Subjective Video Attribute Recognition: Labeling, Training, and Explaining

Ying Ji

Acknowledgements

It took me seven years from the moment I arrived in Japan until the completion of my doctoral dissertation. This difficult journey made me a better person and had a profound impact on my life. Throughout this lengthy and difficult journey, a number of people provided me with invaluable assistance. I would like to express my sincere appreciation to everyone.

I would like to express my gratitude to my supervisors, Professor Kensaku Mori and Professor Jien Kato, for their kind guidance, feedback, and encouragement. Without consistent support from them, I could not reach my destination. I would like to thank my co-author, Associate Professor Yu Wang, for his patience and generous assistance during my Ph.D. study. I am also grateful to my thesis committee, Professor Ichiro Ide, Professor Yoshito Mekada, Associate Professor Hiroaki Kudo, Associate Professor Daisuke Deguchi, and Associate Professor Masahiro Oda for their invaluable contributions. Their constructive suggestions and thought-provoking questions guided me to improve my thesis and enrich the depth of my research.

Finally, I would like to thank my dearest friend and beloved family. To my best friend, Yuqi Luo, thank you for your constant companionship and encouragement. To my parents and grandparents, Yuan Chen, Qing Ji, Budao Ji, and Yuanqin Hua, thank you for your unwavering support and endless love throughout the challenging and arduous journey of my academic endeavors. Without your love, I would not have made it this far. I dedicate my dissertation work to you.

Abstract

The analysis of video attribute recognition can be divided into two classes: Objective Attribute Recognition (OAR) and Subjective Attribute Recognition (SAR). The OAR is primarily concerned with recognizing tangible and stable characteristics, such as actions, objects, or scenes in a video. However, SAR is significantly more challenging since the attributes are defined by human perceptual cognition and are influenced by individual experiences or emotions. Subjective attributes such as image quality, aesthetics, and video popularity, are even difficult to define and recognize. The objective of this thesis is to investigate ground-truth generation, attribute recognition, and model explanations for subjective attributes. With this target, this thesis proposes two subjective video datasets, two training approaches, and one explainable module for SAR.

This thesis contains six chapters.

Chapter 1 describes the background and motivation of the work. The overview of the proposed methods is also concluded.

Chapter 2 introduces the related studies about OAR and SAR. Specifically, research relevant to ground-truth generation, subjective attribute recognition, and explainable artificial intelligence are presented.

To enhance the performance of SAR, this thesis addresses two practical challenges. Chapter 3 introduces an improved dataset labeling method designed for video violence recognition. Chapter 4 provides a training approach for improving the recognition accuracy of social relation atmosphere. Chapter 5 offers an improved model understanding method for explaining the inner procedures of 3D Convolutional Neural Networks (CNNs). By focusing on these three key difficulties of SAR, the recognition can be finally enhanced.

Chapter 3 introduces the process of constructing a clean dataset with reliable and stable ground truth. Currently, data annotations are often provided as single

labels with majority voting. The annotators give an absolute value for the objective attribute. However, the subjective data do not have an exact ground truth. This chapter solves the problem by introducing a pairwise comparison method. The pairwise comparison can reduce the ambiguity and divergence in the annotating process. This chapter takes violence extent analysis as an example and provides a new dataset with subjective violence extent labels. Consequently, a rank learning method is specially designed to estimate the violence extent. Considering the violence extent of each video is a relative attribute and can be compared, the proposed method can learn the relationship between videos at the same level and different levels.

Chapter 4 introduces an efficient way to recognize and represent subjective attributes. End-to-end network is a commonly used method to recognize videos. However, the features directly extracted from the networks are not specific enough to represent the data. A new dataset is first created with both subjective social relation atmosphere attribute and objective social relationships attribute. A 3D explanation module proposed in Chapter 5 is used as a plug-in module in this chapter. The module can be used to extract the most important regions for recognizing social relationships. A framework is proposed to leverage the important information from social relationships to enhance social relation atmosphere recognition. In this way, the recognition of subjective attributes is increased by importing useful objective information. The fused features are more representative than features directly extracted from end-to-end neural networks.

Chapter 5 introduces a spatial-temporal concept-based explanation method for explaining neural networks. Currently, video-based explanation methods mainly focus on pixel-level interpretation. None of them are able to produce a high-level explanation. An STCE (Spatial-Temporal Concept-based Explanation) framework is proposed for interpreting 3D CNNs and explaining them by introducing human-understandable concepts. The concepts are grouped by supervoxels extracted from videos. The framework evaluates the importance score for each concept. The high score represents the network that pays more attention. The proposed framework is utilized in Chapter 4 as a plug-in module to help recognize the subjective attributes.

Chapter 6 gives the summary and prospect of this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The recognition of video attributes can be divided into two classes: Objective Attribute Recognition (OAR) and Subjective Attribute Recognition (SAR). Objective attributes refer to the characteristics that will not be influenced by human perception or personal biases, such as cars and airplanes. They can be measured and evaluated. Subjective attributes, in contrast, are vaguely defined and primarily based on self-perceived knowledge. They are easily influenced by personal viewpoints, educational background, individual experience, and other factors. On the other hand, subjective attributes may be evaluated differently from person to person, and there is no exact numerical value for subjective characteristics. For example, estimating the interestingness or popularity of a video belongs to SAR.

With the rapid development of multimedia and computer vision technologies, it has become easier for people to gain access to an abundance of videos. Estimating subjective attributes in video data can be useful in many real-world applications [1, 2] and has garnered considerable interests. However, due to uncertain and ambiguous properties of subjective attributes, less research has been conducted on the issue of SAR. The challenges of recognizing subjective attributes range from labeling, training, and explaining.

As shown in Figure 1.1, an underlined annotation indicates the objective attribute in a movie segment; social relationships between two people. The social relationships are pre-defined in the scenario and can be uniquely determined

Figure 1.1: Example video frames of SAR and OAR. The frames are from two movie clips. An underlined annotation indicates the objective social relationship. The other annotations indicate the subjective social relation atmosphere. Both two people in the two video examples are in the same social relationship: "friend", but the social relation atmospheres are different.

when given sufficient information. The other annotations indicate the subjective attribute; social relation atmosphere. Obviously, it is difficult for annotators to precisely describe the ground truth of the social relation atmosphere. Thus, providing a reliable annotation for the subjective attributes becomes a major problem.

On the other hand, by observing the samples in Figure 1.1, we can see that even though the objective attribute is the same in various videos, the subjective attributes may still be different. Identifying subjective attributes is a significantly more challenging task. We need to take into account not only human interactions and facial expressions, but also the surrounding environment and numerous attributes. However, the performance of current end-to-end Convolutional Neural Networks (CNNs) is insufficient for recognizing subjective attributes. Thus, effective methods of recognition are urgently required.

Finally, video data contains complex spatial-temporal information, which requires high computation costs. On the other hand, the decision procedure of SAR is complex and opaque. Providing an explainable framework for recognition can help investigate SAR in depth and bring machines one step closer to human cognition. The research questions of SAR can be summarized as the following three questions:

1. How to construct a clean dataset and provide stable and reliable annotations for subjective attributes.

2. How to improve the accuracy of SAR and generate targeted features.

3. How to explain the inner procedure of 3D CNN.

This thesis digs into these issues by solving two real-world SAR tasks. As discussed above, a labeling method is introduced to reduce the ambiguity of subjective attributes. Consequently, a relative fusion framework that employs multiple objective attributes is presented to help recognize subjective attributes. Lastly, a high-level explanation module is proposed to look into the decision procedure of 3D CNNs, which can be used to enhance the SAR performance.

1.2 Research Overview and Thesis Structure

1.2.1 Research Overview

There has been a lot of effort in SAR. However, the performance of annotation, recognition, and explanation can still be greatly improved, as previously stated. To address these three issues, three frameworks are presented in this thesis: Ground-truth generation (Chapter 3), subjective attribute recognition (Chapter 4), and spatial-temporal model explanation (Chapter 5). The first question raised in Section 1.1 is primarily covered in Chapter 3, the second question in Chapter 4, and the third question in Chapter 5.

A large-scale dataset is a crucial component of machine learning algorithms. Crowdsourcing is a popular way of collecting and annotating data because of its low cost, high speed, and diversity of viewpoints [3, 4]. Following crowdsourcing, majority voting is currently considered the optimal technique to obtain a single label from multiple labels [5, 6, 7]. However, the biggest obstacle to majority voting is quality control [8]. Due to the lack of expertise and large bias between annotators, the labels are inconsistent and noisy, so the quality of labels can be low. Since providing annotations for subjective attributes is a more abstract task, utilizing crowdsourcing with multiple single labels is not suitable. Thus, a pairwise comparison labeling method is proposed in this thesis to reduce the bias between labels.

In Chapter 3, a particular SAR task is explored, which is the violence extent of violent videos. As shown in Figure 1.2, traditional crowdsourcing annotation

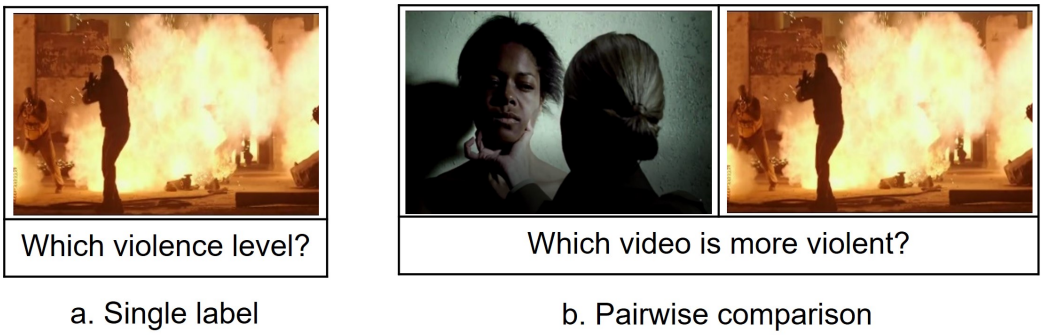


Figure 1.2: Examples of single label and pairwise comparison.

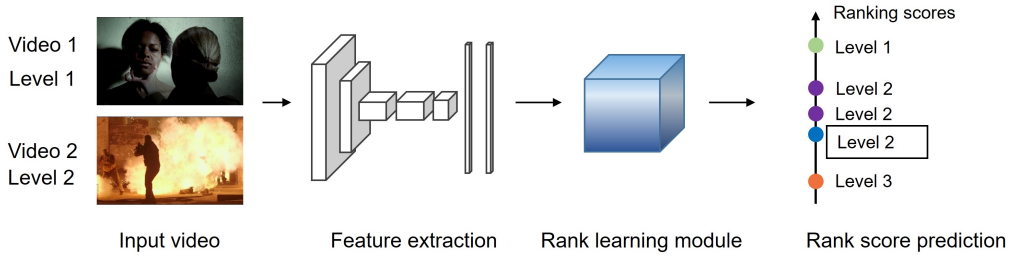


Figure 1.3: Simplified illustration of the proposed rank learning method for recognizing video violence extent in Chapter 3.

prefers single labels. However, when labeling violence extents, the boundaries between different violence extents are very unclear. Individual experiences, cultural backgrounds, and beliefs may influence the labels of various annotators, which result in distinct labels. It is difficult to arrive at a stable judgment on "Which violent level the video belongs to". On the other hand, comparing two videos about which one is more violent is easier to answer. Given a sufficient number of comparisons, the judgment will become more stable than the single label [9, 10]. Therefore, in order to provide a reliable and stable annotation for subjective violence extent, the ground truth is given by pairwise comparison in Chapter 3. A new violent video dataset is also proposed.

In contrast to the regular video classification task, in which the labels are independent, the violence levels between each pair of videos have a strong correlation. Thus, in Chapter 3, different from the end-to-end classification method, a rank learning-based method is specifically designed for automatically estimating the violence extent. A simplified illustration of the proposed method is shown in

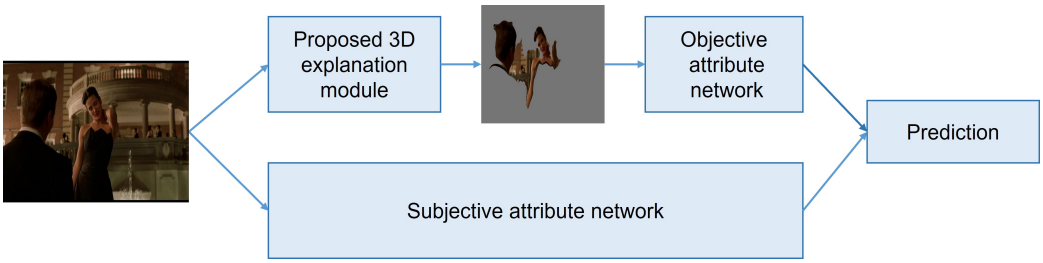


Figure 1.4: Simplified illustration of the proposed feature fusion method in Chapter 4.

Figure 1.3. The proposed rank learning module uses pairs of videos as input to learn the relationship between two videos according to the level of violence they belong to. Based on the predicted violence score, the videos can be classified into different levels. Thus, by making the most of the relationship between videos, the accuracy can be further improved.

In Chapter 3, a unique SAR method is developed for the violence extent estimation. However, there are still many subjective attributes that are independent and can not be measured using the rank learning-based method. A general SAR method is urgently needed. With the development of large-scale datasets, deep neural networks [11] become the first choice for representing images or videos. However, an end-to-end single-column network takes singular information as input without paying special attention to subjective attributes. To address this problem, a relative feature fusion method is proposed in Chapter 4.

A simplified illustration of the proposed method is shown in Figure 1.4. Each video contains both subjective and objective attributes. The proposed explanation module detects the most important information in the raw video for recognizing the objective attribute. The irrelevant pixels are masked and features are extracted from the masked video. A second network is used to extract features for SAR. Heterogeneous information is leveraged to make the final prediction. The proposed method utilizes objective information to supplement SAR, which can provide more representative features than an end-to-end neural network. Since there is no existing dataset that contains labels on both subjective and objective attributes, a new dataset is constructed with both attributes.

Although the proposed method with deep learning technologies shows outstanding performance, the network itself is still a black box, making the prediction

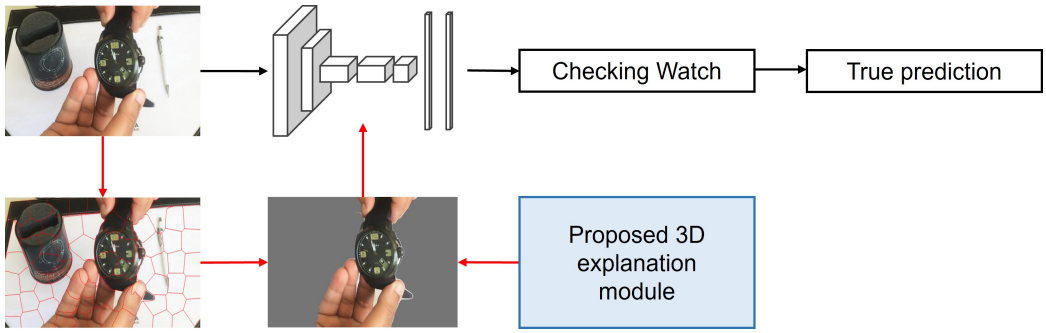


Figure 1.5: Simplified illustration of the proposed explanation method in Chapter 5.

procedure opaque. Understanding the inner principle of deep neural networks is essential for further enhancing the performance of video recognition as well as SAR. Current research on 3D CNN interpretation remains mired in pixel-level explanation, making it still difficult for humans to comprehend.

In Chapter 5, a high-level explanation module is proposed to explain the prediction for video recognition. A simplified illustration is shown in Figure 1.5. Raw videos are segmented into multiple supervoxels. They are video clips that are easy to be visualized and understood. The proposed explanation module calculates the importance score for each supervoxel based on the classification result. The high score supervoxels indicate the network relies more on these parts to make judgments. With only the most important video information, the network can still recognize the masked video as the true prediction. The proposed method not only can make the video recognition procedure more intelligible but also can be used to improve SAR. It is utilized in Chapter 4 as a plug-in module.

1.2.2 Thesis Structure and Chapter Relationship

This thesis is composed of six chapters.

Chapter 1 provided an overview of the background of this thesis, and discussed the existing insufficiency and motivation of this thesis. In addition, a general overview of each proposed strategy was also presented.

Chapter 2 introduces the studies related to ground-truth generation, subjective attribute recognition, and explainable artificial intelligence.

Chapter 3 introduces a violent video dataset with a subjective violence extent

annotation. The annotation is performed by pairwise comparison. It mainly solves the question: ***How to construct a clean dataset and provide stable and reliable annotation for subjective attributes***. Besides, designed for the rating estimation dataset, a rank-learning method that can learn the contrastive relationship is proposed. This chapter provides a better dataset annotating method.

Chapter 4 introduces a dataset and proposes a relative feature fusion method for recognizing subjective attributes. A dataset with both subjective and objective attributes is proposed. Instead of using raw video data as input, only the key information is preserved, and the key information from the objective attribute is employed to assist the SAR. The fused features provide a more representative representation. It mainly solves the question: ***How to improve the accuracy of subjective video attribute recognition and generate targeted features***. This chapter provides a better training method.

Chapter 5 proposes a spatial-temporal concept-based explanation method. The high-level concepts are easy to understand and the visualizations are consistent with human cognition. It mainly solves the question: ***How to explain the inner procedure of 3D CNN***. This chapter provides a better model understanding method.

Chapter 6 concludes the thesis by reviewing the research contribution and proposes future work for this thesis.

Chapter 3 to Chapter 5 are the core chapters of this thesis. With these three core chapters, the subjective video attributes can be better recognized. The relationships and contributions of each chapter are illustrated in Figure 1.6.

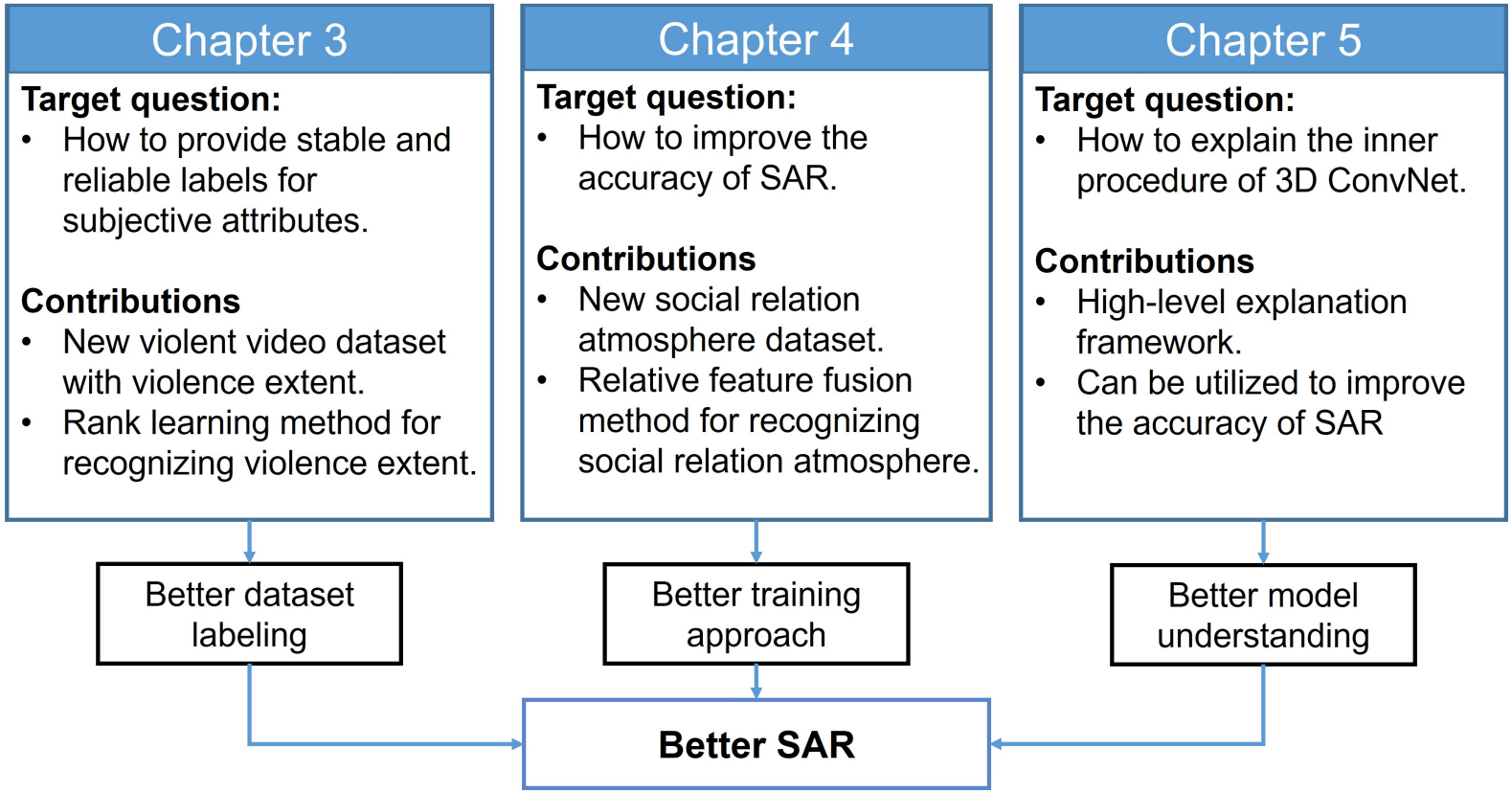


Figure 1.6: Relationships between chapters.

Chapter 2

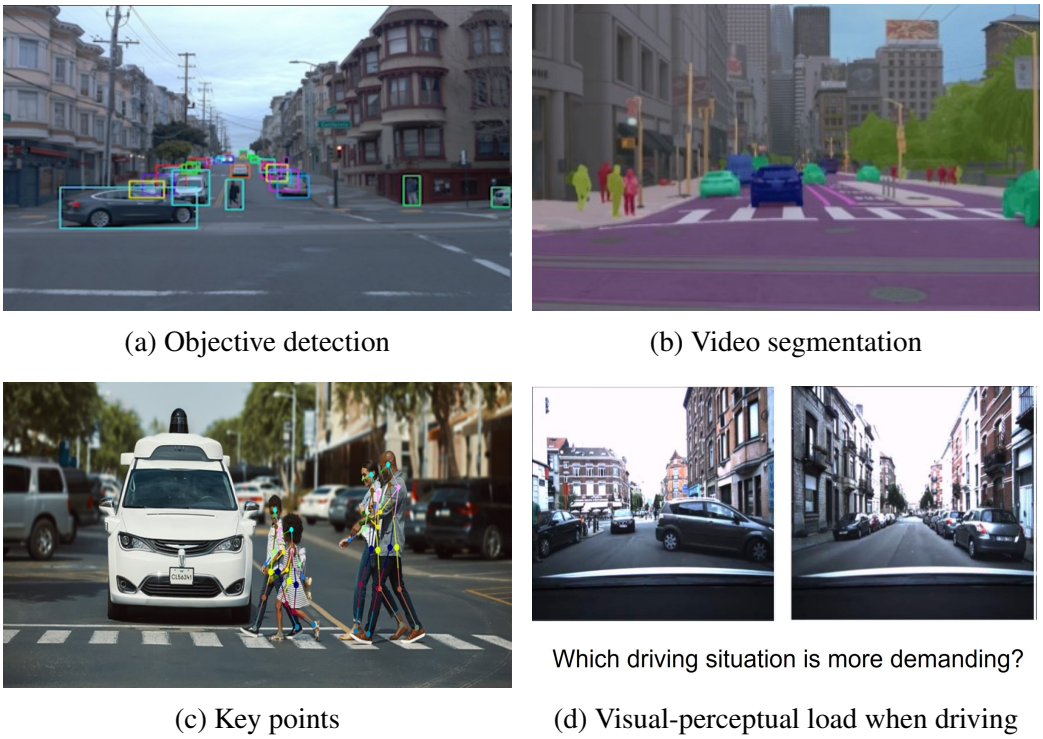
Related Work

2.1 Objective and Subjective Video Attribute Recognition

Video recognition refers to the procedure of automatically analyzing and comprehending the contents of video data, which is a significant research area in computer vision. Depending on the subject of analysis, video recognition can be divided into two classes: Objective Attribute Recognition (OAR) and Subjective Attribute Recognition (SAR).

OAR is the process of identifying or classifying tangible objects or entities in a video. The recognition is based on standardized criteria and will not be affected by personal opinions or evaluations. It includes object detection [12, 13, 14], object segmentation [15, 16, 17], human skeleton detection [18, 19], and so on. Examples of OAR are shown in Figure 2.1a, Figure 2.1b, and Figure 2.1c. Since the recognition results are consistent and impartial, the ground truth can be measured and uniquely confirmed.

