PAPER

Predicting Violence Rating Based on Pairwise Comparison*

Ying JI†, Yu WANG††, Nonmembers, Jien KATO††, Member, and Kensaku MORI†, Nonmember

SUMMARY With the rapid development of multimedia, violent video can be easily accessed in games, movies, websites, and so on. Identifying violent videos and rating violence extent is of great importance to media filtering and children protection. Many previous studies only address the problems of violence scene detection and violent action recognition, yet violence rating problem is still not solved. In this paper, we present a novel video-level rating prediction method to estimate violence extent automatically. It has two main characteristics: (1) a two-stream network is fine-tuned to construct effective representations of violent videos; (2) a violence rating prediction machine is designed to learn the strength relationship among different videos. Furthermore, we present a novel violent video dataset with a total of 1,930 human-involved violent videos designed for violence rating analysis. Each video is annotated with 6 fine-grained objective attributes, which are considered to be closely related to violence extent. The ground-truth of violence rating is given by pairwise comparison method. The dataset is evaluated in both stability and convergence. Experiment results on this dataset demonstrate the effectiveness of our method compared with the state-of-art classification methods.

key words: visual violence analysis, pairwise comparison, rating estimation

1. Introduction

With the advent of multimedia, social media usage among adolescents has increased dramatically. According to Common Sense census [1], [2], children aged 0 to 8 spend more than 3 hours with screen media everyday. Children aged 13 to 18 spend about 9 hours with media, including television, video games, and Internet. However, many videos released on TV or Internet are unsuitable for children to watch, since they may include violent, bloody or adult contents. These videos may lead to a bad influence on children’s behaviours. A number of studies have proven that increasing violent media exposure will result in not only short-term but also long-term harmful effects on youth [3]. Eron [4] and Anderson’s [5] 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 behaviour in their later life, including criminal behaviour, spousal abuse, and assault [6]. Therefore, recognition of violent video is of the essence in multimedia recommendation [7], [8] and multimedia content understanding [9].

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, including either local features or global features. Chen [10] extracts Motion SIFT (MoSIFT) to detect distinctive local features by combining local appearance and temporal information in surveillance video. De Souza [11] presents a violence detector Space-Time Interest Points (STIP [12]) mainly in sports videos. Hassner [13] extracts 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 [14]. Finally, these features will be used for classification by using a linear Support Vector Machine.

Audio information is also extracted as supplementation to visual information. Audio features usually occur in specific scenes as unique signals, such as gunshots, screams or explosions. Lin and Wang [15] apply pLSA method to detect flame and explosion event in violent videos. Blood areas are also detected by calculating motion density. Both the audio features and visual features are used to co-train a classifier. Derbas [16] builds a joint audio-visual codebook that reveals the strong relationship between audio representation and visual representation. Except for visual and audio features, [17] extracts color emotional features based on psychological experiments innovatively.

Existing visual violence researches are only limited to scene-level detection, instead of 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, different from previous objective detection, our research focuses on subjective violence rating prediction.

However, such a requirement can not be met by the existing datasets, none of which provides violence extent label. Previous datasets are only confined to simplex scenes, such as sports video or surveillance video. The number of videos and the varieties of these datasets are also limited. Hockey fight dataset [18], for example, contains 500 violent videos. Crowd violence dataset [13] contains 123 violent videos. And these datasets only divide videos into two classes: violent and non-violent. They lack detailed infor-
mation and can’t distinguish the violence extent of videos. To this end, we also provide a novel dataset with violence extent labels. This paper extends the work in [19] with additional results and analysis. In dataset construction, we perform new experiments to evaluate the validity of Trueskill in violence rating task, and show the influence of comparison orders and comparison times to violence levels. A reliable rating is given with proof. We also present new experiments using Resnet-50 when evaluating our proposed method. We further analyse the effectiveness and limitations of our approach by visualizing prediction results.

Our major contributions are as follows:

(1) Designed for violence rating analysis, we construct a fine-grained violent video dataset called Human Violence Dataset. The dataset consists of 1,930 human-involved violent videos collected from YouTube movie trailers. Each video has 6 objective violent annotations. The subjective violence levels are given by pairwise comparison method. The stability and convergence of Trueskill algorithm in Human Violence Dataset are proved;

(2) A two-step method is developed for violence rating prediction. A two-stream neural network is fine-tuned on Human Violence Dataset and used to extract features for each video. By using different pooling and normalization methods, we assess various representations of two-stream features and validate the advantage of average pooling. Videos are represented by the best combined two-stream features. Then a rating estimation machine is proposed to learn the level relationship between different violent videos. We experimentally show that our method can predict violence rating better than classification methods. The visualization of feature maps and prediction results are also presented.

