. The reason is that when the penalty factor δ is 0, no penalty will be imposed on frame selection in the RLFD mechanism.

In this scenario, frames with strenuous exercise will be selected, and frames with smooth motion will be dropped, which will lead to uneven temporal distribution of motion information in sampled frames, thereby resulting in the low accuracy of action recognition. While, when the penalty factor δ is 1, the NFS method will be similar DFS method (i.e., uniform sampling).

In this case, although the uneven temporal distribution of motion information in sampled frames can be avoided, frames with important action features will be dropped as well, which will also lead to low accuracy on action recognition. Hence, the results show that the penalty factor δ can achieve great efficiency on not only ensuring the rationality of temporal distribution of selected frames, but also selecting out frames with best action features, thereby assisting in achieving great accuracy on action recognition.

3.7 The efficiency of NFS on improving neural network

Method/Datasets	HMDB51	UCF101
I3D	79.86	97.78
I3D+SMART	81.10	98.20
I3D+MDFS	**81.27**	**98.35**

Table 3.5:The action recognition accuracy of I3D [2] without enhancement and with existing SMART [8] and NFS as the enhancement module

The proposed NFS method not only can be used as a frame pre-processing method, but also can be emerged into neural networks as an enhancement module to improve the effectiveness of neural networks on action recognition. Table 3.5 shows the Top(5) –accuracy action recognition accuracy without any enhancement module in the I3D neural network and with the existing SMART method and our NFS method as the enhancement module in the I3D neural network, respectively. Table 5 shows that in both HMDB51 and UCF101 datasets, the I3D neural network with our NFS method as the enhancement module can achieve the greatest action accuracy than that without any enhancement module and with the existing SMART method as the enhancement module. That means our NFS method also can be served as an enhancement module for neural networks to improve the accuracy of action recognition.

4 Conclusion and future development

In this paper, action density based frame sampling is proposed to assist in action recognition. In particular, an action density determination and a reinforcement learning based frame selection mechanism are proposed to select frames with the best action features. Then, a segmented frame sampling (SFS) method and a non-isometric frame sampling (NFS) method are proposed for multi-channel neural networks and single-channel neural networks, respectively. The valuation results show that our frame sampling methods (both SFS and NFS) can effectively preserve the integrity and continuity of actions in the sampled frames and can assist in achieving greater action recognition accuracy in comparison with existing schemes.

The work of this thesis also has optimization research direction. First of all, this paper adopts three-frame difference method and background difference method to extract the video action information, which can maintain a better effect in the video with less interference. When there is noise in the video (for example, the camera moves at high speed, the moving subject has occlusion, etc.), there will be some error in the evaluation of the motion information in the video. In the future work, the extraction method of motion information can be further designed. For example, for the video with strong noise, the extraction method combined with optical flow is adopted. Although it will consume more computing resources, it can maintain a good evaluation range for the high-noise picture.

Bibliography

[1] Sadjad Asghari-Esfeden, Mario Sznaier, and Octavia Camps. Dynamic motion representation for human action recognition. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 546–555, 2020.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, 2017.

[3] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng. Multi-fiber networks for video recognition. 2018 The European conference on computer vision (ECCV), 2018.

[4] Wenkai Dong, Zhaoxiang Zhang, and Tieniu Tan. Attention aware sampling via deep reinforcement learning for action recognition. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 8247–8254. AAAI Press, 2019.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6201–6210, 2019.

[6] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal multiplier networks for video action recognition. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7445–7454, 2017.

[7] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1933–1941, 2016.

[8] Shreyank Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. arXiv preprintarXiv:2012.10671, 12 2020.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[10] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Ana lyzing temporal information in video understanding models and datasets. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7366–7375, 2018.

[11] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):221–231, 2013.

[12] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[13] Bruno Korbar, Du Tran, and Lorenzo and Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6231–6241, 2019.

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In 2011 International Conference on Computer Vision, pages 2556–2563, 2011.

[15] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4694–4702, 2015.

[16] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatiotemporal representation with pseudo-3d residual networks. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 5534–5542, 2017.

[17] L. Sevilla-Lara, Y. Liao, F Guney, V. Jampani, A. Geiger, and M. J. Black. On the integration of optical flow and action recognition. German Conference on Pattern Recognition, 2018.

[18] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. Computer Science, 2014.

[19] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. Computer Science, 2012.

[20] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 4489–4497, 2015.

[21] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classifification with channel-separated convolutional networks. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 5551–5560, 2019.

[22] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6450–6459, 2018.

[23] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L Van Gool. Temporal segment networks: Towards good practices for deep action recognition. Springer, Cham, 2016.

[24] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7794–7803, 2018.

[25] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S. Davis. Adaframe: Adaptive frame selection for fast video recognition. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1278–1287, 2019.

[26] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatitemporal feature learning: Speed-accuracy trade-offs in video classification. 2017 The European conference on computer vision (ECCV), 2017.

[27] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 588–597, 2020.

[28] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei Fei. End-to-end learning of action detection from frame glimpses in videos. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2678–2687, 2016.

[29] S. Zhang, S. Guo, L. Wang, W. Huang, and M. R. Scott. Knowledge integration networks for action recognition. 2020 The AAAI Conference on Artifificial Intelligence, 2020.

[30] Yin-Dong Zheng, Zhaoyang Liu, Tong Lu, and Limin Wang. Dynamic sampling networks for efficient action recognition in videos. IEEE Transactions on Image Processing, 29:7970–7983, 2022

List of Figures

List of Tables

Acknowledgements

As a double-degree postgraduate between POLIMI and XJTU, I am honored to study and live in POLIMI for one and a half years. During this year and a half, I gained knowledge and friendship, and experienced different local conditions and customs. Although I have taken online courses for a period of time due to COVID-19, I am still happy to come to POLIMI for exchange and study.

I would like to thank Professor Luciano Baresi for answering many of my questions as the double degree exchange instructor. I would like to thank Professor Livia De Zan, who has given me great help on various issues concerning the double degree.

Thanks to my supervisor in POLIMI, Professor Marco Marcon, who put forward a lot of revision suggestions during the completion of my thesis. Under his guidance, I could finish the paper smoothly. And I would like to thank my tutor Professor Jie Lin in XJTU for providing the experimental environment and academic suggestions to complete the thesis experiment.

Finally, I would like to express my gratitude for the friendship between POLIMI and XJTU, which gives me the opportunity to carry out this meaningful double-degree exchange visit during my master's degree studying!