SAR, on the other hand, focuses on providing analysis based on personal opinions or feelings. Individual experiences, personal emotions, cultural backgrounds, or biases can influence the recognition results. Since judgment can vary from individual to individual, there is no exact ground truth regarding subjective attributes. Most of the existing subjective recognition methods mainly concentrate on image-based analysis, such as image aesthetics assessment [20, 21, 22], image memorability prediction [23, 24], and so on. However, due to the complexity and



(a) Objective detection

(b) Video segmentation

(c) Key points

(d) Visual-perceptual load when driving

Figure 2.1: Example of objective [28] and subjective video recognition. (a) Object detection. (b) Semantic segmentation. (c) Key points of the human body. (d) Visual-perceptual load when driving, which indicates the amount of visual information received when driving [29].

abundant temporal information of video data, analyzing subjective attributes in videos still remains a challenging task. Only a few studies focus on subjective video recognition [1, 25, 26, 27]. An example is shown in Figure 2.1d; The estimation of the visual-perceptual load when driving is subjective and based on the vast driving environment. There is no ground truth.

The research on SAR is of great social value, since first, it can be used to simulate human cognition and can even achieve recognition ability beyond humans. For example, Palmer *et al.* [29] automatically detected the visual-perceptual load when driving, which estimated the visual information received by drivers when driving. It can compensate for the limitations of human perception. With cognitive abilities that surpass the humans, recognizing subjective attributes in videos also has numerous practical applications. By understanding and ranking the interestingness of videos [1, 30, 31], video search and recommendation systems

can provide more personalized results. Recognizing subjective human emotion has numerous applications in human-computer interaction and health monitoring [32, 33, 34, 35]. Human beauty and attractiveness recognition can be utilized in make-up evaluation, facial image enhancement, or social network recommendation applications [36, 37].

2.2 Ground-Truth Generation

The remarkable success of deep learning in computer vision is significantly influenced by the expansion of large-scale labeled datasets [38, 39, 40], such as ImageNet [41] and UCF-101 [42]. Typically, a dataset consists of data samples and their ground truth. However, as mentioned before, providing ground truth for subjective attributes is a challenging task. In this section, related work about a common labeling technique, majority voting with single label, and an innovative technique, pairwise comparison will be introduced.

2.2.1 Single Label with Majority Voting

With the rapid expansion of crowdsourcing platforms such as *Amazon Mechanical Turk*[1], the use of crowdsourcing to collect or analyze data in various research fields has increased exponentially [8]. In 2016, over 40% of the behavioral studies published in *Journal of Consumer Research* were collected and analyzed by using crowdsourcing Websites. The widespread use of crowdsourcing enables researchers to collect data in a shorter amount of time and at a lower cost. In order to aggregate a single label from multiple labels, majority voting is a common technique in most current subjective attribute datasets.

Datta *et al.* [43] collected $3,581$ photos from an online photo-sharing community *Photo.net*[2]. The registered members of the Website, which include both amateur and professional photographers, were asked to rate the overall aesthetic score of each image on a scale from one to seven. Thus, the average score is calculated as the aesthetic score. AVA (Aesthetic Visual Analysis) is also a novel aesthetic image dataset containing over $250,000$ images with numerous aesthetic

[1] `https://www.mturk.com/` (Accessed: 2023/09/01)
[2] `https://www.photo.net/` (Accessed: 2023/09/01)

scores for each image [44]. The meta-data were collected from a social network *DPChallenge*[3]. To evaluate the aesthetic extent of each image, individual votes from crowdsourcing were collected. On a scale from one to ten, each image was annotated with an average of 210 votes. The final score for each image was determined by averaging the scores from each vote.

However, using the single label with majority voting will cause three problems when labeling subjective attributes. Firstly, since the workers in the crowdsourcing platforms are non-professional and have a wide variety of educational backgrounds and levels of experience, the annotations are much noisier than those obtained from specialists [3]. Since there is no assurance that all workers will complete the annotations to a high standard of quality, low-quality votes will decrease the reliability of annotations. Secondly, it is challenging to precisely define the partition of the extent since subjective attributes do not have an underlying ground truth. For example, various workers may assign varying aesthetic scores for images. Last but not the least, due to the subjectivity of the attributes, the boundaries between different scores are not clear, which makes individual voting difficult. Thus, directly using single label with majority voting is not suitable for collecting ground truth for subjective attributes.

2.2.2 Pairwise Comparison

To solve the problems mentioned in 2.2.1, the pairwise comparison method is introduced here for subjective attributes analysis. Annotators are presented with two images and asked which image is better or worse on the properties. The preference decision is a semantically rich method for humans to observe and evaluate objects in the world [45]. The common pipeline to provide a rank list is: (1) Collect abundant pairwise comparisons, and (2) Employ a rating method [46, 47] to obtain the rank.

Jiang *et al.* [1] collected $1,200$ videos from *YouTube* [4]. Ten evaluators participated in pairwise comparison procedures to assess which video is more interesting. The "interestingness" score of each video was determined by averaging the ratings of all annotators. Dubey *et al.* [48] proposed the Place Pulse 2.0 dataset,

[3] `https://www.dpchallenge.com/` (Accessed: 2023/09/01)

[4] `https://www.youtube.com/` (Accessed: 2023/09/01)

which quantified the six perceptual dimensions of the urban environment. The dataset contains $110,988$ *Google Street View*[5] images and provides an evaluation of the safety, liveliness, boredom, wealth, depression, and beauty of 56 cities. They created a crowdsourced online game and collected over 1.16 million times of pairwise comparisons from $81,630$ online volunteers in three years. Based on the comparison results, TrueSkill [46] algorithm was employed to generate a rank score for each image.

Kiapour *et al.* [49] not only created an online game called *Hipster Wars* to collect pairwise comparison results on clothing outfits styles but also evaluated the effectiveness of pairwise comparison. Using *Amazon Mechanical Turk*[1], another single-value vote was also conducted. Each subjective attribute's extent was divided into ten levels. Five annotators were asked to evaluate the extent of a particular style in each image. Compared to the ranking results obtained from pairwise comparison, the average scores obtained from single votes were considerably noisier. The disparity was very large in single-vote comparisons, but consistent in pairwise comparisons. Thus, they concluded that pairwise comparison is suitable for providing reliable and stable ground truth for subtle and subjective attributes. Kiapour *et al.* [49] also employed the TrueSkill algorithm to gather the final rating from all comparisons.

Other than the TrueSkill algorithm, there are many pairwise comparison methods for generating rating scores, such as the Bradley-Terry model [50], Glicko rating system [51], PageRank [52], and so on. Ponomarenko *et al.* [53] collected $53,000$ times of comparisons on image quality and used the Glicko model to estimate the ground truth for each image. Li *et al.* [54] assessed image beauty with the Bradley-Terry model.

Among the various pairwise ranking methods, TrueSkill remains the most popular for primarily three reasons: Firstly, it is fast in computation speed with comparable high accuracy [55, 56]. It can converge significantly more quickly than other algorithms, making it efficient to evaluate large-scale datasets. The second is its flexibility. In its algorithm, the skill of each player is assumed as a Gaussian distribution, which considers the instability and uncertainty of each player. This makes the ground truth obtained from TrueSkill more stable [57]. Finally, it can update scores based on comparison results in real-time [49]. This

[5] `https://www.google.com/maps` (Accessed: 2023/09/21)

allows for continuous monitoring of extent levels, which makes it possible to capture temporal variations and improve the comparison order arrangement. Based on the listed advantages, TrueSkill is selected as the pairwise comparison method for the violence extent analysis presented in Chapter 3.

2.3 Subjective Attribute Recognition

Following the successful construction of a dataset, the next step is to automatically recognize videos. A look back at the evolution of traditional end-to-end video recognition is first given. Then related work regarding feature fusion is introduced.

2.3.1 End-to-End Video Recognition

Recognizing and understanding human interactions in video data is of great importance for human-computer interaction, health care, video surveillance, autonomous driving, and many other real-world applications. Traditional recognition methods first extract hand-crafted features, such as Histograms of Oriented Gradient (HOG) [58], improved Dense Trajectory (iDT) [59], and Space-Time Interest Point (STIP) [60] to represent data. Then a classifier is used to classify the data based on the features. However, such "hand-crafted" features require heavy human labor and domain expert knowledge to be improved [61]. In recent years, due to the increased computational power and availability of large-scale datasets, deep learning has achieved great success, and extracting features with a deep learning method is gaining increasing attention.

Unlike image classification methods which only consider spatial information, video-based recognition should deal with both spatial and temporal information. 3D Convolutional Neural Networks (CNNs) are introduced to recognize human action. A 3D CNN is constructed by applying convolution using a 3D kernel. Ji *et al.* [62] proposed a 3D CNN that includes one hardwired layer, three convolutional layers, two subsampling layers, and one fully connected layer. The input is seven continuous frames with a size of 60×40 pixels each. The hardwired layer can be used to generate information from multiple channels of input frames, ensuring that the following layers can obtain data about gray, gradient, and op-

tical flow. Following their work, Tran *et al.* [63] proposed a Convolutional 3D (C3D) network that does not need any preprocessing. The designed C3D network consists of eight convolutional layers, five max-pooling layers, and two fully connected layers. Compared to low-level hand-crafted features, the features extracted from C3D contain abundant high-level semantic information.

Although end-to-end networks show superior performance in action recognition and other video recognition tasks, there are still some drawbacks. Firstly, the features extracted directly from the last layer may not be the optimal representation for a given task [64]. Particularly for subjective attributes, current networks are generally trained for action recognition and perform poorly on subjective attribute recognition. Secondly, a deeper network will improve the accuracy of recognition, but it is time-consuming and challenging to train. In order to address these two issues, researchers are increasingly developing methods for fusing features from other related networks.

2.3.2 Feature Fusion

As stated previously, supplementary information has been widely implemented in a variety of recognition tasks in order to improve recognition accuracy and the specificity of features. In literature, various types of additional features are utilized, such as information from different convolutional layers, different models, or different tasks.

Simonyan *et al.* [65] integrated deep features from both spatial and temporal CNNs to recognize human actions more accurately. The spatial stream generates information from still images, while the temporal stream generates motion information from optical flow. Song *et al.* [66] fused low-level, middle-level, and high-level features using a dimension-matching function to ensure the dimensions became the same before feature fusion. Jiang *et al.* [67] compressed face hallucination by employing CNN-, GAN-, and RNN-based underlying super-resolvers to generate candidate Super-Resolution (SR) results. Xu *et al.* [68] combined visual, textual, and audio information from video data to identify emotions in social networks. Bakkali *et al.* [69] proposed a two-stream neural architecture that leverages textual contents and visual features to perform the classification of document images. Wu *et al.* [70] claimed that hand-crafted histogram features could

be complementary to CNN features and help increase the person re-identification accuracy.

Based on the superior performance of feature fusion, subjective attribute recognition could also be improved by fusing related information. Thus, feature fusion will be applied in Chapter 4 to enhance the recognition of the subjective social relation atmosphere. More specifically, this is the first attempt to fuse features from the objective task with the subjective task.

2.4 Explainable Artificial Intelligence

Despite the widespread adoption of CNNs, the decision procedure of the network still lacks transparency and interpretability, making it difficult to enhance the performance further. Hence, there has been considerable interest in providing explanation and interpretability for CNNs over the last few years. Explainable Artificial Intelligence (XAI) investigates the relationship between input images or videos and output predictions. Recent studies have achieved outstanding success in explaining 2D image classification CNNs [71, 72]. On the other hand, due to the high computation cost and complexity of video data, the explanation of 3D video recognition CNNs is relatively less studied. In this section, related works about interpretation for 2D CNNs and 3D CNNs are reviewed.

2.4.1 Interpretation for 2D CNNs

Given an input image and a trained 2D CNN, the objective of the explanation method is to quantify the contribution of each element in the input. On the basis of which attribute the explanation model evaluates, there are mainly two types of techniques: input and concept attribution. The input attribution explains the CNN prediction outcomes in terms of the significance of the input image pixels. Concept attribution, on the other hand, identifies the contribution of human-understandable concepts to the predicted class of an image.

Input attribution

The input attribution method is the most commonly used in recent literature. Activation-based methods, such as Class Activation Mapping (CAM) [71], Grad-

CAM [73], Grad-CAM++ [74], and Score-CAM [75], generate weights by utilizing the activations or gradients from intermediate layers of the neural network, then project the feature maps back to the input size in order to produce a heatmap. Perturbation-based methods [76, 77, 78, 79, 80] focus on perturbing the input image pixels using occlusion, mask, or generative algorithms. The importance of each pixel is quantified according to the output changes. Since the semantic meanings of pixels are diverse and highly dependent upon one another, explanation methods based on input attribution may result in contradictory explanations for different data instances in the same class [72].

Concept attribution

To address this issue, recent research employs human-friendly concepts to interpret 2D CNN predictions. The concepts are generated from training data or user-interested data. Kim *et al.* [72] defined a Concept Activation Vector (CAV) to represent every concept. The importance of the concept is evaluated based on the changes in target images toward the direction of the concept. Ghorbani *et al.* [81] defined the concept as superpixel segmentation extracted from input images in order to compute CAVs without human supervision. Based on [72], Goyal *et al.* [82] utilized a conditional Variational AutoEncoder (VAE) model to measure the causal effect of different concepts. Ge *et al.* [83] discussed the structural relationships between concepts with a Graph Neural Network (GNN)-based graph reasoning network, so that both visual and structural clues can be used for explanation.

2.4.2 Interpretation for 3D CNNs

The goal of interpretation for 3D CNNs is to investigate the essential regions in both spatial and temporal dimensions of video data. Only a few methods visualize the prediction process of 3D CNNs. Several methods understand videos using 2D local input attribution techniques initially designed for images by introducing a temporal domain. Srinivasan *et al.* [84] utilized the Layer-wise Relevance Propagation (LRP) [85] to interpret the action recognition based on handcrafted features and Fisher vector. Hartley *et al.* [86] improved the 2D Superpixels Weighted by Average Gradient (SWAG) [87] to the video version by averaging

and smoothing a saliency map at the superpixel level. Li *et al.* [88] introduced a smoothness loss function to smooth the perturbation results in both spatial and temporal dimensions.

However, these methods are only able to provide coarse video regions that lack exact semantic meaning. To my knowledge, no research has yet been proposed on the concept attribution for 3D action recognition CNNs. Hence, the fundamental idea of Chapter 5 is to provide a concept-based high-level interpretation for video understanding. The proposed structure can be utilized to interpret and improve subjective attribute recognition.

Chapter 3

Ground-Truth Generation

3.1 Overview

This chapter focuses on generating a reliable ground truth for the subjective extent attribute. Specifically, the subjective violence extent of videos is analyzed. With the advent of multimedia, social media usage among adolescents has increased dramatically. According to Common Sense census [89, 90, 91], children aged 0 to 8 spend more than 3 hours with screen media every day, and those aged 13 to 18 spend about 9 hours with media, including television, video games, and Internet. However, many videos released on TV or the Internet are unsuitable for children, since they may include violent, bloody, or adult content. These videos may lead to a bad influence on children's behaviors and development. A number of studies has proven that increasing violent media exposure will result in not only short-term but also long-term harmful effects on youth [92]. Eron's [93] and Anderson's [94] experiments indicate that children who watch violent TV programs or play violent video games in early formative age tend to have a high probability to perform aggressive behavior in their later life, including criminal behavior, spousal abuse, and assault [95]. Therefore recognition of violent video is of the essence in multimedia recommendation [96, 97] and multimedia content understanding [98].

Currently, visual violence studies mainly concentrate on scene detection or action recognition, such as explosion, blood, or fight detection. Generally in violence analysis approaches, extracting features from videos is the first step, includ-

ing either local features or global features. Chen and Hauptmann [99] extracted Motion SIFT (MoSIFT) to detect distinctive local features by combining the local appearance and temporal information in surveillance video. De Souza *et al.* [100] presented a violence detector Space-Time Interest Points (STIP) [60] mainly in sports videos. Hassner *et al.* [101] extracted global features Violent Flows descriptor (ViF) for crowd violence detection. Local features are commonly followed by a coding method to represent the video, such as bag-of-visual-words [102]. Finally, these features will be used for classification by using a linear Support Vector Machine (SVM) [103].

Existing visual violence detection research is only limited to objective scene or action detection, instead of subjective video-level content analysis. However, in reality, different scenes or actions may cause different violence extent for a video, which is the key challenge for rating media violence. In this case, this research focuses on subjective violence rating prediction.

However, such a requirement can not be met by existing datasets, none of which provides a violence extent label. To this end, here, a novel dataset is first constructed with violence extent labels. As introduced in 2.2.2, the pairwise comparison method is used to provide ground truth for violence extent.

The major contributions of this dataset construction are as follows:

- Designed for subjective violence rating analysis, a fine-grained violent video dataset called Human Violence Dataset is constructed. The dataset consists of $1,930$ human-involved violent videos collected from *YouTube* [1] movie trailers. Each video is annotated with six objective violent annotations. The subjective violence levels are given by pairwise comparison. The stability and convergence of TrueSkill [46] in the Human Violence Dataset are proven.

- A two-step method is developed for violence rating prediction. A two-stream neural network is fine-tuned on the Human Violence Dataset and used to extract features for each video. By using different pooling and normalization methods, various representations of two-stream features are assessed and the advantage of average pooling is validated. Videos are represented by the best combined two-stream features. Then a rating estimation

[1] `https://www.youtube.com/` (Accessed: 2023/09/01)

machine is proposed to learn the level relationship between different violent videos. The proposed method is experimentally shown to be able to predict violence ratings better than classification methods. The visualization of feature maps and prediction results are also presented.

3.2 Related Work

3.2.1 Violent Video Dataset

In literature, there already exist some datasets for violent video analysis. Hockey fight dataset [104], crowd violence dataset [101], and Violent Scenes Dataset (VSD) 2014 [105], for example, are the most commonly used datasets. Detailed information of these three datasets are introduced below and summarized in Table 3.1.

Hockey Fight dataset

In 2011, Nievas *et al.* [104] created a dataset containing $1,000$ short clips collected from hockey games of the National Hockey League. All the videos were annotated with two categories: "fight" or "non-fight". There are a total of 500 violent videos and 500 non-violent videos. Each video contains 50 frames with a resolution of 720×576 pixels.

Crowd Violence dataset

In 2012, Hassner *et al.* [101] created a dataset for crowd violence detection in video surveillance systems. 246 real-world video clips were collected from *YouTube* [1] with a resolution of 320×240 pixels. The duration of videos is from 1 second to 6 seconds. Half of the dataset are violent, while the others are non-violent.

Violent Scenes dataset

In 2014, Schedl *et al.* [105] produced a dataset for violent scene detection based on Hollywood movies. This dataset is composed of three parts: (1) Hollywood

Table 3.1: Comparison of different violent video datasets.

Dataset	Annotations	Clips	Resource	# of violent videos	Violent level
Hockey Fight [104]	Fight / Non-fight	1,000	Hockey games	500	—
Crowd Violence [101]	Violent / Non-violent	246	YouTube	123	—
VSD 2014 [105]	7 visual + 3 audio labels	31 movies	Hollywood movies	15% violent scenes	—
Proposed	6 subjective + violence	1,930	Promotion videos	1,930	√

training dataset, (2) Hollywood test dataset, and (3) YouTube dataset. The Hollywood dataset contains 24 movies (In total, 50 hours and 2 seconds) and 7 movies (In total, 13 hours and 53 minutes), respectively. The YouTube dataset contains 86 video clips (In total, 2 hours and 3 minutes) from *YouTube* [1]. The Hollywood dataset has around 15 %, and the YouTube dataset has around 44 % violent scenes, respectively. This Violent Scenes dataset has been used in the MediaEval 2014 workshop [106]. Many researches use this dataset for violent scene detection tasks.

3.2.2 Media Rating Systems

In order to protect children and provide appropriate media for different age groups, many countries have established organizations for media ratings, including film, game, and music rating. Here, Eirin in Japan and MPAA in the United States are introduced below.

Eirin

Eirin is the abbreviation for Film Classification and Rating Organization in Japanese, an independent and non-governmental organization. It was established in 1949 and the classification criteria changed several times over the years [107]. Currently, films in Japan are classified into four classes: G (Suitable for all ages), PG12 (Parental guidance requested for teenagers under twelve), R15+ (Restricted to teenagers over fifteen), and R18+ (Restricted to person aged 18 and above).

MPAA

MPAA is the abbreviation for Motion Picture Association of America, who works for film classification and rating in the United States. Its purpose is to provide parents with information about movies so that they can decide whether the movie can be watched by their children or not [108]. Different from Eirin in Japan, MPAA classifies films into five categories: G (General viewing), PG (Parental guidance needed), PG-13 (Some videos are inappropriate for children under 13), R (Restricted, parental guidance required for teenagers under 17), and NC-17 (No one under 17 admitted). Primary factors that may influence the rating include violence, language, theme, drug abuse, sensuality, and nudity [109].

3.2.3 Visual Violence Analysis

Violence detection is not a novel problem, which has mainly been considered as a task to detect flame or blood flow in previous research. Many research focuses on explosion or blood detection. Chen *et al.* [110] proposed a bloody frame detection approach to determine violent scenes in movies. Giannakopoulos *et al.* [111] extracted 12 kinds of audio features to detect audio violence, such as shots and screams. Later studies began to focus on detecting violent interaction behaviors, for instance, fighting actions between people. De Souza *et al.* [100] extracted Space-Time Interest Points (STIP) to distinguish violent activities from regular activities in sports video. Datta *et al.* [112] defined an Acceleration Measure Vector (AMV) to detect human violence in video, such as fist fighting, kicking, or hitting. Improved dense trajectories [59] is also a widely used feature when detecting violence motion [113, 114].

More recently, researchers have paid more attention to deep learning methods when detecting violence. In MediaEval challenges[2], several groups [115, 116, 117] utilized Convolutional Neural Networks (CNNs) to extract features. Li *et al.* [118] combined CNN features, audio features, and motion features to represent violent video. Several researchers also used convolutional Long Short Term Memory (convLSTM) for violence detection. Dong *et al.* [119] and Hanson *et al.* [120] used it to capture spatio-temporal features.

[2] `http://www.multimediaeval.org/` (Accessed: 2023/09/15)

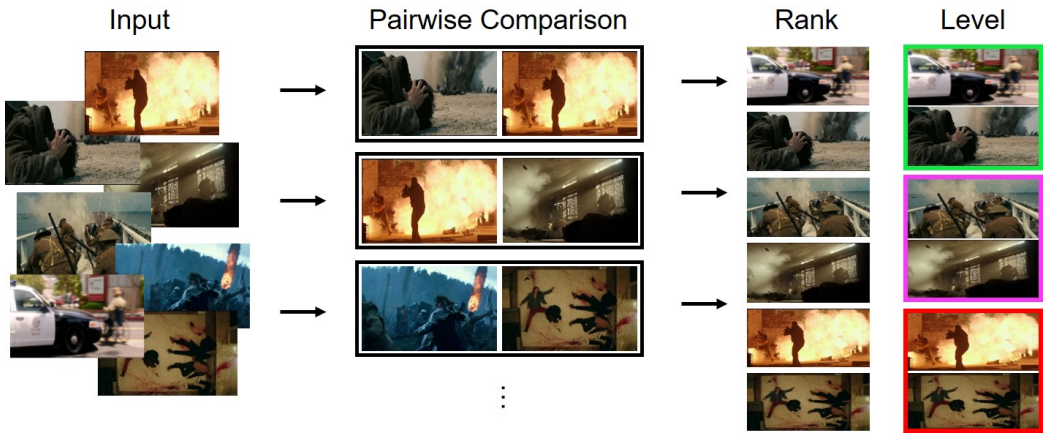


Figure 3.1: Overview of proposed dataset annotation method. With a group of violent videos as input, the videos are ranked according to the violence extent. The final output is the violence level of each video.

Different from previous datasets which contain both violent and non-violent videos, the proposed dataset focuses only on violent videos. Since the number of violent videos in previous datasets is not enough, $1,930$ violent video clips were collected. More specifically, in order to figure out the cause of violence, each video is labeled with six objective attributes. The violence rating of each video is also studied particularly. Table 3.1 also compares the proposed dataset with existing violent video datasets.