The rest of this paper is organized as follows: Section 2 reviews some related works. Section 3 introduces the dataset and evaluates the stability and convergence of the dataset. Section 4 proposes the violence rating method and experiment results are shown in Sect. 5. Finally, Sect. 6 shows the conclusion.

2. Related Work

2.1 Violent Video Dataset

In the literature, there already exist some datasets for violent video analysis. Hockey fight dataset [18], crowd violence dataset [13] and violent scenes dataset (VSD 2014) [20], for example, are the most commonly used datasets. Detailed information of three datasets are summarized in Table 1.

| Dataset                | Year  | Annotations               | Clips   | Resource          | # of violent videos | Violent level |
|------------------------|-------|---------------------------|---------|-------------------|---------------------|--------------|
| Hockey Fight Dataset   | 2011  | Fight / non-fight         | 1,000   | Hockey games      | 500                 | ×            |
| Crowd Violence Dataset | 2012  | Violent / non-violent     | 246     | YouTube           | 123                 | ×            |
| VSD 2014 [20]         | 2014  | 7 visual + 3 audio labels| 31 movies | Hollywood Movies | 15% violent scenes | ×            |
| our Human Violence dataset | 2019 | 6 objective labels + violence level | 1,930 | Promotion videos | 1,930 | ✓ |

2.1.1 Hockey Fight Dataset

Nievas [18] creates a dataset containing 1,000 short clips collected from hockey games of the National Hockey League (NHL). All the videos are labelled as 2 categories: “fight” or “non-fight”. There are totally 500 violent videos and 500 non-violent videos. Each video clip contains 50 frames with a resolution of 720 × 576 pixels.

2.1.2 Crowd Violence Dataset

This dataset is built for crowd violence detection in video surveillance system by Hassner [13] in 2012. 246 real-world video clips are collected from YouTube 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 others are non-violent.

2.1.3 Violent Scenes Dataset

In 2014, Schedl [20] produces a dataset for violent scene detection based on Hollywood Movies. This dataset is composed of 3 parts: (1) Hollywood training dataset; (2) Hollywood test dataset; (3) YouTube dataset. The Hollywood dataset contains 24 movies (totally 50 hours and 2 seconds) and 7 movies (totally 13 hours and 53 minutes) respectively. YouTube dataset contains 86 video clips (totally 2 hours and 3 minutes). Hollywood datasets have around 15 % violent scenes. YouTube dataset has around 44 % violent scenes. This Violent Scenes dataset has been used in MediaEval 2014 workshop. Many researches use this dataset for violent scene detection tasks.

2.2 Media Rating Systems

In order to protect children and provide appropriate media for different age groups, many countries establish organizations for media rating, including film, game and music rating.

2.2.1 Eirin

Eirin is the abbreviated name for Film Classification and Rating Organization in Japan, an independent and non-governmental organization. It was established in 1949 and the classification criteria changed several times over the years [21]. Currently, films in Japan are classified into 4
classes: G (suitable for all ages), PG12 (parental guidance requested for teenagers under twelve), R15+ (restricted to teenagers over fifteen), R18+ (only restricted to person aged 18 and above).

2.2.2 MPAA

MPAA represents the Motion Picture Association of America, who works for film classification and rating in the United States. The purpose of it is to provide parents with information about movies so that parents can decide whether the movie can be watched by their children or not [22]. Different from Eirin in Japan, MPAA classifies films into 5 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), NC-17 (no one under 17 admitted). Primary factors that may influence the rating include violence, language, theme, drug abuse, sensuality and nudity [23].

2.3 Subjective Attribute Recognition

In recent years, many methods are proposed to analyse subjective attributes such as movie popularity [24]–[26], image aesthetics [27] and face attractiveness [28]. In early research, the ground-truth of subjective properties is given by multi-annotators with discrete numeric labels. In [29], to evaluate the originality of a picture, each picture is annotated with a score ranging from one to seven. In [30], each picture is annotated with about 210 votes on the degree of aesthetic.

However, there are some disadvantages when using single value annotation for subjective attributes. Due to the subjectivity of the properties, the boundaries between different scores are unclear, making individual rating label difficult. On the other hand, judgement criteria are different from people to people, which increases the unreliability of final result. To solve these problems, pairwise comparison method is widely used in recent subjective attributes analysis. The annotators are given two images and asked about which image is better or worse on the properties. The common pipeline to provide a rank list is: (1) collecting abundant pairwise comparisons; (2) employing rating method [31], [32] to get the rank.