Existing media rating systems take both visual and auditory information into consideration, while here, only visual information is focused on. Moreover, this thesis focuses on short movie clips without context or story. In this case, four or five categories are too precise, so three categories are considered sufficient for visual distinguishment. Videos in the dataset are annotated with a violence extent on three levels. Since violence extent is a subjective attribute, a pairwise comparison method is employed to provide ground-truth annotations. Different from previous violent scene detection approaches, exact actions or scenes are not detected in a violent video. As shown in Figure 3.1, by contrast, with a video clip as input, the output of the proposed method is the violence extent of the video. Considering the outstanding performance of CNN in violence detection, a two-stream network is used to extract features from violent videos. Most existing research treat violence recognition as a classification problem, while here, the

relationship between different violence levels are learned.

3.3 Human Violence Dataset

The process of collecting the dataset was carried out by two graduate students with a crosscheck mechanism. The current dataset contains $1,930$ violent video clips collected from movie promotion videos on *YouTube* [1]. Considering the copyright of the video, the collected dataset is only used for research and is not published. Each clip has a length of 30 to 100 frames. The frame rate and resolution of video clips are 30 frames per second (fps) and $1,280 \times 720$ pixels, respectively. Additionally, each clip was manually labeled with six objective fine-grained visual attributes. Furthermore, by utilizing the pairwise comparison method, each video was also annotated with one subjective violence extent label. In the following, the details of the dataset creation process will be described.

3.3.1 Data Collection

In order to collect videos with various violent scenes and actions, movie promotion videos were chosen as raw video content because they contain multiple scenes in short time periods. The collection began by selecting action movies released in the last ten years. Then the movie trailers published on *YouTube* [1] by the corresponding official movie companies were downloaded. In total, $1,020$ raw videos were gathered. As shown in Figure 3.2, the duration of each video is mainly around 2 to 3 minutes. Considering the complexity of violence extent, each video clip should contain only a single scene and a complete action. In this case, a segmentation tool [121] was used to divide the videos into over $25,000$ shots. Further, very short clips that did not contain a whole action, clips with multiple scenes, as well as those clips without human-involved violence were manually removed.

In total, $1,930$ human-involved violent video clips were collected. During the labeling process, some other forms of violence without human involvement, such as firearms, fire, or explosion were also included. However, these objective violence attributes have been well-studied in prior research and are also very consistent in different videos. In contrast, most of the violence is accompanied by human interactions. With different interactions between people, violence extent

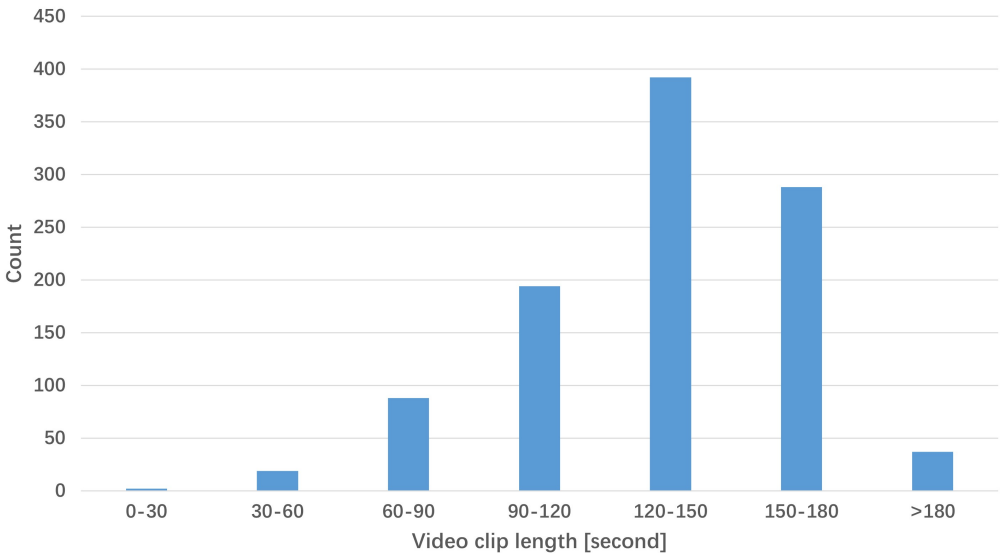


Figure 3.2: Statistics of video clip length of collected movie trailers.

of videos differs from each other. Sometimes even subtle differences may lead to completely different violence levels of videos. Considering these challenges, human-involved visual violence is only focused on in this thesis.

3.3.2 Objective Violent Attribute

Annotations include two parts: Objective violence attributes and Subjective violence ratings. The selected six objective attributes are closely related to the violence extent. The annotation procedure is finished by two annotators independently. The following explains the definition of each attribute:

- **Combat Mode (CM):** There are five subcategories in this attribute: (1) Only attacker appears in the video clip, (2) Only victim appears, (3) One person versus one person, (4) One person versus a group of people, and (5) A group of people versus another group of people.
- **Physical Contact (PC):** There are two subcategories in this attribute: (1) A person brings a part of his/her body into contact with another person, and (2) Others.

Table 3.2: Fine-grained violent attributes and corresponding labels of six video examples in the proposed Human Violence Dataset.

Attributes	Combat Mode	Physical Contact	Weapon Possession	Weapon Direction	Blood	Explosion
Video 1	Attacker	—	√	Other directions	—	—
Video 2	One vs. One	—	√	Opponent	—	—
Video 3	One vs. One	√	√	Opponent	Static	—
Video 4	One vs. Group	√	√	Opponent	—	—
Video 5	Attacker	—	√	Act towards the screen	—	—
Video 6	Victim	—	—	—	—	√

(a) Video 1

(b) Video 2

(c) Video 3

(d) Video 4

(e) Video 5

(f) Video 6

Figure 3.3: Video examples and labels in the proposed Human Violence Dataset. (a) A group of men hold guns, (b) A man attacks another man with a bottle, (c) A man is injured by a knife, (d) A man fights with another two men with a gun on his hand, (e) A woman shoots towards the screen, (f) A group of men are blown up in an explosion. (a) and (b) have the lowest violence level. (c) and (d) have a moderate violence level. (e) and (f) have the highest violence level.

- **Weapon Possession (WP):** There are three subcategories in this attribute: (1) No weapon appears in the video clip, (2) A weapon appears, but is not used, and (3) A weapon is used.
- **Weapon Direction (WD):** A weapon appears in a video, and three subcategories that represent the direction of weapons are annotated: (1) Act on the opponent, (2) Act towards the screen, and (3) Others.
- **Blood:** There are three subcategories in this attribute: (1) No blood, (2) Static blood, and (3) Flowing blood.
- **Explosion:** There are two subcategories in this attribute: (1) No explosion, and (2) Explosion.

Figure 3.3 shows examples of violent videos for each attribute and Table 3.2 shows the corresponding labels in the dataset.

3.3.3 Subjective Violence Rating

As introduced in 2.2.2, pairwise comparison can be used to constrain the instability of subjective attributes rating. Here, the TrueSkill [46] algorithm is employed to annotate the ground-truth violence level. TrueSkill was originally a Bayesian rating system designed for video game matchmaking. When using TrueSkill, for each video, the violence extent will be considered as a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, where μ represents the current estimate of the violence, and σ represents the current uncertainty of the estimate. Every time two videos are compared, the more violent one is decided. After each comparison, μ and σ are updated according to the following equations. Following Herbrich [46], $\mu = 25$ and $\sigma = 25/3$ are set as initial values for each video before any comparison. The following describes the update process:

$$\mu_{\text{win}} \leftarrow \mu_{\text{win}} + \frac{{\sigma^2}_{\text{win}}}{c} \cdot v\left(\frac{(\mu_{\text{win}} - \mu_{\text{lose}})}{c}, \frac{\varepsilon}{c}\right), \tag{3.1}$$

$$\mu_{\text{lose}} \leftarrow \mu_{\text{lose}} + \frac{{\sigma^2}_{\text{lose}}}{c} \cdot v\left(\frac{(\mu_{\text{win}} - \mu_{\text{lose}})}{c}, \frac{\varepsilon}{c}\right), \tag{3.2}$$

$$\sigma^2_{\text{win}} \leftarrow \sigma^2_{\text{win}} \cdot \left[1 - \frac{\sigma^2_{\text{win}}}{c^2} \cdot w\left(\frac{(\mu_{\text{win}} - \mu_{\text{lose}})}{c}, \frac{\varepsilon}{c}\right)\right], \tag{3.3}$$

$$\sigma^2_{\text{lose}} \leftarrow \sigma^2_{\text{lose}} \cdot \left[1 - \frac{\sigma^2_{\text{lose}}}{c^2} \cdot w\left(\frac{(\mu_{\text{win}} - \mu_{lose})}{c}, \frac{\varepsilon}{c}\right)\right], \tag{3.4}$$

$$c^2 = 2\beta^2 + \sigma^2_{\text{win}} + \sigma^2_{\text{lose}}, \tag{3.5}$$

where function $v(\theta) = \mathcal{N}(\theta)/\Phi(\theta)$ and $w(\theta) = v(\theta) \cdot (v(\theta) + \theta)$. They are defined by using the Gaussian Probability Density Function (PDF) $\mathcal{N}(\theta)$ and Cumulative Distribution Function (CDF) $\Phi(\theta)$.

After sufficient comparisons, μ and σ will become stable. According to [29, 48], comparison times around 24 to 36 per video provide a stable ranking. The predicted violent score for each video is calculated as $\mu - 3\sigma$. By sorting violent scores, the violence extent ranking for each video is obtained.

When labeling violence extent, $1,459$ videos with an attribute of WP (2) and (3) subcategories are selected, because videos belonging to these two categories are balanced. Each video is randomly compared to other videos 36 times without any overlap. Graphic User Interface (GUI) was implemented to compare videos efficiently, a snapshot of which is shown in Figure 3.4. Each time it shows two different violent videos randomly and ask the annotator which video is more violent. The observer can only choose one violent video, and it is recorded.

In this experiment, in total, $26,262$ times of different comparisons were collected in about three months by the author to maintain the consistency of the judgment criteria. After all comparisons were performed, TrueSkill violent scores were calculated for each video. Figure 3.5 shows the histogram of violent extent scores after all comparisons. A higher score represents a higher violent extent. These videos are then divided into three levels according to their TrueSkill scores. In practice, the TrueSkill scores averaged over five runs were used as the final score, after proving the stability and convergence of the TrueSkill method.

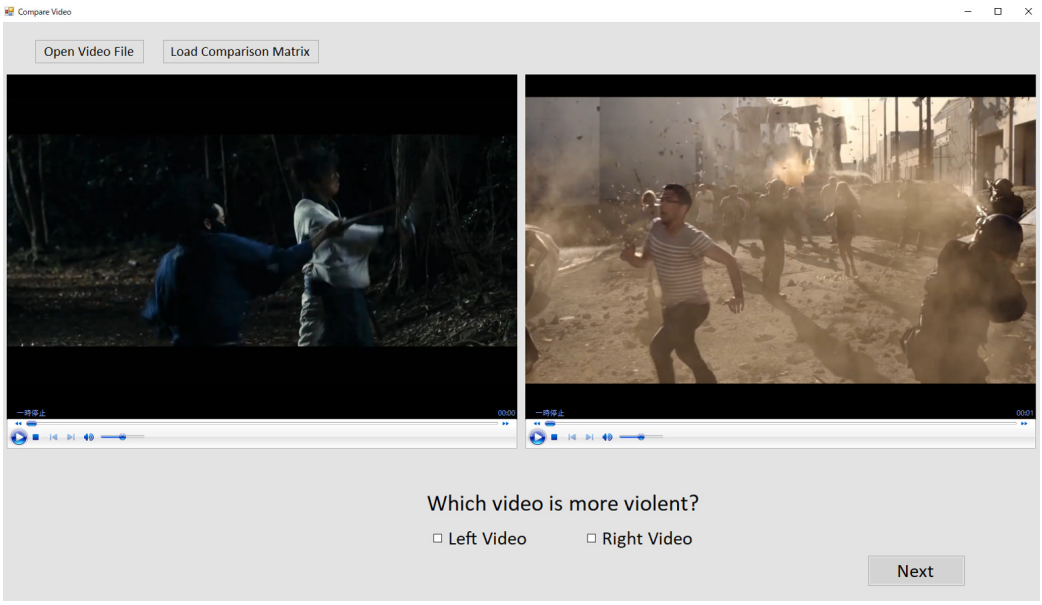


Figure 3.4: Graphical User Interface for pairwise comparison. Two videos are compared on which one is more violent by evaluators.

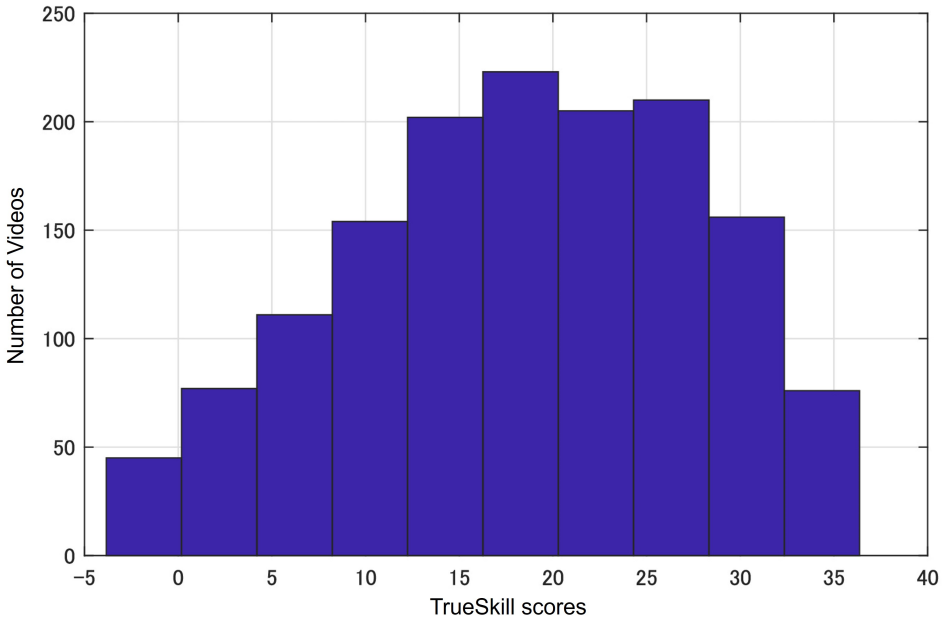


Figure 3.5: Histogram of TrueSkill scores after all comparisons.

Table 3.3: Accuracy of violent level similarity.

Level	Samples	1	2	3	4
1	487	97.33%	96.92%	96.92%	96.92%
2	486	93.83%	93.83%	93.42%	93.83%
3	486	96.50%	96.91%	96.50%	96.91%
All	1,459	95.89%	95.89%	95.61%	95.89%

3.3.4 Dataset Evaluation

The violence level calculated through TrueSkill is used as the ground-truth annotation in the following experiments. In this case, the stability and convergence of the TrueSkill algorithm are proved in the proposed dataset, and below a more convincing violence level annotation is provided. Detailed proofs are shown.

TrueSkill stability in Violent Video dataset

First, the stability of TrueSkill on the proposed dataset is tested. The change of comparison orders is expected not to influence the final violence extent level. In the TrueSkill algorithm, each time only one set of video pairs P_i is compared randomly. Let $L_i = \{L_i^1, L_i^2, ..., L_i^q | q = 1, 2, ..., Q\}$ represent the violence level of each video calculated by P_i, where Q is the video numbers. The sequence of all comparisons can be represented as $S = \{P_1, P_2, ..., P_n | n = 1, 2, ..., N\}$, where N equals to $26,262$, here. The final violence level calculated from sequence S can be represented as $L = \{L^1, L^2, ..., L^q | q = 1, 2, ..., Q\}$. Then the comparison order list S is randomly changed by using a $\mathrm{rand}()$ operation, where the elements in the list will be changed randomly. In this case, a new comparison order list $S_k = \{\mathrm{rand}(P_1, P_2, ..., P_n) | n = 1, 2, ..., N\}$ is obtained. Then, the violence level L_k is calculated for each video according to S_k. By investigating the violence level similarity between L and L_k, we can confirm if the comparison orders will influence the predicted violence extent levels.

The accuracy of violent level similarity when changing comparing orders is shown in Table 3.3. The table includes both the accuracy of three separate levels and the total accuracy. The comparison order is changed four times randomly. From the results, we can conclude that when changing comparison orders, the

violence level will have around 96% similarity between different sequences. It indicates the violence rating calculated by TrueSkill is nearly independent of comparison order.

TrueSkill convergence in Violent Video dataset

Next, the convergence of TrueSkill in the proposed dataset is tested. At present, each video is compared 36 times with other videos. Let each video be compared to other videos t times, the violence level calculated after t times per video can be represented as $L_t = \{L_t^1, L_t^2, ..., L_t^q | q = 1, 2, ..., Q\}$, where Q is the video numbers. The violence level of each video is expected to become stable as the number of comparison times increases, which means the violence levels obtained from two consecutive comparison results should be highly similar after sufficient comparisons. The convergence score of TrueSkill is the similarity between the predicted violence level in L_t and L_{t+1}, which can be defined as:

$$\text{Score}_t = \frac{\left|q \in Q : L_t^q = L_{t+1}^q\right|}{Q} \in [0, 1]. \tag{3.6}$$

Based on previous research [29, 48], the TrueSkill score converged after around $t = 30$ times comparisons for each video. To better observe the trend from non-convergence to convergence, t is set from 20 to 35 here. Figure 3.6 shows the results of the convergence score. After about 28 times comparisons per video, the violence level for each video became stable.

According to previous proof, changing comparison orders will cause a disturbance to the final rating. In order to obtain a more convincing violence level, the procedure is repeated five times and TrueSkill scores are calculated for each time. The average value of five scores is treated as the final score for each video. Figure 3.6 also shows the average convergence scores after five times. By averaging multiple results, the violence level is more accurate and has a higher convergence rate. In this case, the violence level obtained from five times average is selected as the final ground-truth violence rating for the following experiments. Level 1 contains 487 videos with TrueSkill scores ranging from -4.7693 to 14.9794. Level 2 contains 486 videos with scores from 14.9841 to 23.9908. Level 3 contains 486 videos with scores from 24.0114 to 37.5025.

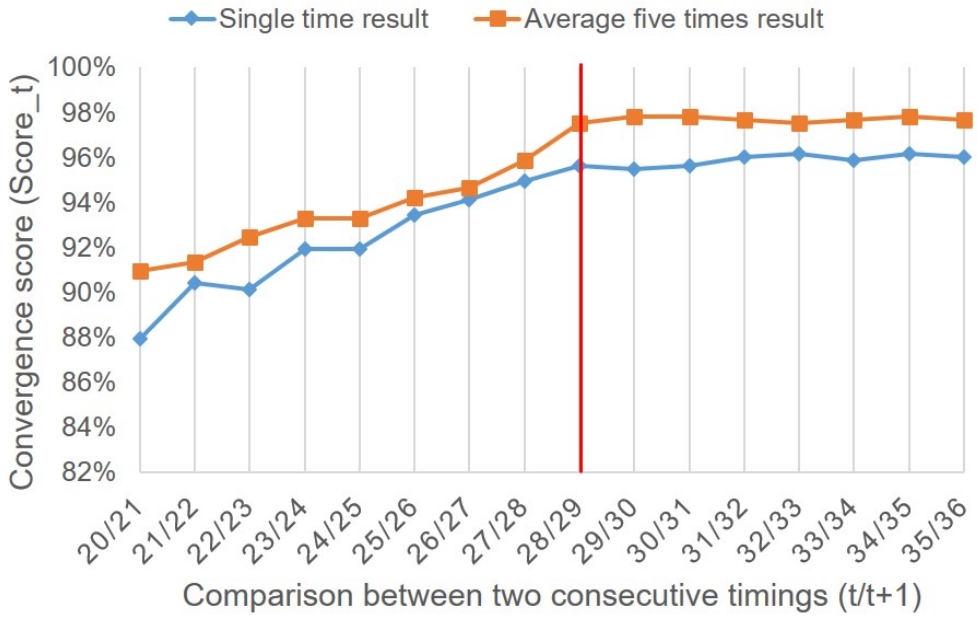


Figure 3.6: Convergence of TrueSkill score. The similarity between consecutive timings are calculated.

Figure 3.3 also shows six video examples from three violence levels.

3.4 Proposed Method

This section proposes a violence rating prediction method designed for recognizing subjective extent. Figure 3.7 illustrates the pipeline of the proposed method. The input is violent video clips and the output is violence levels. The method is composed of two main steps: (1) Fine-tuned two-stream network used to extract features for each video, and (2) Rank learning machine trained to predict the violence rating for a test video.

3.4.1 Two-Stream Network Based Feature Extraction

Two-stream network [65] has shown great success in many computer vision tasks, especially in action recognition. It consists of a spatial stream and a temporal stream. The input of the spatial stream is a single image frame, while the input of

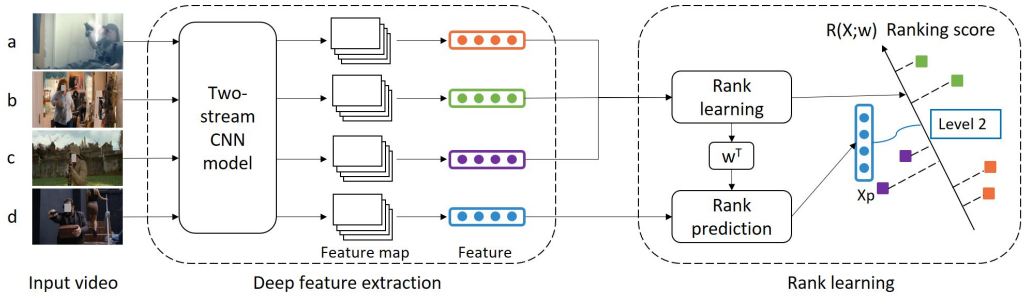


Figure 3.7: Pipeline of the proposed method. Videos a, b, and c are training videos used to train the rank learning machine. Three different colors are used to represent two-stream features extracted from videos in different violence levels. Video d is a test video. The violence level of the test video can be predicted by calculating its ranking score by using the trained rank learning machine.

the temporal stream is a stack of 10 optical-flow frames. A two-stream network is fine-tuned using the Human Violence dataset introduced in Section 3.3, and is used to extract features. Given a raw video i of size $H \times W \times 3 \times F$, where H [pixels] represents the height of the video, W [pixels] represents the width of the video, and F represents the number of frames. Each raw video is down-sampled to be used as the input of the two-stream networks. The number of the down-sampled video frames is set as N selected from the whole number of frames F, making the input sizes of RGB frame and optical flow become $H \times W \times 3 \times N$ and $H \times W \times 10 \times N$, respectively. Then, two-stream features f_s and f_t are extracted from the Rectified Linear Unit (ReLU) layer following the seven Fully Connected (FC) layers. Both f_s and f_t have the size of $H' \times W' \times C \times N$, where C is the channel numbers of the extracted layers, H' is the height of the feature map, and W' is the width. These parameters depend on the video size and the network.

A max pooling or a sum pooling operation is then performed over f_s and f_t on C-dimension. The feature can be further denoted as f'_s and f'_t. Then the two-stream features are normalized and concatenated as the final representation for each video denoted as $\boldsymbol{f} = [B(f'_s); B(f'_t)]$. Here, B denotes two normalization methods: L2 normalization and Square Root (SR) with L2 normalization. This concatenated feature $\boldsymbol{f}$ will be used to predict the level in the following step.

Right now, video i can be represented as $(\boldsymbol{f}_i, l_i)$, where $\boldsymbol{f}_i$ is the two-stream feature and l_i is the ground-truth violence rating. Here, videos are divided into

three levels: L_1, L_2, and L_3, where $L_1 < L_2 < L_3$. The purpose of this research is to predict the violence level l of a test video with feature $\boldsymbol{f}^*$. In previous research, multi-class classification was the most commonly used method. However, classification methods take each sample separately and can not distinguish the difference between different violent levels, such as: $L_1 < L_2$, $L_2 < L_3$, $L_1 < L_3$. In order to make the best of these inner relationships, a rank machine is introduced here to learn the relationship between different violent levels and predict the violent rank.

In the learning stage, let's denote the training dataset which contains V videos as $\boldsymbol{D} = \{(\boldsymbol{f}_1, l_1), (\boldsymbol{f}_2, l_2), ..., (\boldsymbol{f}_v, l_v) | v = 1, 2, .., V\}$. For every two different videos, there are two kinds of relationship according to their violence labels: ordered relationship and similar relationship. Ordered relationship is defined as $O = (\boldsymbol{f}_i, \boldsymbol{f}_j)$, if $l_i > l_j$, which means video i have a higher violence level than video j. Similar relationship is defined as $S = (\boldsymbol{f}_i, \boldsymbol{f}_j)$, if $l_i = l_j$, which means video i and video j are in the same violence level. The purpose here is to learn a ranking function:

$$r(\boldsymbol{f}_i) = \boldsymbol{w}^\intercal \boldsymbol{f}_i, \tag{3.7}$$

where $\boldsymbol{w}^\intercal$ is a coefficient vector. This ranking function should make the maximum number of the following constraints satisfied:

$$\forall (i, j) \in O : \boldsymbol{w}^\intercal \boldsymbol{f}_i > \boldsymbol{w}^\intercal \boldsymbol{f}_j, \tag{3.8}$$

$$\forall (i, j) \in S : \boldsymbol{w}^\intercal \boldsymbol{f}_i = \boldsymbol{w}^\intercal \boldsymbol{f}_j. \tag{3.9}$$

Solving this problem is an NP hard problem. Following the work done by Parikh [45] and Joachims [122], two non-negative slack variables ξ and γ which are similar to SVM are introduced to approximate the results. Equation 3.8 and Equation 3.9 are converted into solving the following optimization problem:

$$\text{minimize} : \left(\frac{1}{2}\|\boldsymbol{w}^\intercal\|^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right) \tag{3.10}$$

$$\text{s.t.} \quad \boldsymbol{w}^\intercal \boldsymbol{f}_i \geq \boldsymbol{w}^\intercal \boldsymbol{f}_j + 1 - \xi_{ij}; \forall (i,j) \in O, \tag{3.11}$$

$$|\boldsymbol{w}^\intercal \boldsymbol{f}_i - \boldsymbol{w}^\intercal \boldsymbol{f}_j| \leq \gamma_{ij}; \forall (i,j) \in S, \tag{3.12}$$

$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0, \tag{3.13}$$

where C is a trade-off constant to maintain the balance between maximizing the margin and meeting the margins of the pairwise label. The Newton method will be used to calculate $\boldsymbol{w}^\intercal$.

3.4.2 Violence Rating Prediction

The ranking function $r(\boldsymbol{f}_i)$ with learned $\boldsymbol{w}^\intercal$ can be utilized to calculate the violence score for each video. Dataset $\boldsymbol{D} = \{(\boldsymbol{f}_1, l_1), (\boldsymbol{f}_2, l_2), ..., (\boldsymbol{f}_v, l_v) | v = 1, 2, ...V\}$ can be represented as $\boldsymbol{D}' = \{(\boldsymbol{w}^\intercal \boldsymbol{f}_1, l_1), (\boldsymbol{w}^\intercal \boldsymbol{f}_2, l_2), ..., (\boldsymbol{w}^\intercal \boldsymbol{f}_v, l_v) | v = 1, 2, ..., V\}$. In the following, three different methods are introduced to predict the violence level with a violence score.

Minimum distance prediction

For each violence level, the ranking score with feature $\boldsymbol{f}$ can be represented as:

$$S_k = \frac{1}{N_k} \sum_{l_v = L_k} \boldsymbol{w}^\intercal \boldsymbol{f}_v \; (k \in \{1, 2, 3\}, v \in \{1, 2, ..., V\}) \tag{3.14}$$

where N_k represents the number of violent videos in level L_k, k denotes the corresponding violence level in the task. The violence level for a new video with feature $\boldsymbol{f}^*$ can be calculated as:

$$L^* = \text{argmin}_k (\boldsymbol{w}^\intercal \boldsymbol{f}^* - S_k)^2 (k \in \{1, 2, 3\}). \tag{3.15}$$

Minimum mean distance prediction

The violence scores for violent videos in each level can be assumed as a Gaussian distribution:

$$F_k(\boldsymbol{w}^\intercal \boldsymbol{f}_k) = \mathcal{N}(\mu_k, \sigma_k) \ \ (k \in \{1, 2, 3\}), \tag{3.16}$$

where μ_k is the mean value of the Gaussian distribution, and σ_k is the standard deviation. Given a new video with feature $\boldsymbol{f}^*$, the violence level can be calculated as:

$$L^* = \text{argmin}_k(\boldsymbol{w}^\intercal \boldsymbol{f}^* - \mu_k)^2 (k \in \{1, 2, 3\}). \tag{3.17}$$

Maximum Gaussian likelihood prediction

The ranking scores in each level also follow a Gaussian distribution $\mathcal{N}(\mu_k, \sigma_k)$. The rating level of a new video can be predicted by computing the maximum likelihood of the rating scores, which can be represented as:

$$L^* = \text{argmax}_k \, P(\boldsymbol{w}^\intercal \boldsymbol{f}^* | \mu_k, \sigma_k) (k \in \{1, 2, 3\}). \tag{3.18}$$

3.5 Experiments

In order to evaluate the effectiveness of the proposed method, experiments implemented in MATLAB are conducted. In total, $1,459$ videos are used with 75% as training data, and the rest as test data. Three networks are used as the backbone Convolutional Neural Network (CNN) and compared: Alexnet [11], VGG16 [123], and ResNet-50 [124]. The structures of these three networks are shown in Figure 3.8, Figure 3.9, and Figure 3.10, respectively. These networks are pre-trained on ImageNet [41] followed by fine-tuning on the proposed dataset. Alexnet has 5 convolutional layers and 3 FC layers, followed by ReLU. VGG16 has 13 convolutional layers and 3 FC layers. Max-pooling is performed over a 2×2 pixels window with stride 2. All hidden layers use ReLU as an activation function. ResNet-50 is a 50 layers residual network. Each stream is trained separately. Softmax scores of two streams are combined by averaging fusion [65].

All images are resized to 256×342 pixels beforehand. When implementing VGG16 and ResNet-50, a 224×224 pixels sub-image is cropped from the selected

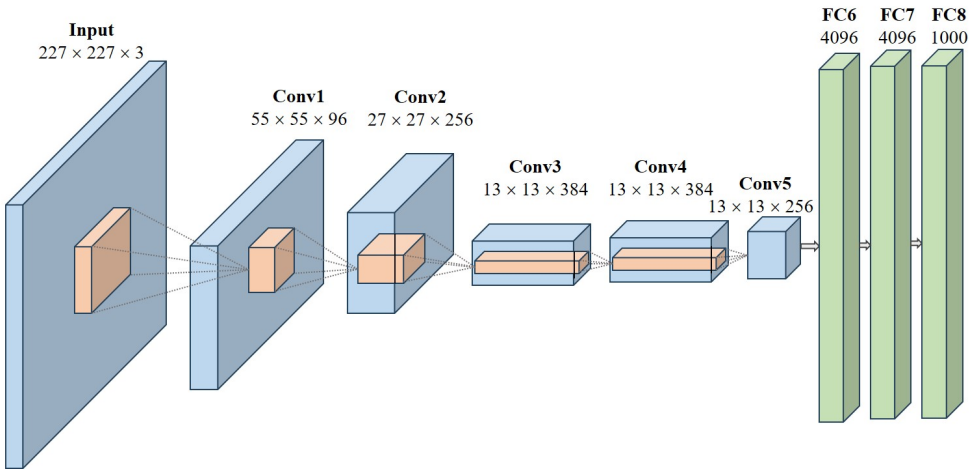


Figure 3.8: Network structure of Alexnet.

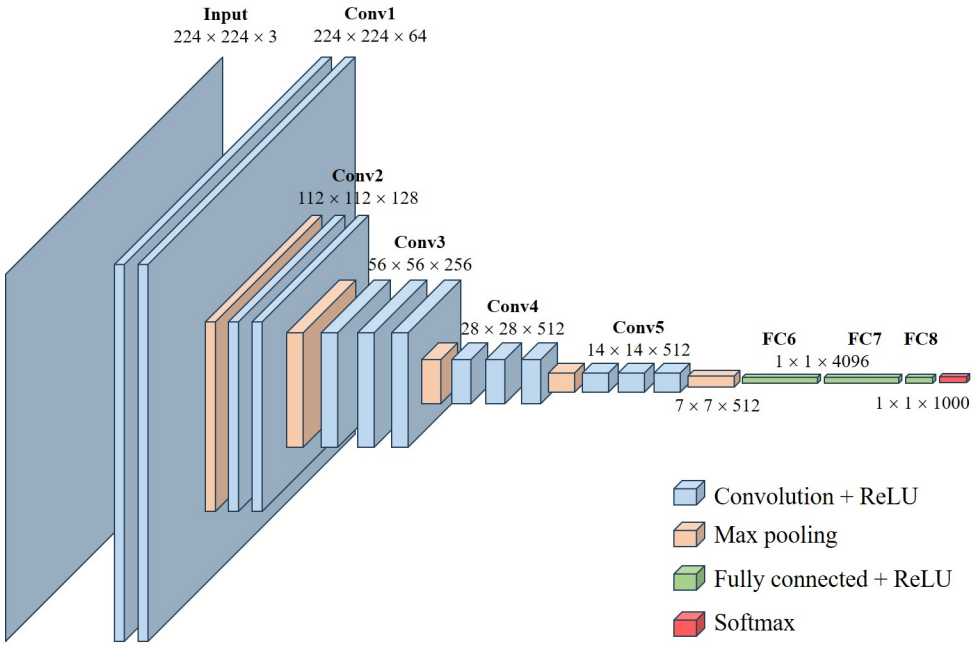


Figure 3.9: Network structure of VGG16.

image randomly. When using Alexnet, a 227×227 pixels sub-image is cropped. In the training stage, for the spatial network, one frame is randomly chosen from each video and resized to the required size. For the temporal network, 10 continuous optical-flow frames are randomly chosen from each video. In the test stage, for the spatial network, the middle frame of the video is used. For the temporal network,

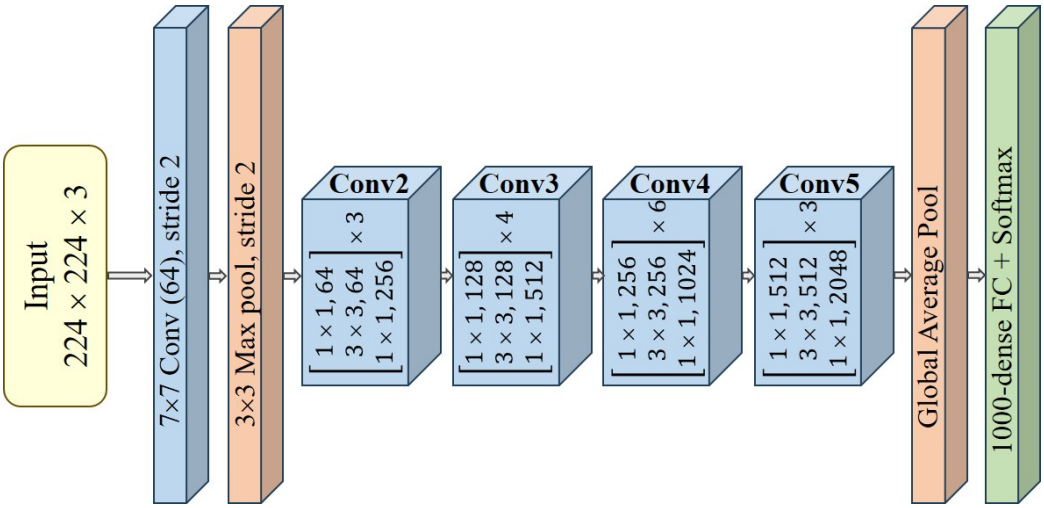


Figure 3.10: Network structure of ResNet-50.

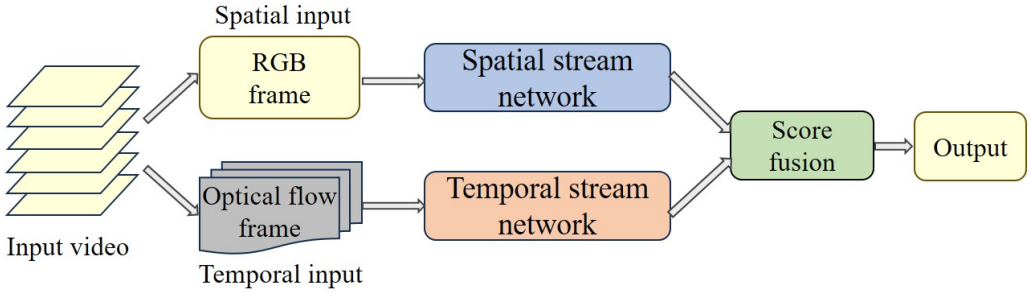


Figure 3.11: Network structure of Two-Stream Convolutional Network.

the middle 10 continuous optical-flow frames are used. The learning rate for two streams starts from 10^{-3}, and is reduced by a factor of 10 every 30 epochs until the 90th epoch. When using Alexnet and VGG16, a dropout layer is added after FC layer for two streams. According to the good practices in [125, 126], the dropout ratio for the spatial network is set as 0.8, and the temporal network 0.9.

3.5.1 Evaluation of Two-Stream Network

In general, deep learning methods perform better than traditional methods, and with the increase of the network depth, the accuracy becomes higher. First, two experiments are conducted: (1) Improved Dense Trajectories (IDT) [59] is used to extract trajectory features, and (2) Two-stream network is used for violence

Table 3.4: Evaluation of classification methods and deep features.

Method	End-to-End		Feature			
			Pooling	Raw	L2-norm	SR + L2-norm
IDT	49.17%		—			
Alexnet	Spatial Network	39.84%	Average	39.29%	39.84%	**40.93%**
			Max	40.11%	**41.21%**	38.46%
	Temporal Network	41.75%	Average	40.48%	41.23%	**42.03%**
			Max	42.31%	**44.78%**	42.03%
	Two-stream Network	**46.40%**	Average	44.23%	46.70%	**46.70%**
			Max	45.33%	45.33%	46.15%
VGG16	Spatial Network	42.86%	Average	45.60%	**47.80%**	46.98%
			Max	45.05%	45.88%	**46.70%**
	Temporal Network	46.43%	Average	47.53%	**49.18%**	47.53%
			Max	42.31%	47.80%	**48.90%**
	Two-stream Network	**50.28%**	Average	47.25%	**51.65%**	51.10%
			Max	49.18%	51.10%	50.27%
ResNet-50	Spatial Network	44.23%	Average	41.48%	43.96%	**48.63%**
			Max	42.03%	46.70%	**48.08%**
	Temporal Network	48.90%	Average	46.70%	47.53%	**50.27%**
			Max	49.18%	**49.73%**	49.45%
	Two-stream Network	**50.82%**	Average	49.73%	49.73%	**53.02%**
			Max	49.18%	50.82%	48.90%

classification. The structures of two-stream convolutional network is shown in Figure 3.11.

In IDT, four different descriptors are computed: Histogram of Oriented Gradient (HOG) [58], Histogram of Flow (HOF) [127], Motion Boundary Histogram (MBH) [128], and trajectory. Here, trajectory length is set as 3 frames, and in total 402 dimensional features are calculated for each video. The dimensions are decreased to 201 by using Principal Component Analysis (PCA), and fisher vector is used to encode the features. Finally, a linear SVM is used for classification. The accuracy of IDT was 49.17% as shown in Table 3.4.

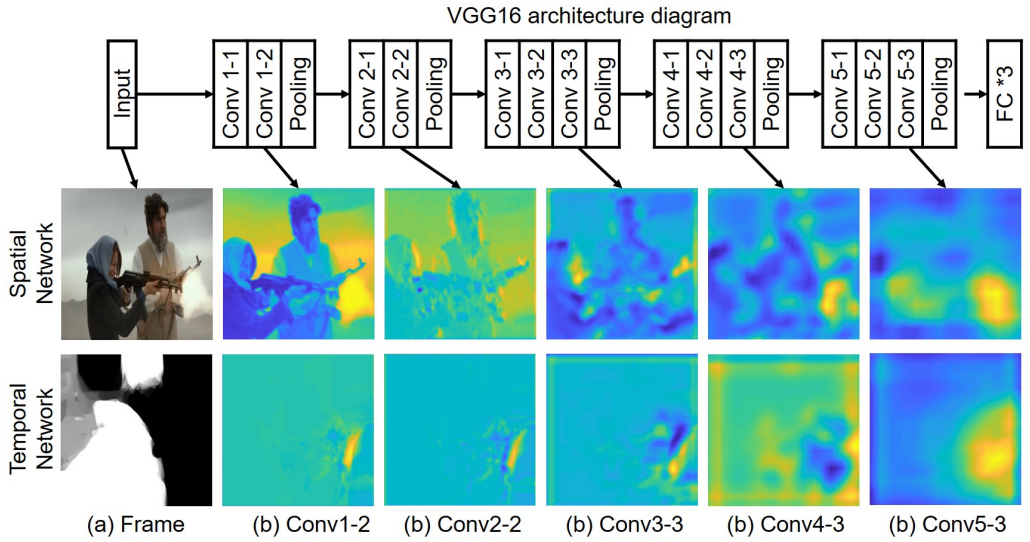


Figure 3.12: Visualization of feature maps from different layers using VGG16. (a) Original violent frame and optical-flow frame, (b)–(f) Feature maps extracted from Conv1-2, Conv2-2, Conv3-3, Conv4-3, and Conv5-3 layers. We can see that deeper convolution layers can provide more typical features.

The end-to-end results of the two-stream network are also included in Table 3.4. A deeper network can better predict violence rating. ResNet-50 showed the best performance. IDT performed worse than VGG16 and ResNet-50. A two-stream network can learn from both image information and motion information. The result validates that it can provide more violent representations than trajectory features. Figure 3.12 visualizes the feature maps of two streams using VGG16. For each convolutional layer, the mean value of each channel is calculated and the most active channel is visualized with a pseudocolor image. We can see that the shallower layers mainly provide edge or texture patterns, while the deeper layers can provide more discriminative information such as a gun shot.

3.5.2 Evaluation of Deep Features

CNN representation has been proven to be a powerful descriptor in previous research [129, 130]. As mentioned above, deeper layers can produce more characteristic features than shallower ones. The activations extracted from the FC layer with a linear SVM usually show outstanding performance [131]. Normally the ex-

tracted features are normalized to small scale such as 0 to 1, in order to maintain the balance and prevent numerical difficulties before classification.

Here, a fine-tuned two-stream network is used as a feature extractor and features are extracted from the ReLU7 layer. As mentioned in 3.4.1, N is set as the number of frames for each video. Two pooling operations and two normalization operations are conducted on the extracted features. Since violent activity is a continuous action, violent extent judgment depends on the information from all frames in a single video. Average pooling will better utilize all violence features than max pooling in this case. Finally, violence level is predicted by feeding the features into a linear SVM.

Table 3.4 shows the evaluation results of different pooling and normalization methods for the spatial network and the temporal network. For each pooling method, the best-performed normalization method in each single stream is concatenated as two-stream features, and fed into a linear SVM. Table 3.4 also evaluates the performance of two-stream features. The results prove that average pooling can retain more violence information and can predict better in all three networks, while normalization methods do not make much difference. Furthermore, the combined two-stream features perform better than end-to-end two-stream results.

3.5.3 Evaluation of Proposed Method

Now let us evaluate the proposed method. The best performed two-stream features in Table 3.4 are used to train the rank learning machine proposed in 3.4.1. Three different methods are conducted to predict violence rating. The comparison results between the proposed method and existing methods are summarized in Table 3.5. We can see that predicting violence rating by calculating maximum Gaussian likelihood performs best. Using mean distance prediction also performs better than classification methods. This is because the ground-truth violence scores are predicted by TrueSkill [46] and estimated as Gaussian distribution. So Gaussian distribution can better match the relationship of rating difference. However, the minimum distance prediction method performs worse. This can be hypothesized that because learning rank relationship is complicated, directly using the rank score difference can not reflect the inner relationship.

Table 3.5: Evaluation of the proposed method.

Methods	Alexnet	VGG16	ResNet-50
Chance rate	33.33%	33.33%	33.33%
Two-stream End-to-end	46.40%	50.28%	50.82%
Two-stream feature + SVM	46.70%	51.65%	53.02%
Minimum distance prediction	40.38%	45.60%	49.45%
Mean distance prediction	49.18%	53.37%	57.69%
Maximum Gaussian likelihood	**51.10%**	**53.85%**	**57.97%**

Table 3.6: Confusion matrix of VGG16. Horizontal axis is the predicted label. Vertical axis is the true label. (a) Confusion matrix of end-to-end two-stream network, (b) Confusion matrix of the best performed two stream features with an SVM classifier, (c) Confusion matrix of the proposed method using the maximum likelihood estimation method.

End-to-end	Level 1	Level 2	Level 3
Level 1	0.4672	0.459	0.0738
Level 2	0.314	0.4793	0.2066
Level 3	0.124	0.314	0.562

(a) End-to-end

SVM	Level 1	Level 2	Level 3
Level 1	0.6311	0.2459	0.123
Level 2	0.4463	0.3388	0.2149
Level 3	0.1488	0.2727	0.5785

(b) SVM

Proposed	Level 1	Level 2	Level 3
Level 1	0.6639	0.2623	0.0738
Level 2	0.438	0.3471	0.2149
Level 3	0.1405	0.2562	0.6033

(c) Proposed method

Table 3.6 shows three confusion matrices of using VGG16 network. The matrices prove that the proposed method can better predict violence rating in all three levels. Especially, the prediction accuracy in level 1 and level 3 are improved a lot. However, the videos in level 2 have the lowest prediction accuracy in all methods. Videos in level 1 have a high possibility to be predicted as level 2, while videos in level 3 are more likely to be judged as level 2 than level 1. This can be hypothesized that videos in level 1 and level 3 usually have a strong evidence, while videos in level 2 do not have a clear boundary with the nearby two levels.

Figure 3.13 shows some prediction examples using the proposed method. In each level, videos in the box of the same color with the ground-truth level color are the successful examples. However, there are still some failures. For example, the first failure example in level 1 is a man hitting the head of a woman with a kettle. It is a low violence video according to its TrueSkill score, while it is considered as level 2. This could be because the proposed method does not detect exact objects in the video, so it may fuse the kettle with some other aggressive weapons,

Figure 3.13: Examples of violence rating estimation results using the proposed method. The border color of each image indicates the estimation result. Three different colors are used to represent the predicted violence level: Green represents level 1, purple represents level 2, red represents level 3. The TrueSkill score for each video is labelled at the bottom right of each image. The left four videos in each level are the successful examples, the right two videos in each level are failure examples.

leading the attacking behavior to become a high violence action. By observing the remaining failure examples and their TrueSkill scores, we can see that the boundaries between videos that have close scores are very unclear. Sometimes the actions are very similar, and only the direction of the weapon or the frequency of using the weapon has a slight difference. So it is hard to predict the level of videos near the border. These examples are consistent with the above confusion matrices.

3.6 Summary

This chapter focused on providing reliable ground truth for the subjective violence attribute. A new dataset was constructed. The proposed Human Violence Dataset consists of $1,930$ violent videos. Besides, six objective violence attributes and one subjective violence rating level were annotated to each video. Based on TrueSkill [46] pairwise comparison algorithm, the violence extent can be measured. The ground truth was evaluated in both stability and convergence.

Designed for this dataset, a rank learning-based method was proposed to estimate the violence extent of each video. It is mainly composed of two parts: (1) Visual features are extracted using a fine-tuned two-stream network. Spatial features and temporal features with different pooling and normalization methods are investigated, and (2) Violence rating prediction machine is learned by utilizing deep features and pairwise relationship between videos.

Experimental results on violence rating prediction showed that the proposed method performed better than existing classification methods. This indicates that the proposed method better reflects the relationship between different violence levels and can produce more representations for violent videos.

The main advantage of the proposed method over the previous studies is that it focuses on video-level analysis instead of single scene-level detection. On the other hand, the subjective violence rating label is mapped to a ranking problem with pairwise comparison. The ground-truth violence rating provided by multi-time comparisons is more reliable than a single-score evaluation. Finally, the proposed method is suitable for predicting subjective attributes which have an inner extent relationship.

Chapter 4

Subjective Attribute Recognition

4.1 Overview

In this chapter, a detailed task to investigate subjective attribute recognition; social relation atmosphere recognition is introduced. Social relation atmosphere plays an important role in the human society. It refers to the integration of multiple cues in social relationships, such as communication states, interpersonal interactions, emotional expressions, and body language of humans. By automatically recognizing the social relation atmosphere in social media, machines can interpret human behaviors more precisely, provide more semantic information, and develop more intelligent applications.