In [33], 10 assessors are involved in pairwise comparison procedure to predict the interestingness of videos. Dubey et al. [34] quantify the safety and beauty extent of urban environment through collecting crowd-sourcing pairwise comparisons. Kiapour et al. [35] use Amazon Mechanical Turk to gather judgement for clothing style recognition. More recently, pairwise comparison method is also used to predict human perception when driving [36].

2.4 Visual Violence Analysis

Violence detection is not a novel problem, violence was mainly considered as flame or blood flow in previous researches. Many researches focus on explosion or blood detection. [37] proposes a bloody frame detection approach to determine violent scene in movie. [38] extracts 12 kinds of audio features to detect audio violence, such as shots and screams. Later studies begin to focus on detecting violent interaction behaviours, for instance, fight action between people. [11] extracts Space-Time Interest Points (STIP) to distinguish violence activity from regular activities in sports video. Datta et al. define an Acceleration Measure Vector (AMV) to detect human violence in video, such as fist fighting, kicking or hitting [39]. Improved dense trajectories [40] is also a widely used feature when detecting violence motion [41], [42].

More recently, researchers pay more attention to deep learning methods when detecting violence. In MediaEval challenges, [43]–[45] utilize convolutional neural network (CNN) to extract features. [46] combines CNN features, audio features and motion features to represent violent video. Several researches also use convolutional long short term memory (convLSTM) for violence detection. [47] and [48] use convLSTM to capture spatio-temporal features.

Different from previous datasets which contain both violent and non-violent videos, our work focus on providing a dataset only containing violent videos. We think the number of violent videos in previous datasets are not enough. In our case, 1,930 violent video clips are collected. More specifically, in order to figure out the cause of violence, each video is labelled with 6 objective attributes. Violence rating of each video is also studied particularly. Table 1 also compares our dataset with existing violent video datasets.

Existing media rating systems take both visual and auditory information into consideration, while we only focus on visual information. Moreover, our research objects are short movie clips without context and story. In this case, 4 or 5 categories are too precise for our research. 3 categories are enough for visual distinguishment. Videos in our dataset are divided into 3 violence levels. Since violence extent is also a kind of subjective attribute, we will also employ pairwise comparison method to provide ground-truth violence extent.

![Image](image_url)

**Fig. 1** With a group of violent videos as input, we rank the videos according to the violence extent. The final output of our research is violence rating of each video.
for each video. Different from previous violent scene detection approaches, we do not detect exact actions or scenes in a violent video. As shown in Fig. 1, by contrast, with a video clip as input, our output is the violence extent of the video. Considering the outstanding performance of CNN in violence detection, we use a two-stream network to extract features from violent videos. Most existing researches treat violence recognition as a classification problem, while in our case, we innovatively learn the relationship between different violence levels.

3. Human Violence Dataset

The current dataset contains 1,930 violent video clips collected from movie promotion videos on YouTube. Each clip has a length of 30 to 100 frames. The frame rate and resolution of video clips are 30 FPS and 1,280 × 720 respectively. Additionally, each clip is manually labelled with 6 objective fine-grained visual attributes. Furthermore, by utilizing pairwise comparison method, each video is also annotated with one subjective violence extent label. In the following, we will describe the details of the dataset creation process.

3.1 Data Collection

In order to collect videos with various violent scenes and actions, movie promotion videos are chosen as raw video content because it contains multiple scenes in short time periods. We begin the collection by selecting action movies released in recent ten years. We then download the movie trailers published on YouTube by the corresponding official movie companies. Totally 1,020 raw videos are gathered. The duration of each video is around 2 to 3 minutes. Considering the complexity of violence extent, we aim to build a dataset in which each video clip contains only single scene and complete action. In this case, we apply the segmentation tool [49] to divide the videos into over 25,000 shots. We further manually exclude very short clips, clips with multiple scenes, as well as those clips without human involved violence.

A total of 1,930 human-involved violent video clips are collected. During labelling process, there also includes some other forms of violence without human-involved, such as firearms, fire or explosion. However, we identify these objective violence attributes have been well-studied in prior researches. They are also very consistent in different videos. By contrast, we observe that most of the violence are accompanied by human interactions. With different interactions between people, violence extent of videos become different from each other. Sometimes even subtle differences may lead to completely different violence levels of videos. Considering these challenges, our work only focuses on human-involved visual violence.