In recent years, some research focuses on recognizing objective social relationships from images and videos. Existing techniques on image-based data primarily utilize multiple visual cues. Sun *et al.* [132] divided social life into 5 domains and 16 social relationships based on Burgental's social psychology theory [133]. Age, gender, and head position were employed as intermediate attributes to aid in the prediction of both social domains and relationships. Similar to this, Goel *et al.* [134] believes that intermediate attributes are helpful for understanding social relationships. They first generated a Multi-Network Convolutional Neural Network (MN-CNN) for extracting body attributes and human activity representation. The relationship is then depicted using a structured graph. More recently, rather than directly concatenating multi-source attributes, Wang *et al.* [135] proposed a Deep Supervised Feature Selection (DSFS) framework to abstract the

Figure 4.1: Examples of different social relationships and social relation atmospheres. The underlined labels identify social relationships from the ViSR dataset. The other labels indicate the social relation atmosphere in the proposed annotations. The definition of social relation atmosphere between individuals is closely related to social relationships and other social clues.

attributes as a subset of discriminative features.

Compared to image-based analysis, a video-based scenario has received less attention. Considering the vast amount of information in video, research on video data focuses more on extracting spatial-temporal information. Liu *et al.* [136] proposed a Multi-scale Spatial-Temporal Reasoning (MSTR) framework to discover both local and global representations from the spatial-temporal domain. Additionally, audio information is an essential supplementation to video. Lv *et al.* [137] introduced a multi-stream model that integrated spatial, temporal, and audio characteristics for identifying social relationships in videos.

Although objective social relationship recognition has been widely studied, none of the existing research pays attention to the subjective social relation atmosphere analysis. However, this faces some obstacles. First, there is no well-annotated dataset. Existing datasets only consider social relationships. In reality, people with different relationships have distinct social relation atmospheres, which is a significant challenge. Second, because video data contain redundancy and noise, a person can appear in any spatial-temporal location and from various

angles, making it difficult to track the essential semantic information from video.

To this end, this chapter focuses on social relation atmosphere analysis and extracting essential visual information from video. As illustrated in Figure 4.1, the social relation atmosphere among the same relationships may differ, whereas the same atmosphere may be produced by different relationships. In general, negative relationships tend to create a negative atmosphere, and positive relationships tend to produce a positive atmosphere. Social relationships and social relation atmosphere are closely related because both attributes take into account human interactions, facial expressions, and the surrounding environment. Thus, social relationships could provide additional information regarding the social relation atmosphere. Here, first, annotations for the social relation atmosphere is introduced. Four kinds of labels are prepared and new labels are annotated for each video in addition to the existing Video Social Relation (ViSR) Dataset [136]. Secondly, a method for identifying the subjective social relation atmospheres is proposed based on the Relevant Visual Concepts (RVC) from the objective social relationship task. Specifically, to avoid video redundancies, an explainable module proposed in Chapter 5 is employed for extracting the most crucial spatial-temporal visual information from videos. Finally, the proposed method is evaluated using popular 3D CNNs and the results are visualized.

To summarize, the contributions of this chapter include:

- Four subjective social relation atmosphere attributes are annotated on the ViSR dataset. To the best of my knowledge, this is the first attempt to understand the social relation atmosphere in videos.

- Relationship between subjective and objective attributes is investigated. The most significant video volumes from the social relationship recognition task can be used to supplement social relation atmosphere recognition.

- Effectiveness of the proposed framework is evaluated on three 3D CNNs [63, 138, 139]. Both quantitative and qualitative results indicate that the proposed method outperforms the end-to-end 3D CNNs.

4.2 Related Work

4.2.1 Social Understanding

Computer scientists and psychologists have paid considerable attention to the study of social understanding for many years. Popular research topics include gaze detection [140, 141, 142], facial expression recognition [143, 144], group activity recognition [145, 146, 147], and so on. In recent years, there has been an explosion of interest in recognizing social relationships through multimedia. Recognition in still images has achieved remarkable success, whereas recognition in video is still limited. In contrast to image-based methods that heavily rely on various semantic attributes, it is challenging to extract specific clues from the video data, such as actions or body part positions. In this case, existing video-based approaches prioritize spatial-temporal information more.

Lv *et al.* [137] introduced a multi-stream model. Initially, a spatial CNN is used to discover video scenes and people representations from images. In the second step, a temporal segment CNN [125] is used to extract action features. Finally, GoogleNet [41] is applied to learn audio features using the audio spectrum. Each network is trained independently. In a late fusion approach, the prediction score from each stream is combined to generate the final decision for social relationships. However, they only considered global and local features, disregarding the relationships between characters, objects, and scenes in the video.

Liu *et al.* [136] proposed a Multi-scale Spatial-Temporal Reasoning (MSTR) framework to integrate the spatial and temporal features of a person or an object from videos. Using a Mask R-CNN [148], the people and objects in a video clip are first cropped. To investigate the interaction between various individuals and objects, triple graphs are constructed. An Inter-Person Graph is constructed to capture the interactions from different individuals. An Intra-Person Graph is designed for the same individual, and a Person-Object Graph is designed for the co-existence of persons and objects. In addition, 3D CNN is utilized to extract global features from the original video frames. The social relationship in a video is predicted by combining features from multiple graphs.

Most of these works deal with objective social relationships, and none consider the social relation atmosphere between people, which is vital for future human

interaction applications. Therefore, the subjective social relation atmosphere is investigated in this chapter.

4.2.2 Video-Based Social Relationship Dataset

Existing social relationship recognition datasets are predominantly image-based, while video-based datasets are uncommon. The most popular video datasets are the Social Relation in Videos (SRIV) dataset [137], MovieGraphs dataset [149], and Video Social Relation (ViSR) dataset [136]. Detailed information of these three datasets are introduced below.

Social Relation in Videos dataset

Social Relation in Videos (SRIV) dataset was created in 2018 by Lv *et al.* [137] and contains 3,124 videos. The videos were collected from 69 TV dramas and movies. Each video clip has a duration between 5 and 50 seconds. The dataset is labeled with eight subjective relations according to [150] and eight objective relations based on the study in [151]. In this dataset, 16 subclasses are defined in total.

MovieGraphs dataset

Vicol *et al.* [149] proposed the MovieGraphs dataset in 2018. The dataset consists of $7,637$ clips collected from 51 movies with abundant annotations. Each clip is labeled with four components: (1) Graph representing characters' detailed attributes, interactions, relationships, and time stamps, (2) Situation label summarizing the interactions, (3) Scene label indicating the location of the action, and (4) Multi-sentence, natural language description of the clip. Within these labels, over 50 kinds of social relationships are annotated.

Video Social Relation dataset

The ViSR dataset was created in 2019 by Liu *et al.* [136]. It comprises $8,240$ video clips collected from over 200 movies with a wide variety of types. The length of each clip is around $10 \sim 30$ seconds. The dataset is annotated with eight types of social relationships derived from the domain-based theory [133];

Figure 4.2: Example of video frames from the ViSR dataset with eight types of social relationships.

"leader-subordinate", "colleague", "service", "parent-offspring", "sibling", "couple", "friend", and "opponent".

4.3 Social Relation Atmosphere Dataset

4.3.1 Overview

As introduced in 4.2.2, ViSR [136] is the latest dataset with the most number of video clips for social relationship research, so it is chosen as the basic dataset in this chapter. Each video was originally labeled with eight social relationships. Figure 4.2 displays examples of the relationships. The videos are collected from movies with a resolution of $1,280 \times 720$ pixels. Based on the ViSR dataset, here, each individual video segment is annotated with four kinds of social relation atmosphere labels. In the following, details of the labeling process are described.

4.3.2 Social Relation Atmosphere Annotation

Previous studies [32, 33, 152, 153] only analyzed social relationships, while ignoring the physical or emotional interactions between people and their surrounding

environment. Considering this, here, a new definition of *social relation atmosphere* is proposed. This aims to describe the overall conversational environment between two individuals on the basis of various visual cues, such as dialogue emotions, action activity, social status, and so on. According to the research by Toyoda *et al.* [154], the dialogue atmosphere with voice information can be defined by six characteristics: "cheerful", "serious", "miscommunication", "excited", "close", and "counterpart". However, since this thesis focuses on video frame data without voice or text analysis, the "miscommunication" attribute is difficult to analyze in the video, so it is discarded in the experiments. On the other hand, among all the $8,240$ video segments, only 318 segments show the "cheerful" attribute, and over 75% of these "cheerful" segments behave similarly to the "excited" attribute. As shown in the far right of Figure 4.3, the individuals show rich body movements and happy facial expressions. In this case, the attributes "cheerful" and "excited" overlap. Thus, only the "excited" label is retained which contains more video segments and rich situations including "cheerful". Therefore, the annotation of social relation atmosphere in video data is annotated down to four attributes, with each attribute labeled independently. The annotations are performed by two annotators. Below are the definitions of the four attributes.

- **Excited**: The individuals are pleased with one another or are enthusiastic about their surrounding environment. Typically, the characters tend to have large movement ranges or expressive facial expressions.
- **Serious**: The individuals behave and express themselves in a considered manner. The characters usually have no or very few facial expressions during the conversation. The interactions between two people are sincere and earnest.
- **Close**: The individuals are familiar with one another and maintain a friendly relationship. They usually have pleasant expressions on their face and speak with a smile. The gesture interactions between characters are kind and harmonious.
- **Counterpart**: During the conversation, the characters are dressed similarly, are of the same generation, or have the same social standing.

Figure 4.3 to Figure 4.10 illustrate examples of each attribute. Below each image, the original social relationship labels are also highlighted. From the sample

video frames, we can see that the same social relationship may present a different social relation atmosphere, while the characters belonging to the same social relation atmosphere may form different social relationships. This indicates that the social relationships and social relation atmosphere tasks are closely related but distinct. The distribution of the social relation atmosphere labels is shown in Figure 4.11.

4.4 Proposed Method

4.4.1 Overview

This section introduces the details of the proposed Relevant Visual Concept (RVC) for social relation atmosphere recognition. The pipeline is shown in Figure 4.12. Given a video dataset with both social relationship and social relation atmosphere labels, the social relation atmosphere is identified by utilizing the visual concepts from the social relationship recognition task as supplementary information. It consists of two steps: (1) Videos in training data are segmented into multiple supervoxels. The supervoxels are grouped into different visual concepts, such as "grass" or "head". A 3D explanation module computes a rank for each concept according to its importance when CNN predicts the social relationship, where a higher rank indicates that the concept is more important for social relationship recognition; and (2) Every video in the dataset is segmented into supervoxels. By matching each supervoxel to the concept calculated in Step 1, each supervoxel is classified into one concept. The video is saved as a masked video that only contains the highest-ranking concepts. A trained social relationship network is used to extract features from the masked video. Another trained social relation atmosphere network is used to extract features from raw videos. The final prediction is based on the combination of the two features.

4.4.2 3D Explanation Module

The purpose of the 3D explanation module is to extract the important regions for recognizing social relationships. For each social relation atmosphere attribute, let $\boldsymbol{V} = \{(v_1, y_R^1, y_A^1), (v_2, y_R^2, y_A^2), ..., (v_n, y_R^n, y_A^n) | n = 1, 2, ..., N\}$ be the training

Friend Leader-subordinate Opponent Service

Figure 4.3: Example of excited video frames.

Friend Colleague Sibling Parent-offspring

Figure 4.4: Example of not-excited video frames.

Friend Friend Colleague Parent-offspring

Figure 4.5: Example of counterpart video frames.

Leader-subordinate Parent-offspring Service Opponent

Figure 4.6: Example of not-counterpart video frames.

Couple Colleague Friend Parent-offspring

Figure 4.7: Example of close video frames.

Friend Leader-subordinate Couple Parent-offspring

Figure 4.8: Example of not-close video frames.

Leader-subordinate Colleague Parent-offspring Opponent

Figure 4.9: Example of serious video frames.

Opponent Sibling Leader-subordinate Parent-offspring

Figure 4.10: Example of not-serious video frames.

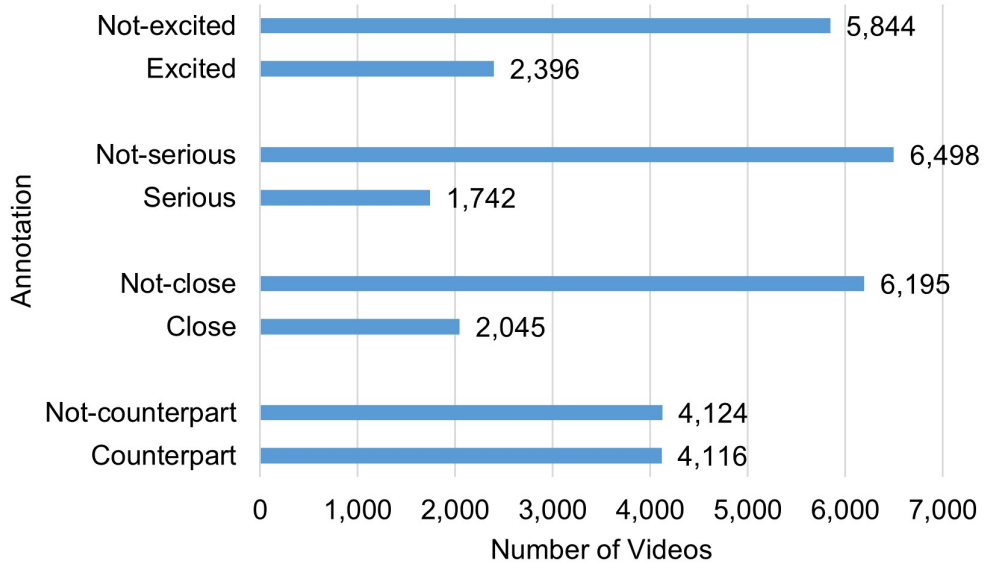


Figure 4.11: Distribution of the Social Relation Atmosphere Dataset.

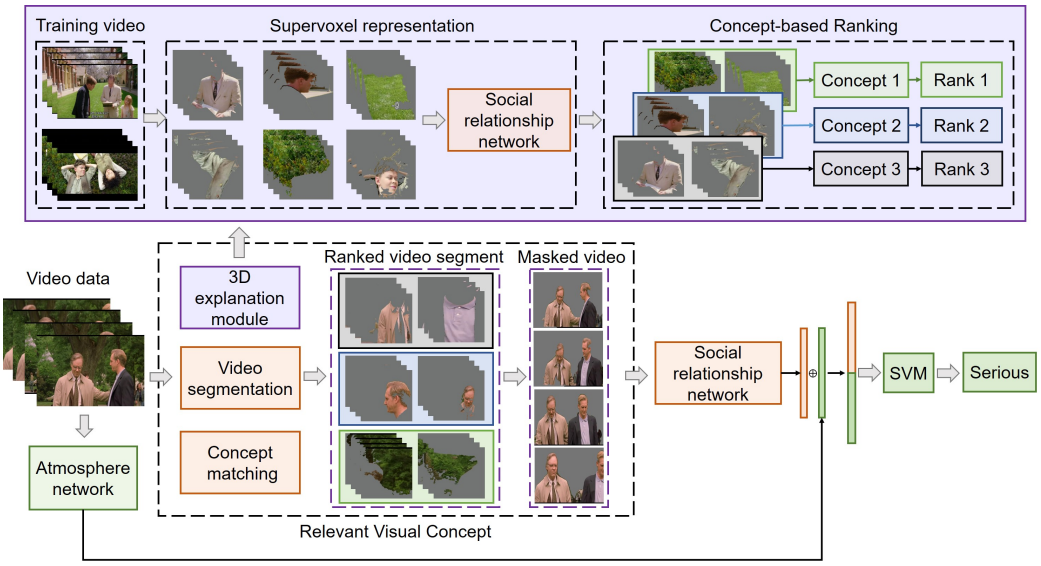


Figure 4.12: Overview of the proposed Relevant Visual Concept (RVC)-based method for recognizing social relation atmosphere.

video data in the ViSR dataset, which contains N videos. v_n is the n-th video, $y_R^n \in (1, 8)$ is the social relationship label, and $y_A^n \in (0, 1)$ is the binary label for the social relation atmosphere.

Supervoxel representation

First, a 3D Simple Linear Iterative Clustering (SLIC) [155] is used to divide videos into multiple spatial-temporal supervoxels. A 3D CNN F_R is trained from scratch on (v_n, y_R^n) to classify social relationships and used as a feature extractor. Each supervoxel is resized to the standard input size of the network. The empty regions in every frame are filled with average image value, as depicted in grey in Figure 4.12. The feature vectors are extracted from the top layer in F_R for each supervoxel.

Concept-based ranking

After extracting the deep features, supervoxels in social relationship class y_R^n are categorized into distinct concepts. Similar supervoxels are grouped into a concept. To determine which concept the network pays more attention to when predicting social relationships, the importance rating is evaluated for each concept. By utilizing the explanation module proposed in Chapter 5, the importance rank of each concept in class y_R^n can be determined.

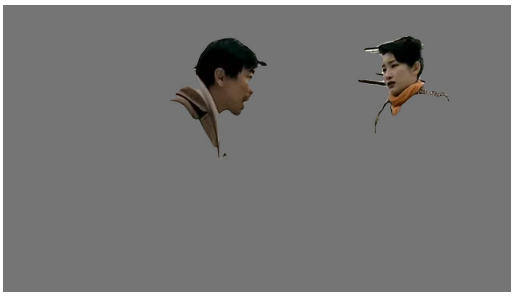
4.4.3 Relevant Visual Concept Extraction

After obtaining the concept ranking in the training data with social relationship label y_R^n, the high-ranked spatial-temporal volumes is highlighted on the raw video data to extract relevant social relationship visual concepts. The c-th concept for class y_R^n can be represented as $(r_c^{y_R^n}, \boldsymbol{f}_c^{y_R^n})$, where $r_c^{y_R^n}$ is the importance rank of the concept and $\boldsymbol{f}_c^{y_R^n}$ is the feature vector of the cluster center.

Let $\boldsymbol{T} = \{(t_1, y_R^1, y_A^1), (t_2, y_R^2, y_A^2), ..., (t_x, y_R^x, y_A^x) | x = 1, 2, ..., X\}$ represent the whole video data, which includes the test video. t_x is also segmented into P supervoxels. The p-th segment s_p^x can be represented as t_x masked with a mask m_p^x:

$$s_p^x = m_p^x \odot t_x. \tag{4.1}$$

(a) Raw video frame



(b) Masked frame with the Top 1 important visual concept as visible part.

(c) Masked frame with the Top 2 important visual concept as visible part.

(d) Masked frame with the Top 3 important visual concept as visible part.

Figure 4.13: Example of adding visual concept step by step. In (b), only the supervoxels that belong to the Top 1 rank concept are visible. Other irrelevant parts are masked with grey. (c) and (d) are masked frame examples of the Top 2 and the Top 3 visual concepts.

As demonstrated by Equation 4.2, each supervoxel is assigned to the closest concept c by calculating the distance between it and each cluster center.

$$c^* = \underset{c}{\operatorname{argmin}} D(F_R(s_p^x), \boldsymbol{f}_c^{y_R^n}), \tag{4.2}$$

where $F_R(s_p^x)$ is the feature vector of s_p^x extracted by social relationship CNN F_R.

Assume that a blank video volume has the same size as t_x. When supervoxels from different concepts are added to the blank video, the visible regions of the video can be generated as a spatial-temporal volume:

$$D_x^q = \sum_{j=1}^{q} M_j^x \odot t_x, \tag{4.3}$$

where M_j^x is the sum of supervoxel masks that belongs to concept j, and q is the number of visible concepts.

As shown in Figure 4.13, supervoxels are added based on their rankings to a blank video one by one. The intermediate video examples in Figure 4.13(b)–(d) represent D_x^q with different values of q. When the highest-ranked supervoxels are added to a blank video, a masked video that conveys the most important information for recognizing social relationships is obtained. Right now, video with a mask can be represented as $(D_x^q, \boldsymbol{f}_R^x)$, where $\boldsymbol{f}_R^x = F_R(D_x^q)$ is the feature vector extracted by F_R. $\boldsymbol{f}_R^x$ is the explainable visual concept feature used as supplementary information for the following social relation atmosphere recognition.

4.4.4 Social Relation Atmosphere Prediction

A 3D CNN F_A is trained from scratch on (v_n, y_A^n) to classify the social relation atmosphere. For each video t_x, $\boldsymbol{f}_A^x = F_A(t_x)$ is the feature vector extracted with social relation atmosphere CNN F_A. Finally, feature $\boldsymbol{f}_R^x$ is fused with $\boldsymbol{f}_A^x$ to recognize the social relation atmosphere.

4.5 Experiments

To evaluate the proposed relevant feature fusion method for recognizing social relation atmosphere, experiments were performed.

4.5.1 Implementation Details

Dataset

The proposed method is evaluated on the ViSR dataset [136] with social relation atmosphere attributes. The current dataset contains $8,240$ videos. 70% are used as training data, 10% are used as validation data, and the rest 20% are used as test data.

3D CNN

To evaluate the effectiveness of the extracted Relevant Visual Concept (RVC), experiments on three standard 3D CNN architectures: Convolutional 3D (C3D) [63], Residual 3D (R3D)-18 [138], and Inflated 3D (I3D) [139] networks are performed. Each is trained from scratch. Following [63], video frames are resized into 128×171 pixels. Random horizontal flipping and random cropping are applied for data augmentation. The training video frames are randomly cropped to the standard input size of 112×112 pixels, while the test video frames are center cropped. In the training stage, 16 continuous frames are randomly chosen as input. In the test stage, the middle 16 continuous frames are fed into the network. All the CNNs are optimized using Stochastic Gradient Descent (SGD) with the momentum set to 0.9. The total number of iterations is 150 epochs, the batch size is 64, and the learning rate starts from 0.01 for the first 50 epochs and decreases by a factor of 10 for every 50 epoch. The accuracy derived from the end-to-end CNN is considered as the baseline in the experiments.

3D explanation module

The standard settings in [156] are followed in order to ensure that the 3D explanation module is sufficient to extract RVCs. 200 videos from the training set were randomly selected to generate concepts per class. Three different resolution levels are employed to segment videos. Each video is segmented into 15, 50, and 80 supervoxels, separately. Similar supervoxels within a single video are eliminated. The number of clusters for each class is set to 25, with each cluster representing a concept. Only 40 supervoxels are retained in each concept. The activation for each supervoxel is extracted from the top layer. For C3D, the features from the last fully connected layer (FC7) are extracted. The global average pooling layer is used to extract features for R3D and I3D. Furthermore, 50 groups of random videos are generated from the HMDB database [157], which are used to differentiate the concept voxels and calculate concept activation vectors. All experiments are implemented in TensorFlow framework with two 24G NVIDIA RTX 3090 GPUs.

Feature fusion

The social relationship features extracted from RVCs with social relation atmosphere features are concatenated using the concatenation fusion method. The concatenation of deep features extracted from various CNNs has been proven to be a credible method for enhancing recognition accuracy [158].

Concatenation fusion

The two features are fused using concatenation, represented as $\boldsymbol{f}_{\text{fuse}} = [\boldsymbol{f}_R^x; \boldsymbol{f}_A^x]$.

4.5.2 Quantitative Analysis

The social relationship recognition accuracy on the ViSR dataset by using C3D, R3D, and I3D were 27.35%, 30.26%, and 33.11% respectively, which are comparable to the standard 3D CNN performance [159]. This ensures that the social relationship network is efficient, allowing us to extract valuable RVCs. The proposed method is quantitatively evaluated in the following three aspects.

Evaluation of RVCs

In the experiments, q was set to 5. According to Ghorbani *et al.* [81], the top 5 important Relevant Visual Concepts are sufficient to characterize the raw videos. Each spatial-temporal volume D_x^q is a masked video that represents the essential social relationship information. D_x^q is fed into F_R to extract features. The fused features $\boldsymbol{f}_{\text{fuse}}$ are fed into a linear SVM to make the final prediction. To evaluate the proposed Top 5 Relevant Visual Concepts, let's first compare it with three different methods:

- **End-to-end model**: The recognition accuracy obtained from the end-to-end social relation atmosphere model is used as the baseline.
- **Atmosphere + Relation**: The social relationship features are extracted from the entire raw videos, and concatenated with the social relation atmosphere features to make the predictions.

Table 4.1: Recognition accuracy [%] by using C3D network.

C3D	Excited	Serious	Close	Counterpart
Baseline (End-to-End)	72.77	80.72	75.13	54.76
Atmosphere+Relation	75.56	81.50	75.32	55.79
Atmosphere+Least 5	72.58	79.56	73.25	53.06
Proposed (Atmosphere+Top 5)	**77.32**	**81.93**	**75.68**	**59.19**

Table 4.2: Recognition accuracy [%] by using R3D network.

R3D	Excited	Serious	Close	Counterpart
Baseline (End-to-End)	73.01	81.81	75.19	55.91
Atmosphere+Relation	75.74	82.47	75.68	59.49
Atmosphere+Least 5	73.92	80.65	74.53	56.22
Proposed (Atmosphere+Top 5)	**76.90**	**84.05**	**76.47**	**66.04**

Table 4.3: Recognition accuracy [%] by using I3D network.

I3D	Excited	Serious	Close	Counterpart
Baseline (End-to-End)	75.50	83.74	77.93	62.40
Atmosphere+Relation	76.22	84.41	78.53	64.22
Atmosphere+Least 5	73.01	83.62	75.20	61.86
Proposed (Atmosphere+Top 5)	**78.90**	**85.81**	**79.56**	**66.65**

- **Atmosphere + Least 5**: The social relationship features are extracted from the five visual concepts with the least ranking.
- **Atmosphere + Top 5**: The proposed method; The social relationship features are extracted from the five visual concepts with the highest ranking.

The results are shown in Table 4.1, Table 4.2, and Table 4.3 by using C3D, R3D, and I3D, respectively. The social relationship features extracted from whole videos are shown to help improve the social relation atmosphere accuracy, which means social relationship tasks can provide useful information. Moreover, the proposed method performs the best, and shows a higher accuracy than using the whole video. It indicates that the most important RVCs which preserve the core information of social relationships can be discovered in the proposed method. Furthermore, when the least significant concepts are used, the recognition accuracy

decreases and even falls below the baseline. This demonstrates that some of the information in the raw videos is invalid and that not all of the data from the social relationships can be utilized to improve the social relation atmosphere. By utilizing the 3D explanation module, meaningless information can also be excluded.

Evaluation of concept numbers

The top 5 RVCs have been shown to be capable of representing the necessary information for the social relationship recognition task. Here, comparison experiments are further performed with varying numbers of concepts to explore how many concepts are sufficient to maintain the performance. Based on the initial five concepts with the highest rankings, let's add the next four Top N important concepts ($N = 6, 7, 8, 9$), and observe the cumulative effect. Table 4.4, Table 4.5, and Table 4.6 present the recognition accuracy by using C3D, R3D, and I3D, respectively. For each model, the bold with underline is the highest accuracy in each social relation atmosphere attribute. We can observe that with a different number of concepts, there is a slight perturbation in recognition accuracy. However, there is no specific best choice of concept numbers. This could be because the portion of useful information in each video differs. However, for each attribute, the majority of the highest accuracy is derived from the top six or seven RVCs. Instead, as more concepts are added, the accuracy begins to decline. This reflects that only very few RVCs (e.g. six or seven) are sufficient to maintain the information from the social relationship recognition task. The RVCs with a ranking of nine begin to contain some invalid information.

Evaluation of fusion methods

Finally, let's investigate the impact of fusion methods to determine whether they affect the performance of social relation atmosphere recognition. Besides the concatenation operation, max fusion, and mean fusion are compared.

- **Max fusion**: The two features are fused using max pooling, represented as $\boldsymbol{f}_{\text{fuse}} = \max(\boldsymbol{f}_R^x, \boldsymbol{f}_A^x)$.

Table 4.4: Evaluation of different number of concepts and fusion methods by using C3D.

C3D	Excited			Serious			Close			Counterpart		
Atmosphere+Top N	Mean	Con	Max	Mean	Con	Max	Mean	Con	Max	Mean	Con	Max
$N = 5$	77.20	77.32	77.14	**82.23**	81.93	81.67	75.74	75.68	**75.74**	58.82	59.19	58.04
$N = 6$	**77.32**	77.26	76.41	81.56	**82.53**	**81.87**	**75.80**	**75.99**	75.20	58.34	59.85	58.52
$N = 7$	77.01	**77.56**	77.20	81.32	81.75	81.56	75.50	75.56	75.50	**59.43**	**60.33**	**59.37**
$N = 8$	77.08	77.26	77.02	81.75	81.56	81.75	75.44	75.56	75.38	59.07	59.49	59.13
$N = 9$	76.53	77.44	**77.26**	81.67	81.75	81.63	75.68	75.80	75.62	59.07	59.24	58.70

Table 4.5: Evaluation of different number of concepts and fusion methods by using R3D.

R3D	Excited			Serious			Close			Counterpart		
Atmosphere+Top N	Mean	Con	Max	Mean	Con	Max	Mean	Con	Max	Mean	Con	Max
$N = 5$	77.01	76.90	76.77	**84.11**	84.05	83.93	76.47	76.47	76.96	65.86	66.04	66.89
$N = 6$	76.22	76.75	76.89	84.05	84.11	**84.29**	76.96	**77.50**	76.96	66.16	65.80	66.70
$N = 7$	76.40	77.13	76.59	83.87	84.35	84.11	**77.01**	**77.50**	**77.01**	66.26	65.73	66.22
$N = 8$	77.08	**77.20**	**77.08**	83.74	**84.41**	84.11	76.90	77.38	76.83	66.89	**67.13**	**66.94**
$N = 9$	**77.13**	77.08	76.53	**84.11**	84.35	83.87	76.71	76.77	76.65	**67.01**	66.76	66.52

Table 4.6: Evaluation of different number of concepts and fusion methods by using I3D.

I3D	Excited			Serious			Close			Counterpart		
Atmosphere+Top N	Mean	Con	Max	Mean	Con	Max	Mean	Con	Max	Mean	Con	Max
$N = 5$	78.05	78.90	77.80	85.02	**85.81**	85.26	**79.93**	79.56	79.26	66.53	66.65	**66.47**
$N = 6$	78.59	**78.96**	**78.29**	85.39	85.45	**85.32**	79.44	**80.17**	79.07	**66.83**	66.34	65.80
$N = 7$	**78.65**	78.23	77.93	**85.57**	85.02	85.20	79.56	79.44	79.38	66.04	**67.07**	65.68
$N = 8$	77.87	77.87	78.05	85.20	85.26	85.08	79.32	79.68	**79.50**	66.77	66.16	65.37
$N = 9$	78.17	78.60	77.44	85.39	85.08	84.78	79.62	79.68	79.26	66.04	66.22	66.04

- **Mean fusion**: The two features are fused using mean average pooling, represented as $\boldsymbol{f}_{\text{fuse}} = \text{mean}(\boldsymbol{f}_R^x + \boldsymbol{f}_A^x)$.

Table 4.4, Table 4.5, and Table 4.6 also present the recognition accuracy with different fusion methods. The accuracy in bold denotes the highest accuracy in each fusion method. Concatenation fusion performs the best of the three fusion techniques, followed by mean fusion and max fusion. This is because concatenation fusion retains all information from both social relationships and social relation atmosphere, whereas max fusion and mean fusion reduce the dimension of the features. However, it is evident that all the recognition accuracy performs better than the baseline and raw video feature fusion. This indicates that the proposed method not only can explore the most important visual concepts from social relationships but is also stable for any fusion method.

Consequently, based on three quantitative analyses, we can conclude that the proposed method is the most effective at identifying the social relation atmosphere. Using six or seven RVCs may enhance the accuracy, but more concepts will also result in more redundant information. The fusion methods will not influence the performance of the proposed method.

4.5.3 Qualitative Analysis

In order to qualitatively evaluate the proposed method, the masked video frames are visualized with different importance ranks. The advantage of the concept-based representation is high-level and human-understandable. Figure 4.14 and Figure 4.15 visualize two videos in the dataset. The social relationship of two individuals in Figure 4.14 is "sibling". And the two individuals in Figure 4.15 are "friend". In each figure, the first row is the raw video frames. The second row is the masked video frames with the most five important concepts, such as "face" and "hair". The third row is the masked video frames with the most nine important concepts. The fourth row is the masked video frames with the least five important concepts, such as "trees" and "grasses". We can see that the most important concepts mainly concentrate on humans and their interactions, while Figure 4.14d and 4.15d are primarily located in the background. We can also observe that the top 9 masked videos contain too much abundant information in some cases, which may lead to lower accuracy. This explains why the highest

accuracy appears in the different concept numbers in 4.5.2 and why the accuracy decreases with a large number of concepts. However, the existing segmentation method is not specifically designed for person-centered action. Thus, we can see from the visualization results that some supervoxels are still rough.

4.6 Summary

In this chapter, the improvement of subjective recognition accuracy as well as the generation of representative features were investigated. A novel Relevant Visual Concept (RVC)-based method was proposed for identifying subjective social relation atmosphere. Social relationship features were extracted to enhance the performance of social atmosphere recognition. In contrast to previous research that directly fuses features from raw videos, the proposed method innovatively extracts the most important spatial-temporal volume in the raw video data in order to represent the objective social relationships attribute. Videos are segmented into multiple supervoxels and similar supervoxels are clustered as a concept. A 3D explanation module proposed in Chapter 5 is utilized to provide a rank for each concept according to the contribution when the network makes a prediction. The videos are masked with only high-ranked concepts. Deep features extracted by social relationship CNN on the important visual concepts are used as relevant features. Then the subjective social atmosphere features are obtained with the objective social relationship feature as the final representation.

To evaluate the proposed method, based on the ViSR dataset, four social atmosphere labels were annotated to each video. Extensive experiments on this new dataset with three different 3D CNNs demonstrated that the proposed method performed better than the end-to-end prediction and raw video feature fusion. Finally, the detected video concepts were visualized with different numbers of important concepts. The visualization results also indicate the proposed 3D explanation module can disclose the most essential and least important regions for social relationships. In conclusion, the recognition of subjective attributes can be enhanced by utilizing objective relative features as supplementary information.

(a) Raw video frames

(b) Extracted top 5 important masked video frames

(c) Extracted top 9 important masked video frames

(d) Extracted least 5 important masked video frames

Figure 4.14: Masked videos frames with different numbers of concepts and different important ranks.

(a) Raw video frames

(b) Extracted top 5 important masked video frames

(c) Extracted top 9 important masked video frames

(d) Extracted least 5 important masked video frames

Figure 4.15: Masked videos frames with different numbers of concepts and different important ranks.

Chapter 5

Spatial-Temporal Model Explanation

5.1 Overview

After constructing a subjective violence extent dataset and proposing a feature fusion method for subjective social relation atmosphere recognition, explaining Convolutional Neural Networks (CNNs) becomes the top priority because it can help understand the inner decision procedure of CNNs and help improve the recognition accuracy further. CNNs have been widely used in various computer vision tasks, such as image classification [160, 161, 162], semantic segmentation [163, 164], object detection [165, 166], and so on. Despite the fact that CNN models show competitive performance in these tasks, current neural networks are still regarded as black boxes. Due to the large number of parameters and high nonlinearity [167], the underlying prediction mechanism is opaque. This reduces the reliability of neural networks in high-stakes real-world applications such as autonomous driving and medical image analysis [168, 169]. In recent years, Explainable Artificial Intelligence (XAI) has become a popular topic to help comprehend model predictions and increase the credibility of CNNs.

In general, the explanation methods can be divided into local and global methods. Local methods concentrate on understanding predictions on individual data instances, while global methods attempt to explain the overall logic of the target CNNs at the class or dataset level. This chapter focuses on the global explanation,

which is crucial to comprehend the overall behavior of the black boxes.

There are already some methods that provide explanations for 2D image classification CNNs [78, 87, 170, 171, 172], but most of them are local techniques. Zhou *et al.* [71] generated a Class Activation Map (CAM) using global average pooling for each image to highlight the discriminative regions that are used for the 2D CNN to predict class. Ribeiro *et al.* [170] proposed Local Interpretable Model-agnostic Explanations (LIME) to interpret the model by approximating the predictions in a local similarity neighborhood of a target image. However, these methods are not only limited to a single prediction, but they are also difficult for humans to comprehend. The highlighted regions are pixel-level, devoid of human-understandable semantic interpretation. More recently, interpretation with high-level concepts has attracted considerable attention. Kim *et al.* [72] introduced Concept Activation Vectors (CAVs) which use the directional derivatives to quantify the importance of the network prediction to user-defined concepts. Based on this, Ghorbani *et al.* [81] proposed ACE (Automatic Concept-based Explanation) to discover the relationship between image segments and image classification prediction.

Despite solid achievements in 2D image classification interpretation, only a few studies have attempted to interpret 3D action recognition CNNs, primarily due to the huge computational cost and rich spatial-temporal content of video data. Existing 3D explanation methods are mainly extended from 2D local explanation methods. Stergiou *et al.* [173] proposed Saliency Tubes, which applied Grad-CAM [73] to 3D CNNs. The activation maps of the 3D CNN's final convolutional layer are combined to produce heatmaps of input videos. Li *et al.* [88] adopted Extremal Perturbations (EP) [76] to the video case by adding a spatial-temporal smoothness constraint. However, these methods have two major drawbacks: (1) Discriminative 3D regions are based on a single frame and lack spatial-temporal consistency, and (2) Regions are pixel-level and lack high-level semantic information.

To address these issues, here, 2D ACE [81] is extended to 3D and a high-level global interpretation is proposed. For each class, videos are segmented into multiple spatial-temporal supervoxels. Similar supervoxels are grouped to form a meaningful concept. The proposed method can assign a score for each concept according to its contribution when the network makes a prediction. When

interpreting the decision procedure of 3D action recognition CNNs, instead of highlighting essential pixels for a single video, it can answer two fundamental questions at the class level: *which objects or motions in the video are significant for a particular action recognition class* and *which object or motion is the most crucial clue in this class*.

The contributions of this chapter include:

1. Spatial-temporal Concept-based Explanation (STCE) for 3D CNNs is proposed. The discriminative regions are spatial-temporal continuous and human-understandable. STCE is among the first to achieve action recognition interpretation based on high-level video supervoxels.
2. The proposed method is validated using 3D CNNs on the Kinetics-700 [174] and KTH [175] datasets. Both qualitative and quantitative results demonstrate that it can explain the 3D action recognition CNNs consistent with human cognition.
3. The proposed STCE can be used as a plug-in module for interpreting and enhancing the procedure of subjective video attribute recognition.
4. The source code is publicly available[1], making this work serve as a starting point in the research area of 3D XAI.

5.2 Proposed Method

In this section, the details of the proposed Spatial-Temporal Concept-based Explanation (STCE) method are introduced. Given a video classification dataset and a 3D CNN that has been trained using the dataset, the network is interpreted by investigating the most important spatial-temporal volumes from the training videos. The pipeline is shown in Figure 5.1. The procedure consists of two steps: (1) Raw videos are first segmented into multi-resolution spatial-temporal volumes. The green, blue, and orange colors shown in step (1.1) indicate that videos are segmented into 15, 50, and 80 supervoxels, respectively. A 3D CNN trained on the dataset is then used to extract the feature vector of each supervoxel, and (2) Supervoxels are grouped into different clusters. Each cluster is a meaningful concept, such as “hand”, “sausages”, or “grass”. STCE finally evaluates the importance

[1] `https://github.com/yingji425/STCE` (Accessed: 2023/09/04)

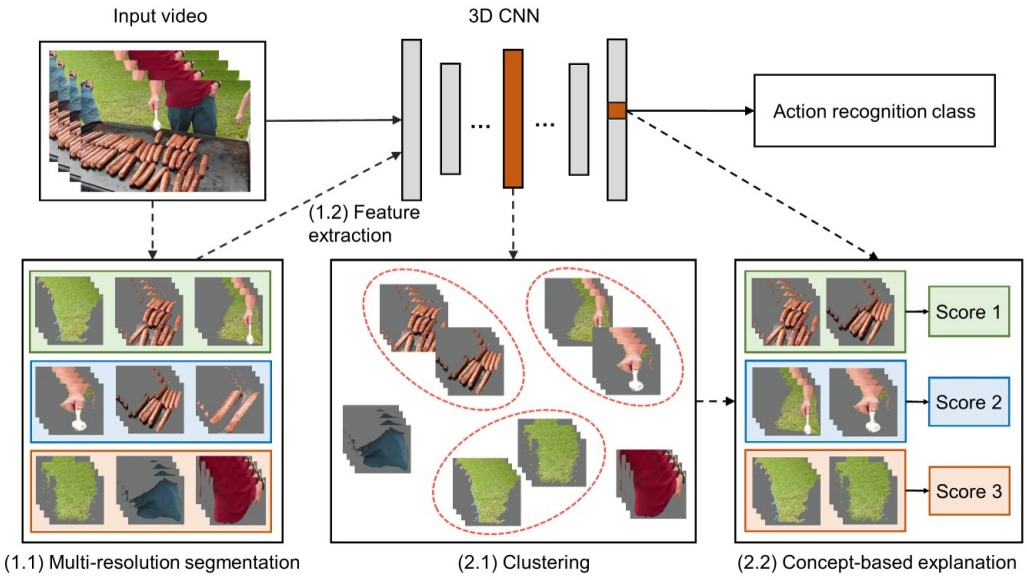


Figure 5.1: Overview of the proposed Spatial-Temporal Concept-based Explanation (STCE) method. The input is videos from the same class. The video shown here is from the "cooking sausages" class.

score of each concept with respect to the class it belongs. Within the prediction-making procedure, the network pays more attention to the concepts with high scores.

5.2.1 Supervoxel Representation

Let $\boldsymbol{V} = \{(v_1, y_1), (v_2, y_2), ..., (v_n, y_n) | n = 1, 2, ..., N\}$ be an action recognition dataset which contains N videos. v_n is the n-th video with a label $y_n \in (1, Y)$. Each video is first segmented into supervoxels. In contrast to previous research [176, 177], which simply divided videos into segments with equal time intervals, the proposed method uses 3D Simple Linear Iterative Clustering (SLIC) [155] to divide videos due to its superior performance in video segmentation [178]. In this case, videos are segmented into meaningful spatial-temporal volumes, such as a wheel of a moving car or a swinging arm. Since a video contains information ranging from fine-grained still texture to coarse-grained continuous action motion, each video is segmented three times with different levels of resolution to preserve the hierarchical information. For each video v_n, $[s_n^{\text{small}}, s_n^{\text{middle}}, s_n^{\text{large}}]$ contains

different size of segments. To avoid calculational cost for redundant supervoxels, the similarity between every two supervoxels is calculated. When the Jaccard index score [179] between two supervoxels is larger than a threshold (set to 0.5 in the experiments), these two segments are recognized as similar pairs. Duplicate segments will be removed, and only the most distinguishable supervoxels will remain.

A 3D CNN is trained from scratch on $\boldsymbol{V}$ and is used as a feature extractor. Each supervoxel is resized to the standard input size of the network. The empty regions in each frame are filled with average image value, as depicted in grey in Figure 5.1. The feature vectors are extracted from the top layer l for each supervoxel.

5.2.2 Concept-Based Explanation

After extracting the deep features, supervoxels of class y are categorized into distinct concepts. By calculating the Euclidean distance between every pair of supervoxels, similar supervoxels are grouped as a single concept. To preserve the distinctiveness between different clusters, only a small number of segments (40 in the experiment) that are close to the center of each concept are retained. The remaining segments are discarded. Videos in class y can be represented as C groups of concepts, where $\text{Conceps} = \{\text{concept}_c | c = 1, 2, .., C\}$, and each concept concept_c contains 40 supervoxels. s_c^y represents all the segments belonging to the c-th concept.

To determine which concept the network pays more attention to when making the prediction, the importance rating for each concept is evaluated. To this end, a Concept Activation Vector (CAV) [72] is caculated to characterize the concept. The pipeline to generate the vector $\boldsymbol{v}_c^l$ is illustrated in Figure 5.2. All the segments s_c^y are put into the trained CNN as positive samples, while a group of random videos from irrelevant datasets is used as negative samples. Using the 3D CNN, features are extracted from both concept supervoxels and random videos. A linear classifier is learned to separate the positive and negative samples. The vector $\boldsymbol{v}_c^l$ orthogonal to the decision boundary is used to represent the c-th concept.

In order to figure out the impact of the concept_c given to a video v_n from class y, the idea from [72] is followed to calculate the gradient of logit with respect to

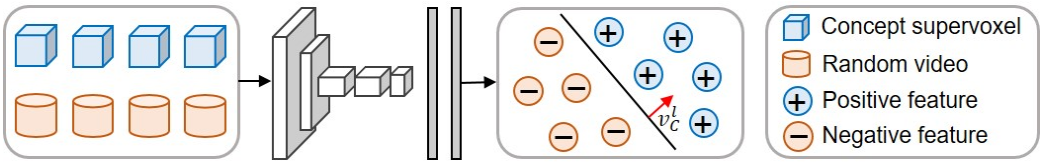


Figure 5.2: Pipeline to generate a Concept Activation Vector (CAV). The inputs are concept supervoxels and the same number of random videos. The direction of the red arrow is orthogonal to the decision boundary of the classifier. The vector $\boldsymbol{v}_c^l$ is used to represent the concept.

the activations of v_n in layer l. Thus the importance score of a particular concept can be computed as:

$$\begin{aligned} I_{c,y,l}(v_n) &= \lim_{\epsilon \to 0} \frac{p_{l,y}(\boldsymbol{f}_l(v_n) + \epsilon \boldsymbol{v}_c^l) - p_{l,y}(\boldsymbol{f}_l(v_n))}{\epsilon} \\ &= \nabla p_{l,y}(\boldsymbol{f}_l(v_n)) \cdot \boldsymbol{v}_c^l, \end{aligned} \tag{5.1}$$

where $\boldsymbol{f}_l(v_n)$ is the feature vector of the input video, $p_{l,y}$ is the logit for the video v_n from class y, and $\boldsymbol{v}_c^l$ is the concept vector.

When $I_{c,y,l}(v_n)$ is greater than zero, it indicates that this concept positively affects the CNN's prediction for video v_n. If $I_{c,y,l}(v_n)$ is less than zero, the concept has a negative impact.

For one class with K input videos, the directional derivatives for each video is calculated. The total importance score for one concept is defined as:

$$S_{c,y,l} = \frac{|v_n \in \boldsymbol{V} : I_{c,y,l}(v_n) > 0|}{K} \in [0, 1]. \tag{5.2}$$

For each concept concept_c, the score $S_{c,y,l}$ computes the proportion of input videos that are positively influenced by the concept, where a higher S indicates the most concerning part for a 3D CNN to recognize the video. By sorting the scores, the importance rank of each concept for class y can be determined. Unlike previous research, which assessed the importance score of each pixel, the proposed method interprets the CNN using concepts with videos from the entire class.

5.3 Experiment

In this section, empirical evaluations of the proposed STCE interpretation method for the 3D CNNs are presented. 5.3.1 describes the dataset and system set-up, 5.3.2 introduces evaluation metrics for the experiments, 5.3.3 presents the quantitative results of adding and removing concepts, and 5.3.4 interprets the CNN by visualizing the concept frames compared to raw videos. Finally, 5.3.5 discusses the influence of different parameters.

5.3.1 Implementation Details

Datasets

The proposed method is evaluated on two popular datasets: Kinetics-700 human action recognition dataset [174] and KTH Action dataset [175].

The Kinetics dataset contains 700 action classes. The proposed STCE interprets the performance of CNN at the class level. Thus, ten classes are randomly selected from the raw dataset to conduct the interpretability experiment. As training data, a total of $6,846$ videos are utilized, while as test data, 480 videos are utilized. The video clips have variable high resolutions.

The current KTH dataset includes six types of human actions: walking, jogging, running, boxing, hand waving, and hand clapping. In total, the dataset contains $2,391$ video sequences. Each video has a low resolution of 160×120 pixels. The experiment setup by Liu *et al.* [180] is followed, 80% of the dataset ($1,528$ videos) are used for training, and the remaining 20% (863 videos) are used for validation.

3D CNN

Experiments are conducted on three standard 3D CNN architectures: Convolutional 3D (C3D) [63], Residual 3D (R3D)-18 [138], and Inflated 3D (I3D) [139] networks. Each is trained from scratch. Following Tran *et al.* [63], video frames in the Kinetics dataset are resized into 128×171 pixels. Due to the low resolution, videos in the KTH dataset are resized to 120×120 pixels. Random horizontal flipping and random cropping are applied for data augmentation. The training video frames are randomly cropped to the standard input size of 112×112 pixels, while

the test video frames are center cropped. In the training stage, 16 continuous frames are randomly chosen as input. In the test stage, the middle 16 continuous frames are fed into the network. All the CNNs are optimized using Stochastic Gradient Descent (SGD) with a momentum set to 0.9. The total number of iterations is 150 epochs, the batch size is 64, and the learning rate starts from 0.01 for the first 50 epochs and decreases by a factor of 10 for every 50 epoch. The accuracy derived from the end-to-end CNN is considered as the baseline in the experiments.

STCE configuration

After training a 3D CNN, the next step is to interpret the prediction procedure. 200 videos were randomly selected from the training set to generate concepts per class. Three different resolution levels were set to segment the videos. Each video was segmented into 15, 50, and 80 supervoxels, separately. Similar supervoxels within a single video were eliminated. The number of clusters for each class was set to 25 with each cluster being a concept. Only 40 supervoxels were retained in each concept. The activation for each supervoxel was extracted from the top layer l. For C3D, the features from the last Fully Connected layer (FC7) were extracted. The global average pooling layer was used to extract features for R3D and I3D networks. Furthermore, 50 groups of random videos were also generated from the Human Metabolome Database (HMDB) [157]. The random videos were used to differentiate the concept voxels and calculate concept activation vectors, as described in Figure 5.2. All experiments were implemented in the TensorFlow framework with two 24G NVIDIA RTX 3090 GPUs.

5.3.2 Evaluation Overview

This section introduces the evaluation procedure for the experiment. The concepts calculated in 5.2.2 are validated on the test data. After calculating the importance score $I_{c,y,l}$ with training data, the c-th concept for class y can be represented as $(r_c^y, \boldsymbol{f}_c^y)$, where r_c^y represents the importance rank of the concept, and $\boldsymbol{f}_c^y$ is the feature vector of the clustering center that has the same dimension as the supervoxel's activation. To quantitatively evaluate the influence of each concept, the recognition accuracy is evaluated by adding and removing video concepts one by

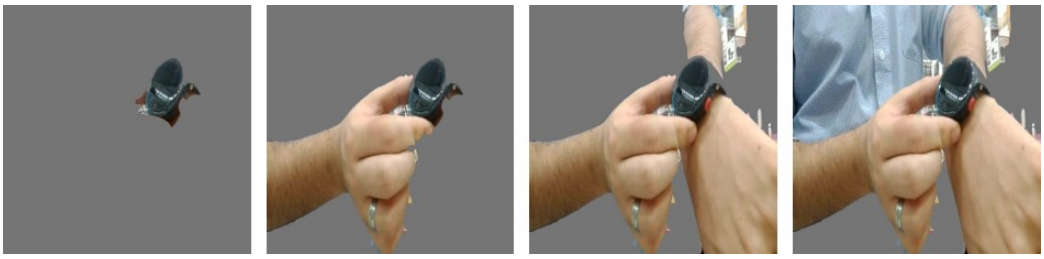


Figure 5.3: Example of adding concepts from a blank video. The sample frame is from the "checking watch" class. In each step, supervoxels belonging to a specific concept are added to the existing video. For example, the first video represents adding supervoxels belonging to the "watch" concept. The second video represents adding supervoxels that belong to the "left hand" concept to the first video.

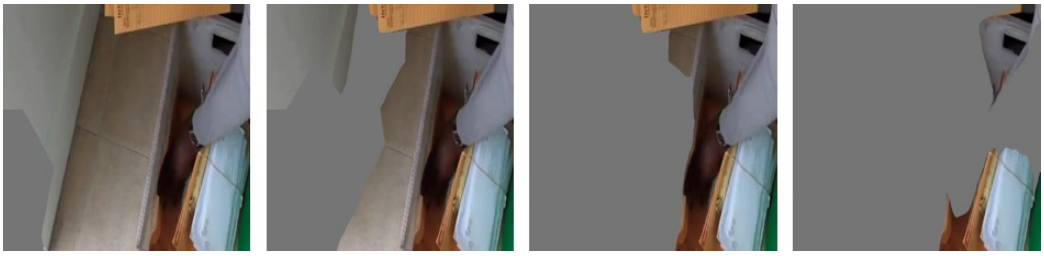


Figure 5.4: Example of removing concepts from a test video. The sample frame is extracted from the "delivering mail" class. In each step, all supervoxels from one concept are removed from the raw video.

one from the test video.

For each test video t_x, the video is also segmented into P supervoxels. The p-th segment s_p^x can be represented as t_x masked with a mask m_p^x:

$$s_p^x = m_p^x \odot t_x. \tag{5.3}$$

As demonstrated by Equation 5.4, each supervoxel is assigned to the closest concept c by calculating the distance between it and each clustering center:

$$c^* = \underset{c}{\operatorname{argmin}} D(f_p^x, \boldsymbol{f}_c^y), \tag{5.4}$$

where $\boldsymbol{f}_p^x$ is the feature vector of s_p^x.

Let's assume that a blank video volume has the same size as the test video t_x. When supervoxels from different concepts are added to the blank video, the

Table 5.1: Recognition accuracy of adding concepts using the Kinetics dataset. The baseline is the end-to-end accuracy [%] by 3D CNNs.

Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	22.29	34.79	43.33	50.83	**58.54**	79.58
	Random	21.88	33.33	39.79	49.17	55.83	
	Least	23.33	30.63	37.29	45.83	52.29	
R3D	Top	11.67	23.96	32.92	39.38	**46.67**	75.62
	Random	10.63	21.25	32.50	37.71	41.25	
	Least	9.79	16.04	26.04	33.13	41.46	
I3D	Top	23.33	37.92	46.88	54.38	**61.88**	85.63
	Random	25.83	37.50	46.04	52.71	56.67	
	Least	25.42	37.50	47.29	51.46	55.83	

visible regions of the video can be generated as a spatial-temporal volume:

$$R_x^q = \sum_{j=1}^{q} M_j^x \odot t_x, \tag{5.5}$$

where M_j^x is the sum of supervoxel masks that belongs to concept j. q is the number of concepts that will be set in the following experiments.

As shown in Figure 5.3, supervoxels are added to a blank video one by one. The intermediate examples are R_x^q with different values of q. When adding all the supervoxels from t_x, the blank video will be the same as the test video t_x.

In contrast, when supervoxels are removed from raw video t_x, the visible regions are represented as $(1 - M_j^x) \odot t_x$. Figure 5.4 demonstrates the procedure of removing different concepts.

5.3.3 Quantitative Analysis

In this experiment, q in Equation 5.5 was set to 5, which indicates at most five different concepts will be removed from the raw video. For each test video, when adding and removing the concept, the spatial-temporal volume R_x^q is fed into the CNN and a prediction is made.

Table 5.2: Recognition accuracy of removing concepts in the Kinetics Dataset. The accuracy decreases the most when the most significant concepts are removed.

Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	74.38	60.21	55.42	47.50	**43.54**	79.58
	Random	74.38	64.38	59.38	51.88	45.00	
	Least	75.21	66.04	60.21	53.75	47.71	
R3D	Top	69.79	66.25	50.83	39.58	**24.38**	75.62
	Random	72.29	64.38	51.04	39.79	28.13	
	Least	73.33	64.38	51.67	42.29	28.13	
I3D	Top	74.58	65.63	58.96	49.79	**41.88**	85.63
	Random	78.33	70.83	60.42	53.75	43.96	
	Least	80.21	71.25	65.00	57.92	46.25	

Table 5.1 represents the experimental results of adding concepts using the Kinetics dataset. For each model, the first row represents the accuracy of adding concepts with the highest scores, the second raw, adding concepts with random scores, and the third row, adding concepts with the lowest scores. We can see that adding the most important concepts can lead to higher recognition accuracy for the CNNs, whereas the concepts with the lowest importance score can offer very little information. In addition, after adding five important concepts, the accuracy exceeds 70% of the baseline for C3D and I3D, and 60% of the baseline for R3D.

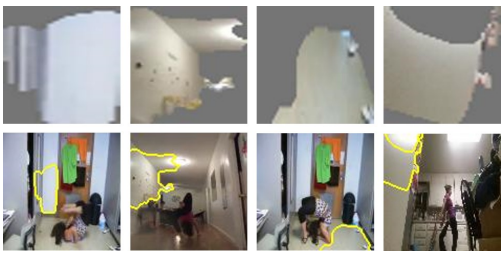
Table 5.2 demonstrates the influence of removing concepts. It is evident that removing the essential concepts results in a reduction in accuracy. Especially for R3D, the accuracy is only 30% of the baseline after removing five concepts. These experimental results indicate the proposed STCE is capable of revealing which concept the CNN focuses on and how much role it plays during the prediction.

5.3.4 Qualitative Analysis

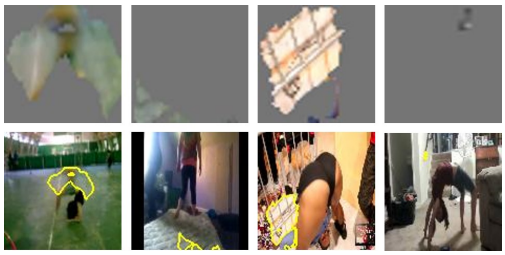
In order to qualitatively evaluate the proposed model, video frames of the detected concepts are visualized in Figure 5.5. In particular, the most and the least significant concept examples from the "bending back" class are illustrated in the Kinetics dataset. Figures 5.5a and 5.5b present the supervoxel frames belonging to the top two important concepts. The highest importance score is close to 1,

(a) Importance score = 0.94

(b) Importance score = 0.89



(c) Importance score = 0.26



(d) Importance score = 0.25

Figure 5.5: Visualization of four concepts from the "bending back" class using the C3D network. The first row of each subfigure is highlighted supervoxel frames. The second row is video frames from raw videos. (a) and (b) are the most two important concepts for CNN prediction. (c) and (d) are two concepts with the least significance.

indicating that this concept positively influenced nearly all of the test videos in this class. The first row of each figure shows the highlighted regions, while the second row displays the corresponding raw video frames. It is evident that the dominant actions are body parts and bending actions for predicting the "bending back" class.

Similarly, two groups of supervoxel frames belonging to the least important concepts are presented in Figures 5.5c and 5.5d. In contrast, these highlighted regions are primarily located in the background and lack significance. The visualization results interpret what the 3D CNN focuses on when recognizing actions. It is obvious that the concepts are intuitive and consistent with human understanding. The remarkable consistency of both quantitative and qualitative results confirms that the proposed STCE is effective for interpreting 3D CNNs.

Table 5.3: Recognition accuracy of adding concepts on the KTH dataset with Standard setting.

Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	21.21	23.52	29.43	39.17	**46.23**	91.31
	Random	19.35	23.52	25.26	28.27	32.91	
	Least	17.27	18.77	20.97	25.26	31.87	

Table 5.4: Recognition accuracy of removing concepts on the KTH dataset with Standard setting.

Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	89.92	84.94	81.11	76.83	**70.45**	91.31
	Random	90.50	89.80	86.67	82.04	74.74	
	Least	90.50	89.80	89.46	86.44	81.23	

5.3.5 Discussion

In this section, the influence of various parameter settings is examined. The number of concepts and supervoxels is mainly explored through comparative experiments on the KTH dataset. In particular, two types of parameter settings are explored for extracting important concepts.

- **Standard setting**: This setting is the same as experiments on the Kinetics-700 dataset in 5.3.3. Each video is divided into 15, 50, and 80 segments, separately. The number of concept clusters is set to 25.
- **Small setting**: STCE is also conducted with small parameters because the KTH dataset has a relatively low resolution. In this instance, each video is segmented into 15, 30, and 60 segments, respectively. All the supervoxels in the same class are clustered into 15 concepts.

Table 5.3 and Table 5.4 illustrate the accuracy of action recognition with the standard setting, while Table 5.5 and Table 5.6 show the accuracy with the small setting. We can see that both settings are consistent with the tendency demonstrated in 5.3.3. However, we can also see that despite the fact that adding concepts will undoubtedly improve accuracy, the accuracy only reaches 50% of the

Table 5.5: Recognition accuracy of adding concepts on the KTH dataset with Small setting.

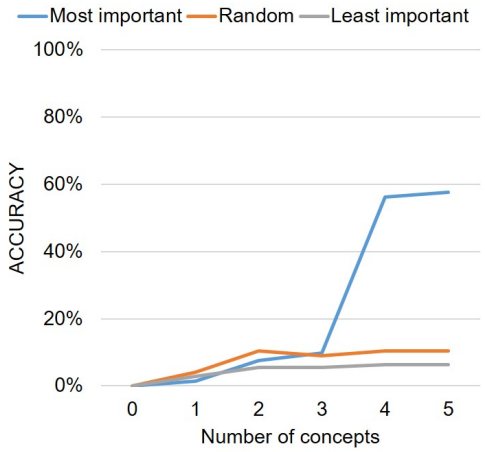
Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	35.46	54.35	66.63	69.52	**73.81**	91.31
	Random	27.69	43.92	59.68	67.44	71.84	
	Least	22.94	34.07	48.44	65.82	68.37	

Table 5.6: Recognition accuracy of removing concepts on the KTH dataset with Small setting.

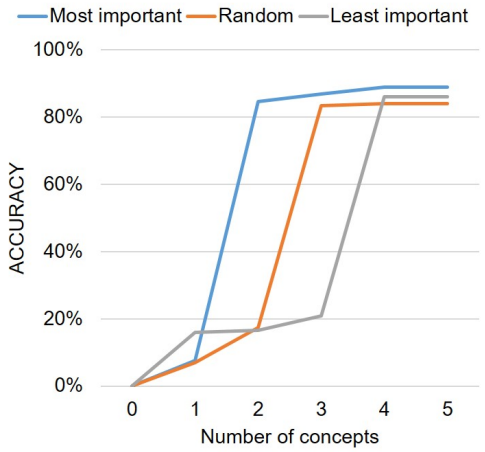
Model	Concepts	1	2	3	4	5	Baseline
C3D	Top	89.69	86.10	78.10	66.86	**59.68**	91.31
	Random	90.50	88.88	78.68	73.93	69.18	
	Least	91.43	89.92	85.75	78.91	71.73	

baseline in the standard setting. The phenomenon is the same when concepts are removed from the test video. On the other hand, the small setting can improve the effectiveness of concepts more than the standard setting, which reaches 80% of the baseline. Here, the "jogging" class is taken as an example, and their statistical charts are shown in Figure 5.6 and Figure 5.7. From these statistical charts, we can conclude that for low-resolution datasets, the CNNs obtain more information from large-scale concepts.

To more intuitively visualize the difference between two settings, Figure 5.8 shows the concept results with both settings extracted from the "boxing" class from the same raw video. Due to the low resolution and large blank background, it is evident that most of the essential regions for the KTH dataset are located on human body parts. This means that using the standard setting will result in quite a number of backgrounds, which can not improve the recognition accuracy. When using a small setting, the clustered concepts are easier to recognize. For comparison, the concept from the "checking watch" class in the Kinetics dataset is also visualized in Figure 5.9. Since the high-resolution dataset contains abundant information such as watch bands, hands, desks, and watches, even small concepts are sufficient to provide enough information.

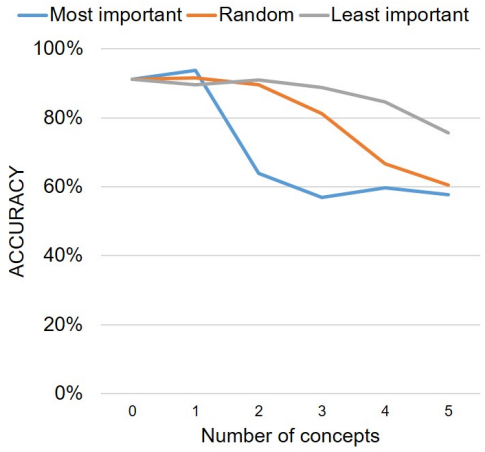


(a) KTH standard setting

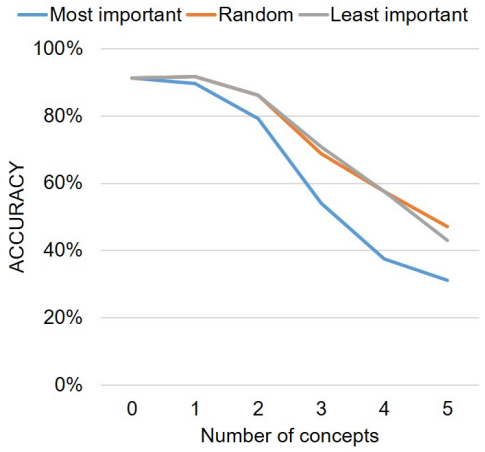


(b) KTH small setting

Figure 5.6: Performance of adding concepts using Standard and Small settings in the "jogging" class from the KTH dataset.



(a) KTH standard setting



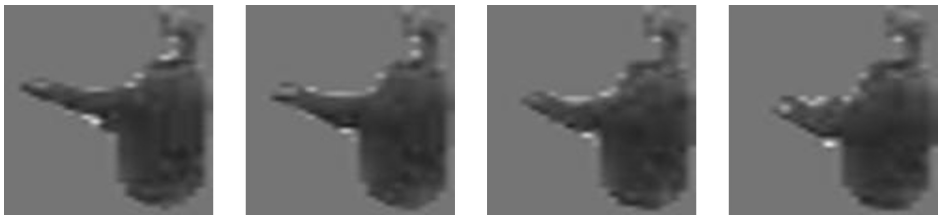
(b) KTH small setting

Figure 5.7: Performance of removing concepts using Standard and Small settings in the "jogging" class from the KTH dataset.

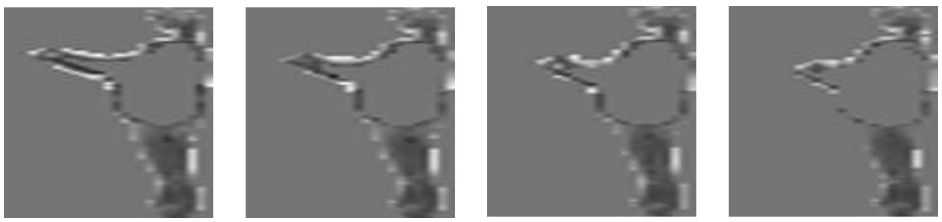
5.4 Summary

In this chapter, a Spatial-Temporal Concept-based Explanation (STCE) method for interpreting 3D CNN was proposed. In contrast to the prior pixel-level strategy which focuses on a single instance, the proposed method is the first attempt

(a) Raw video frames



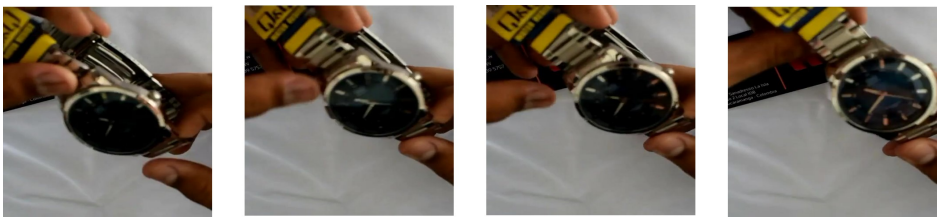
(b) Concept supervoxel frames excrated using the Standard setting



(c) Concept supervoxel frames extracted using the Small setting

Figure 5.8: Concept frames from the "boxing" class in the KTH dataset with Standard and Small settings.

to offer a human-understandable high-level explanation. Concretely, first, videos from an entire class are segmented and clustered into concepts. Each concept comprises similar meaningful supervoxels. Then, importance scores for each concept are calculated. Extensive experiments on three different 3D CNNs demonstrated the efficiency of the proposed STCE. Later, the detected concepts were visualized according to the scores, where the most and the least essential concepts were consistent with human perception. Finally, the choice of various parameters for the low-resolution dataset was investigated. The number of concepts and clusters did not affect the tendency reported in the experiments. Thus the proposed method successfully disclosed the prediction mechanism under the 3D CNN.



(a) Raw video frames

(b) Concept supervoxel frames using Standard setting

Figure 5.9: Concept frames from the “checking watch” class in the Kinetics-700 dataset with the Standard setting.

Chapter 6

Conclusion and Future Plan

6.1 Conclusion

The main objective of the work introduced in this thesis was to recognize subjective video attributes. In contrast to objective attributes, which can be definitively determined, subjective attributes are intangible and reliant on personal opinions, lacking a ground truth. When annotating subjective attributes, the annotations of various individuals are easily influenced by their individual experiences, cultural backgrounds, and subjective factors. When recognizing subjective attributes, the features extracted from end-to-end neural networks are not representative enough. When explaining subjective attributes, there is no high-level explanation for 3D Convolutional Neural Networks (CNNs). The difficulties in recognizing subjective attributes are from labeling, and training to explaining. Thus, three research questions were raised in Chapter 1:

1. How to construct a clean dataset and provide stable and reliable annotation for subjective attributes.

2. How to improve the accuracy of subjective video attribute recognition and generate targeted features.

3. How to explain the inner procedure of 3D CNN.

Chapter 3 mainly tackled the first question, Chapter 4, the second question, and Chapter 5, the third question.

In Chapter 3, different from previous objective violent object or action detection, the analysis of subjective video violence ratings was explored. To provide a reliable and stable violent dataset with subjective annotation, the TrueSkill [46] pairwise comparison was used to obtain the ground-truth violence score for each video. The convergence and stability of the TrueSkill score have been verified. On the other hand, designed for recognizing the subjective extent attribute, a rank learning method was proposed. The proposed method could learn the relationship between videos from different levels or the same level. With the proposed method, we can recognize violence extent better than by taking the classification approach.

In Chapter 4, the subjective attribute recognition on social relation atmosphere recognition was explored. Since end-to-end neural networks show low accuracy on subjective recognition, feature fusion was used to increase the discrimination of deep features. For each video data, both subjective and objective attributes exist simultaneously. The relationship between both attributes was investigated and the possibility of using objective attributes to help enhance the recognition of subjective attributes was explored. A Relevant Visual Concept (RVC) module was proposed for analyzing social relation atmospheres. The most significant video volumes from the social relationship recognition task are used to supplement social relation atmosphere recognition. The combined features were shown to represent subjective attributes better than end-to-end networks.

In Chapter 5, a global STCE (Spatial-Temporal Concept-based Explanation) method was proposed for interpreting 3D CNNs. In this method, (1) Videos are represented with high-level supervoxels, which are clustered as a concept. This is straightforward for humans to understand, and (2) Interpretation framework calculates a score for each concept, which reflects its significance in the CNN decision procedure. The explanation module allows us to investigate the impact of the concepts on a target task in-depth, such as social relation atmosphere recognition in Chapter 4.

In summary, two subjective datasets were constructed for analyzing video violence rating and social relation atmosphere, and two training approaches were proposed for recognizing the extent attributes and general subjective attributes, respectively. Finally, a plug-in explanation module was proposed for interpreting and enhancing the procedure of subjective video attribute recognition.

6.2 Limitation and Future Plan

The questions investigated in this thesis may pave the way for further subjective attribute recognition. In this section, potential new research ideas are discussed based on the findings of this thesis.

Advanced dataset

The currently proposed datasets were constrained by the limited number of videos and manpower. Especially the size of the violent video dataset presented in Chapter 3 is insufficient for popular deep neural networks. Using crowdsourcing effectively is one way to improve the data collection and annotation process. To improve the reliability of crowdsourcing, a golden standard from experts can be used. Annotations that deviate excessively from the standard will be eliminated. At the same time, a setwise comparison can also be proposed so that three or more videos can be compared simultaneously. Large-scale datasets with stable annotations will be a priority for subjective attribute recognition.

Multimodal recognition

When recognizing video violence extent and social relation atmosphere, the proposed work primarily concentrated on utilizing visual information with basic neural networks. However, in reality, textual, auditory, and other relevant data also play important roles in video. The integration of information from multiple modalities can improve the accuracy and understanding of subjective attributes. Textual and auditory information as well as physiological information could be collected for comprehensive subjective attribute recognition.

Vision transformer for video recognition

The recognition approaches proposed in the thesis are CNN-based deep learning architectures, which operate on the whole video. However, more recently, Vision Transformers (ViTs) have outperformed CNNs in multiple tasks because they are patch-based processing methods. Since ViTs process the entire video as a sequence of patches, applying ViT for subjective attribute recognition makes

it possible to analyze the contextual relationship between different regions and better generate explainable attention maps.

Explainable AI for subjective attributes

An explanation module was proposed in Chapter 5 and has been utilized in Chapter 4. However, the current explainable artificial intelligence is not specially designed for subjective attributes. The hints for analyzing subjective attributes are always concealed in human interactions. Thus, in Chapter 5, the segmented supervoxels had a significant impact on the explanation. In this case, a human-centered or action-centered segmentation method can be developed so that supervoxels can more accurately represent subjective attributes. The explanations can emphasize more on identifying emotions, sentiments, or other subjective attributes.

Nevertheless, the work presented in this thesis has established the groundwork for the discussed future work, and I eagerly anticipate future research that further improves the recognition of subjective attributes in video data.

Publication List

Journal Papers

1. **Ying Ji**, Yu Wang, Kensaku Mori, and Jien Kato. Social relation atmosphere recognition with relevant visual concepts. *IEICE Transactions on Information and Systems*, E106(10): 2023.

2. **Ying Ji**, Yu Wang, Jien Kato, and Kensaku Mori. Predicting violence rating based on pairwise comparison. *IEICE Transactions on Information and Systems*, E103(12):2578–2589, 2020.

Refereed International Conference Papers

1. **Ying Ji**, Yu Wang, and Jien Kato. Spatial-temporal concept based explanation of 3D CNNs. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15444–15453, 2023.

2. Yinan Yang, Yu Wang, **Ying Ji**, Heng Qi, and Jien Kato. One-shot network pruning at initialization with discriminative image patches. In *Proceedings of the 33rd British Machine Vision Conference*, 715, 14 pages, 2022.

3. **Ying Ji**, Yu Wang, and Jien Kato. Visual violence rating with pairwise comparison. In *Proceedings of the 2019 IEEE International Conference on Image Processing*, pages 3332–3336, 2019.

Domestic Meetings in Japan

1. 木場竣哉, **吉嶺**, 劉家慶, 王彧, 加藤ジェーン. 野球中継における投球動作の詳細認識. 第26回画像の認識・理解シンポジウム, IS1-110, 2023.

2. Yinan Yang, **Ying Ji**, Yu Wang, and Jien Kato. Discriminative data matters: Enhancing one-shot pruning at initialization to avoid layer collapse. 第26回画像の認識・理解シンポジウム, OS6B-S2, 2023.

3. Yinan Yang, **Ying Ji**, Yu Wang, Heng Qi, and Jien Kato. Content Matters: Concept-based explainable network pruning. 第25回画像の認識・理解シンポジウム, IS3-39, 2022.

4. **Ying Ji**, Yu Wang, Kensaku Mori, and Jien Kato. Supervoxel-based explanation for action recognition. 電子情報通信学会技術研究報告, PRMU2021-42, 2021.

5. 杉山瑠菜, **吉嶺**, 王彧, 加藤ジェーン. 時空間TCAVを用いた社会関係認識タスクにおける映像認識モデルの判断根拠分析. 電子情報通信学会技術研究報告, PRMU2021-45, 2021.

6. **Ying Ji**, Yu Wang, Jien Kato, and Kensaku Mori. Violence rating prediction with rank learning. 第22回画像の認識・理解シンポジウム, PS2-33, 2019.

7. **Ying Ji**, Yu Wang, and Jien Kato. A fine-grained dataset for visual violence analysis. 第21回画像の認識・理解シンポジウム, PS3-68, 2018.

Bibliography

[1] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. Understanding and predicting interestingness of videos. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, volume 1, pages 1113–1119, 2013.

[2] Mihai Gabriel Constantin, Liviu-Daniel Ştefan, Bogdan Ionescu, Ngoc QK Duong, Claire-Héléne Demarty, and Mats Sjöberg. Visual interestingness prediction: A benchmark framework and literature review. *International Journal of Computer Vision*, 129:1526–1550, 2021.

[3] Alexey Drutsa, Viktoriya Farafonova, Valentina Fedorova, Olga Megorskaya, Evfrosiniya Zerminova, and Olga Zhilinskaya. Practice of efficient data collection via crowdsourcing at large-scale. *Computing Research Repository arXiv Preprint*, arXiv:1912.04444, 2019.

[4] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[5] Dapeng Tao, Jun Cheng, Zhengtao Yu, Kun Yue, and Lizhen Wang. Domain-weighted majority voting for crowdsourcing. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1):163–174, 2018.

[6] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, 2008.

[7] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. CDnet 2014: An expanded change detection benchmark dataset. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 387–394, 2014.

[8] Yukino Baba and Hisashi Kashima. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 554–562, 2013.

[9] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2019.

[10] Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781, 2020.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[12] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7210–7218, 2018.

[13] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[14] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3D object detection for autonomous driving. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2156, 2016.

[15] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 221–230, 2017.

[16] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *Proceedings of the 13th International Conference on Computer Vision*, pages 1995–2002, 2011.

[17] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2663–2672, 2017.

[18] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019.

[19] Yukun Su, Guosheng Lin, Jinhui Zhu, and Qingyao Wu. Human interaction learning on 3D skeleton point clouds for video violence recognition. In *Proceedings of the 16th European Conference on Computer Vision*, volume 16, pages 74–90, 2020.

[20] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015.

[21] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *Proceedings of the 16th IEEE International Conference on Computer Vision*, pages 638–647, 2017.

[22] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19861–19869, 2022.

[23] Aditya Khosla, Wilma A Bainbridge, Antonio Torralba, and Aude Oliva. Modifying the memorability of face photographs. In *Proceedings of the 14th IEEE International Conference on Computer Vision*, pages 3200–3207, 2013.

[24] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the 15th IEEE International Conference on Computer Vision*, pages 2390–2398, 2015.

[25] Qi Kuang, Xin Jin, Qinping Zhao, and Bin Zhou. Deep multimodality learning for UAV video aesthetic quality assessment. *IEEE Transactions on Multimedia*, 22(10):2623–2634, 2019.

[26] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Proceedings of the 16th European Conference on Computer Vision*, volume 16, pages 223–240, 2020.

[27] Romain Cohendet, Karthik Yadati, Ngoc QK Duong, and Claire-Hélène Demarty. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval*, pages 178–186, 2018.

[28] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[29] Luke Palmer, Alina Bialkowski, Gabriel J Brostow, Jonas Ambeck-Madsen, and Nilli Lavie. Predicting the perceptual demands of urban driving with video regression. In *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision*, pages 409–417, 2017.

[30] Shuai Wang, Shizhe Chen, Jinming Zhao, and Qin Jin. Video interestingness prediction based on ranking model. In *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and the 1st Workshop on Multi-Modal Affective Computing of Large-Scale Multimedia Data*, pages 55–61, 2018.

[31] Yuesong Shen, Claire-Héiène Demarty, and Ngoc QK Duong. Deep learning for multimodal-based video interestingness prediction. In *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo*, pages 1003–1008, 2017.

[32] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 17th ACM International Conference on Multimodal Interaction*, pages 467–474, 2015.

[33] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450, 2016.

[34] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2011.

[35] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, 10(1):60–75, 2017.

[36] Sacide Kalayci, Hazim Kemal Ekenel, and Hatice Gunes. Automatic analysis of facial attractiveness from video. In *Proceedings of the 2014 IEEE International Conference on Image Processing*, pages 4191–4195, 2014.

[37] Jwan Saeed and Adnan Mohsin Abdulazeez. Facial beauty prediction and analysis based on deep convolutional neural network: A review. *Journal of Soft Computing and Data Mining*, 2(1):1–12, 2021.

[38] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

[39] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *Computing Research Repository arXiv Preprint*, arXiv:1506.03365, 2015.

[40] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the 16th IEEE International Conference on Computer Vision*, pages 843–852, 2017.

[41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *Computing Research Repository arXiv Preprint*, arXiv:1212.0402, 2012.

[43] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European Conference on Computer Vision*, volume 3, pages 288–301, 2006.

[44] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012.

[45] Devi Parikh and Kristen Grauman. Relative attributes. In *Proceedings of the 13th International Conference on Computer Vision*, pages 503–510, 2011.

[46] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill: A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, 19:569–576, 2007.

[47] David R Hunter. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, 32(1):384–406, 2004.

[48] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *Proceedings of the 14th European Conference on Computer Vision*, volume 1, pages 196–212, 2016.

[49] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Hipster wars: Discovering elements of fashion styles. In *Proceedings of the 13th European Conference on Computer Vision*, volume 1, pages 472–488, 2014.

[50] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 1952.

[51] Mark E Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 48(3):377–394, 1999.

[52] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. In *Technical Report*. Stanford InfoLab, Stanford, CA, USA, 1999-66, 1999.

[53] Mykola Ponomarenko, Sheyda Ghanbaralizadeh Bahnemiri, Karen Egiazarian, Oleg Ieremeiev, Vladimir Lukin, Veli-Tapani Peltoketo, and Jussi Hakala. Color image database HTID for verification of no-reference metrics: Peculiarities and preliminary results. In *Proceedings of the 9th European Workshop on Visual Information Processing*, pages 1–6, 2021.

[54] Shiyu Li, Hao Ma, and Xiangyu Hu. Neural image beauty predictor based on bradley-terry model. *Computing Research Repository arXiv Preprint*, arXiv:2111.10127, 2021.

[55] Rémi Coulom. Whole-history rating: A Bayesian rating system for players of time-varying strength. In *Proceedings of the 2008 International Conference on Computers and Games*, pages 113–124, 2008.

[56] Rui Zhao, Ruoqi Dang, and Yinliang Zhao. A neural network go rating model considering winning rate. In *Proceedings of the 3rd International Conference on Computer Science and Artificial Intelligence*, pages 23–27, 2019.

[57] Ilya Makarov, Dmitry Savostyanov, Boris Litvyakov, and Dmitry I Ignatov. Predicting winning team and probabilistic ratings in "Dota 2" and "Counter-Strike: Global Offensive" video games. In *Proceedings of the 6th International Joint Conference on the Analysis of Images, Social Networks and Texts*, pages 183–196, 2018.

[58] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[59] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the 14th IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[60] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2–3):107–123, 2005.

[61] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[62] 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, 2012.

[63] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the 15th IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[64] Fei Gao, Jun Yu, Suguo Zhu, Qingming Huang, and Qi Tian. Blind image quality prediction by exploiting multi-level deep representations. *Pattern Recognition*, 81:432–442, 2018.

[65] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 27:568–576, 2014.

[66] Weiwei Song, Shutao Li, Leyuan Fang, and Ting Lu. Hyperspectral image classification with deep feature fusion network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3173–3184, 2018.

[67] Kui Jiang, Zhongyuan Wang, Peng Yi, Guangcheng Wang, Ke Gu, and Junjun Jiang. ATMFN: Adaptive-Threshold-based Multi-model Fusion Network for compressed face hallucination. *IEEE Transactions on Multimedia*, 22(10):2734–2747, 2019.

[68] Guangxia Xu, Weifeng Li, and Jun Liu. A social emotion classification approach using multi-model fusion. *Future Generation Computer Systems*, 102:347–356, 2020.

[69] Souhail Bakkali, Zuheng Ming, Mickaël Coustaty, and Marçal Rusiñol. Visual and textual deep feature fusion for document image classification. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 562–563, 2020.

[70] Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng. An enhanced deep feature representation for person re-identification. In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision*, pages 1–8, 2016.

[71] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[72] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *Proceedings of Machine Learning Research*, 80:2668–2677, 2018.

[73] 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 *Proceedings of the 16th IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[74] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018.

[75] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

[76] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision*, pages 2950–2958, 2019.

[77] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the 16th IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[78] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of black-box models. *Computing Research Repository arXiv Preprint*, arXiv:1806.07421, 2018.

[79] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *Computing Research Repository arXiv Preprint*, arXiv:1702.04595, 2017.

[80] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in Neural Information Processing Systems*, 30:6970–6979, 2017.

[81] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. *Computing Research Repository arXiv Preprint arXiv:1902.03129*, 2019.

[82] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (CACE). *Computing Research Repository arXiv Preprint*, arXiv:1907.07165, 2019.

[83] Yunhao Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyan Wu. A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2195–2204, 2021.

[84] Vignesh Srinivasan, Sebastian Lapuschkin, Cornelius Hellge, Klaus-Robert Müller, and Wojciech Samek. Interpretable human action recognition in compressed domain. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1692–1696, 2017.

[85] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[86] Thomas Hartley, Kirill Sidorov, Christopher Willis, and David Marshall. SWAG-V: Explanations for video using Superpixels Weighted by Average Gradients. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 604–613, 2022.

[87] Thomas Hartley, Kirill Sidorov, Christopher Willis, and David Marshall. SWAG: Superpixels Weighted by Average Gradients for explanations of CNNs. In *Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 423–432, 2021.

[88] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1120–1129, 2021.

[89] Victoria Rideout, Alanna Peebles, Supreet Mann, and Michael B Robb. *2021 The Common Sense Census: Media Use by Tweens and Teens*. Common Sense Media, San Francisco, CA, USA, 2021.

[90] Victoria Rideout and Michael B Robb. *2017 The Common Sense Census: Media Use by Kids Age Zero to Eight*. Common Sense Media, San Francisco, CA, USA, 2017.

[91] Victoria Rideout. *2015 The Common Sense Census: Media Use by Tweens and Teens*. Common Sense Media, San Francisco, CA, USA, 2015.

[92] Craig A Anderson, Brad J Bushman, Bruce D Bartholow, Joanne Cantor, Dimitri Christakis, Sarah M Coyne, Edward Donnerstein, Jeanne Funk Brockmyer, Douglas A Gentile, C Shawn Green, Rowell Huesmann, Tom Hummer, Barbara Krahe, Victor C Strasburger, Warburton Wayne, Wilson Barbara J, and Ybarra Michele. Screen violence and youth behavior. *Pediatrics*, 140(Supplement 2):S142–S147, 2017.

[93] Leonard D Eron, L Rowell Huesmann, Monroe M Lefkowitz, and Leopold O Walder. Does television violence cause aggression? *American Psychologist*, 27(4):253, 1972.

[94] Craig A Anderson, Akira Sakamoto, Douglas A Gentile, Nobuko Ihori, Akiko Shibuya, Shintaro Yukawa, Mayumi Naito, and Kumiko Kobayashi. Longitudinal effects of violent video games on aggression in Japan and the United States. *Pediatrics*, 122(5):e1067–e1072, 2008.

[95] L Rowell Huesmann, Jessica Moise-Titus, Cheryl-Lynn Podolski, and Leonard D Eron. Longitudinal relations between children's exposure to TV violence and their aggressive and violent behavior in young adulthood: 1977–1992. *Developmental Psychology*, 39(2):201–221, 2003.

[96] Sergio Benini, Luca Canini, and Riccardo Leonardi. A connotative space for supporting movie affective recommendation. *IEEE Transactions on Multimedia*, 13(6):1356–1370, 2011.

[97] Yipeng Zhou, Jiqiang Wu, Terence H Chan, Siu-Wai Ho, Dah-Ming Chiu, and Di Wu. Interpreting video recommendation mechanisms by mining view count traces. *IEEE Transactions on Multimedia*, 20(8):2153–2165, 2017.

[98] Ahmad Babaeian Jelodar, David Paulius, and Yu Sun. Long activity video understanding using functional object-oriented network. *IEEE Transactions on Multimedia*, 21(7):1813–1824, 2018.

[99] Ming-yu Chen and Alexander Hauptmann. MoSIFT: Recognizing human actions in surveillance videos. In *Technical Report*. Carnegie Mellon University, Pittsburgh, PA, USA, CMU-CS-09-161, 1995.

[100] Fillipe DM De Souza, Guillermo C Chavez, Eduardo A do Valle Jr, and Arnaldo de A Araújo. Violence detection in video using spatio-temporal features. In *Proceedings of the 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, pages 224–230, 2010.

[101] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2012.

[102] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 197–206, 2007.

[103] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[104] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns*, volume 2, pages 332–339, 2011.

[105] Markus Schedl, Mats Sjöberg, Ionut Mironica, Bogdan Ionescu, Vu Lam Quang, and Yu-Gang Jiang. VSD2014: A dataset for violent scenes detection in Hollywood movies and Web videos. *Sixth Sense*, 6(2.00):12–40, 2015.

[106] Georgios Petkos, Symeon Papadopoulos, Vasileios Mezaris, and Yiannis Kompatsiaris. Social Event Detection at MediaEval 2014: Challenges, Datasets, and Evaluation. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, 2014.

[107] Liam Grealy, Catherine Driscoll, and Kirsten Cather. A history of age-based film classification in Japan. *Japan Forum*, 34(4):443–468, 2020.

[108] Douglas A Gentile. The rating systems for media products. *Handbook of Children, Media, and Development*, pages 527–551, Blackwell Publishing, Oxford, England, UK. 2008.

[109] Lucille Jenkins, Theresa Webb, Nick Browne, Abdelmonem A Afifi, and Jess Kraus. An evaluation of the motion picture association of America's treatment of violence in PG-, PG-13–, and R-rated films. *Pediatrics*, 115(5):e512–e517, 2005.

[110] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. Violence detection in movies. In *Proceedings of the 8th International Conference on Computer Graphics, Imaging and Visualization*, pages 119–124, 2011.

[111] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis, and Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In *Proceedings of the 2010 Hellenic Conference on Artificial Intelligence*, pages 91–100, 2010.

[112] Ankur Datta, Mubarak Shah, and N Da Vitoria Lobo. Person-on-person violence detection in video data. In *Proceedings of the 2002 International Conference on Pattern Recognition*, volume 1, pages 433–438, 2002.

[113] Diego Castán, Mario Rodríguez, Alfonso Ortega, Carlos Orrite, and Eduardo Lleida. ViVoLab and CVLab-MediaEval 2014: Violent scenes detection affect task. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, 2014.

[114] Vu Lam, Sang Phan, Duy-Dinh Le, Duc Anh Duong, and Shin'ichi Satoh. Evaluation of multiple features for violent scenes detection. *Multimedia Tools and Applications*, 76(5):7041–7065, 2017.

[115] Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-Gang Jiang. Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[116] Qi Dai, Zuxuan Wu, Yu-Gang Jiang, Xiangyang Xue, and Jinhui Tang. Fudan-NJUST at MediaEval 2014: Violent scenes detection using deep neural networks. In *Working Notes Proceedings of MediaEval 2014 Workshop*, 2014.

[117] Omar Seddati, Emre Kulah, Gueorgui Pironkov, Stéphane Dupont, Saïd Mahmoudi, and Thierry Dutoit. UMons at MediaEval 2015 Affective impact of movies task including violent scenes detection. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[118] Xirong Li, Yujia Huo, Qin Jin, and Jieping Xu. Detecting violence in video using subclasses. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 586–590, 2016.

[119] Zhihong Dong, Jie Qin, and Yunhong Wang. Multi-stream deep networks for person to person violence detection in videos. In *Proceedings of the 7th Chinese Conference on Pattern Recognition*, volume 1, pages 517–531, 2016.

[120] Alex Hanson, Koutilya PNVR, Sanjukta Krishnagopal, and Larry Davis. Bidirectional convolutional LSTM for the detection of violence in videos. In *Proceedings of the 15th European Conference on Computer Vision Workshops*, volume 2, pages 280–295, 2018.

[121] Evlampios Apostolidis and Vasileios Mezaris. Fast shot segmentation combining global and local visual descriptors. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6583–6587, 2014.

[122] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[123] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository arXiv Preprint*, arXiv:1409.1556, 2014.

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

[125] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the 14th European Conference on Computer Vision*, volume 8, pages 20–36, 2016.

[126] Yifan Wang, Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Two-Stream SR-CNNs for action recognition in videos. In *Proceedings of the 2016 British Machine Vision Conference*, 2016.

[127] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[128] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *Proceedings of the 9th European Conference on Computer Vision*, volume 2, pages 428–441, 2006.

[129] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. *Proceedings of Machine Learning Research*, 32(1):647–655, 2014.

[130] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in Neural Information Processing Systems*, 27:487–495, 2014.

[131] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the 13th European Conference on Computer Vision*, volume 1, pages 818–833, 2014.

[132] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 435–444, 2017.

[133] Daphne Blunt Bugental. Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin*, 126(2):187, 2000.

[134] Arushi Goel, Keng Teck Ma, and Cheston Tan. An end-to-end network for generating social relationship graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11195, 2019.

[135] Mengyin Wang, Xiaoyu Du, Xiangbo Shu, Xun Wang, and Jinhui Tang. Deep supervised feature selection for social relationship recognition. *Pattern Recognition Letters*, 138:410–416, 2020.

[136] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3566–3574, 2019.

[137] Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, and Huadong Ma. Multi-stream fusion model for social relation recognition from videos. In *Proceedings of the 23rd International Conference on Multimedia Modeling*, volume 1, pages 355–368, 2018.

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

[139] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[140] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. LAEO-Net: Revisiting people looking at each other in videos. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019.

[141] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[142] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11390–11399, 2021.

[143] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision*, pages 10143–10152, 2019.

[144] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. EmotiCon: Context-aware multimodal emotion recognition using Frege's principle. In *Proceedings of the 2020*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14234–14243, 2020.

[145] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019.

[146] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 839–848, 2020.

[147] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. GroupFormer: Group activity recognition with clustered spatial-temporal transformer. In *Proceedings of the 18th IEEE/CVF International Conference on Computer Vision*, pages 13668–13677, 2021.

[148] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the 16th IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[149] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 8581–8590, 2018.

[150] Donald J Kiesler. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90(3):185, 1983.

[151] David YF Ho. Interpersonal relationships and relationship dominance: An analysis based on methodological relationism. *Asian Journal of Social Psychology*, 1(1):1–16, 1998.

[152] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, Mehdi Mirza, Sébastien Jean, Pierre-Luc Carrier, Yann Dauphin, Nicolas Boulanger-Lewandowski,

Abhishek Aggarwal, Jeremie Zumer, Pascal Lamblin, Jean-Philippe Raymond, Guillaume Desjardins, Razvan Pascanu, David Warde-Farley, Atousa Torabi, Arjun Sharma, Emmanuel Bengio, Myriam Côté, Kishore Reddy Konda, and Zhenzhou Wu. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 543–550, 2013.

[153] Xun Gao, Yin Zhao, Jie Zhang, and Longjun Cai. Pairwise emotional relationship recognition in drama videos: Dataset and benchmark. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3380–3389, 2021.

[154] Kaoru Toyoda, Yoshihiro Miyakoshi, Ryosuke Yamanishi, and Shohei Kato. Dialogue mood estimation focusing on intervals of utterance state. *Transactions of the Japanese Society for Artificial Intelligence*, 27(2):16–21, 2012.

[155] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[156] Ying Ji, Yu Wang, and Jien Kato. Spatial-temporal concept based explanation of 3D ConvNets. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15444–15453, 2023.

[157] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *Proceedings of the 13th IEEE International Conference on Computer Vision*, pages 2556–2563, 2011.

[158] Long D Nguyen, Dongyun Lin, Zhiping Lin, and Jiuwen Cao. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In *Proceedings of the 2018 IEEE International Symposium on Circuits and Systems*, pages 1–5, 2018.

[159] Zihe Liu, Weiying Hou, Jiayi Zhang, Chenyu Cao, and Bin Wu. A multimodal approach for multiple-relation extraction in videos. *Multimedia Tools and Applications*, 81(4):1–26, 2022.

[160] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the 15th IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[161] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[162] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[163] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[164] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the 15th European Conference on Computer Vision*, volume 8, pages 325–341, 2018.

[165] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[166] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015.

[167] Jan Macdonald, Mathieu E. Besançon, and Sebastian Pokutta. Interpretable neural networks with Frank-Wolfe: Sparse relevance maps and relevance

orderings. *Proceedings of Machine Learning Research*, 162:14699–14716, 2022.

[168] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-MIR: Explainable Medical Image Retrieval. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 440–450, 2022.

[169] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[170] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[171] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[172] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9097–9107, 2019.

[173] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Remco Veltkamp, and Ronald Poppe. Saliency Tubes: Visual explanations for spatio-temporal convolutions. In *Proceedings of the 2019 IEEE International Conference on Image Processing*, pages 1830–1834, 2019.

[174] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the Kinetics-700 human action dataset. *Computing Research Repository arXiv Preprint*, arXiv:1907.06987, 2019.

[175] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 32–36, 2004.

[176] Yilin Wang, Suhang Wang, Jiliang Tang, Neil O'Hare, Yi Chang, and Baoxin Li. Hierarchical attention network for action recognition in videos. *Computing Research Repository arXiv Preprint*, arXiv:1607.06416, 2016.

[177] Hao Yang, Chunfeng Yuan, Li Zhang, Yunda Sun, Weiming Hu, and Stephen J Maybank. STA-CNN: Convolutional spatial-temporal attention learning for action recognition. *IEEE Transactions on Image Processing*, 29:5783–5793, 2020.

[178] Murong Wang, Xiabi Liu, Yixuan Gao, Xiao Ma, and Nouman Q Soomro. Superpixel segmentation: A benchmark. *Signal Processing: Image Communication*, 56:28–39, 2017.

[179] Feng Ge, Song Wang, and Tiecheng Liu. New benchmark for image segmentation evaluation. *Journal of Electronic Imaging*, 16(3):033011, 2007.

[180] Yiqing Liu, Tao Zhang, and Zhen Li. 3DCNN-based real-time driver fatigue behavior detection in urban rail transit. *IEEE Access*, 7:144648–144662, 2019.