3.2 Objective Violent Attribute

Annotations of Human Violence Dataset include two parts: objective violence attributes and subjective violence rating. The selected 6 objective attributes are closely related to violence extent. The annotation procedure is finished by two annotators independently. The following part explains the definition of the attributes.

**Combat Mode (CM)** There are 5 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; 5) a group of people versus another group of people.

**Physical Contact (PC)** There are 2 subcategories in this attribute: 1) If a person bring a part of his/her body into contact with other person; 2) else.

**Weapon possession (WP)** There are 3 subcategories in WP category: 1) no weapon appears in the video clip; 2) there is a weapon appears, but hasn’t been used; 3) a weapon has been used.

**Weapon direction (WD)** If there appears a weapon in the video, the direction of weapons will be annotated in detail. 3 kinds of directions are given: 1) act on the opponent; 2) act towards the screen; 3) others.

**Blood** There are 3 subcategories in this attribute: 1) there is no blood in the video; 2) static blood; 3) flowing blood.

**Explosion** There are 2 subcategories in this attribute: 1) no explosion appears in the video; 2) an explosion happens in the video.

Figure 2 shows examples of violent videos and Table 2 shows the corresponding labels in our dataset.

3.3 Subjective Violence Rating

As introduced in Sect. 2.3, pairwise comparison can be used

---

**Fig. 2** Video examples and labels in 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 shots towards the screen; (f) a group of men are blown up in an explosion. Three different colors are used to represent three violence levels. Videos with green border have the lowest level of violence. Videos with purple border have the moderate level of violence. Videos with red border have the highest level of violence.
Table 2  Fine-grained violent attributes and corresponding labels of 6 video examples in 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| x         |
| Video 4    | One Vs. Group | ✓               | ✓                 | Opponent         | ×     | x         |
| Video 5    | Attacker    | ×                 | ✓                 | Act towards the screen | ×     | x         |
| Video 6    | Victim      | ×                 | ×                 | -                | ×     | ✓         |

to constrain the instability of subjective attributes rating. In our work, we employ Trueskill [31] algorithm to obtain ground-truth violence level. Trueskill is a Bayesian rating system designed for video game matchmaking originally. When using Trueskill, for each video, the violence extent will be considered as a Gaussian distribution, \( N(\mu, \sigma) \). \( \mu \) represents current estimate of the violence. \( \sigma \) represents current uncertainty of the estimate. Every time we compare two videos and decide which one is more violent. After each comparison, we will update \( \mu \) and \( \sigma \) according to the following equations. Following Herbrich [31], we use \( \mu = 25 \) and \( \sigma = 25/3 \) as initial values for each video before any comparisons.

\[
\begin{align*}
\mu_{\text{win}} & \leftarrow \mu_{\text{win}} + \frac{\sigma_{\text{win}}^2}{c} \cdot v \left( \frac{\mu_{\text{win}} - \mu_{\text{lose}}}{c}, \frac{\epsilon}{c} \right) \\
\mu_{\text{lose}} & \leftarrow \mu_{\text{lose}} + \frac{\sigma_{\text{lose}}^2}{c} \cdot v \left( \frac{\mu_{\text{win}} - \mu_{\text{lose}}}{c}, \frac{\epsilon}{c} \right) \\
\sigma_{\text{win}}^2 & \leftarrow \sigma_{\text{win}}^2 - \frac{\sigma_{\text{win}}^2}{c^2} \cdot w \left( \frac{\mu_{\text{win}} - \mu_{\text{lose}}}{c}, \frac{\epsilon}{c} \right) \\
\sigma_{\text{lose}}^2 & \leftarrow \sigma_{\text{lose}}^2 - \frac{\sigma_{\text{lose}}^2}{c^2} \cdot w \left( \frac{\mu_{\text{win}} - \mu_{\text{lose}}}{c}, \frac{\epsilon}{c} \right) \\
c^2 & = 2\beta^2 + \sigma_{\text{win}}^2 + \sigma_{\text{lose}}^2
\end{align*}
\]

After sufficient comparisons, \( \mu \) and \( \sigma \) will become stable. According to [34], [36], comparison times around 24 to 36 per video are able to provide a stable ranking. The predicted violent score for each video can be calculated as \( \mu - 3\sigma \). By sorting violent scores, we can obtain the violence extent ranking for each video clip.

When labelling violence extent, we use 1,459 videos from “have not use the weapon” and “weapon has been used are” subcategories out of WP category, because videos belong to these two categories are balanced. Each video is randomly compared to other videos 36 times without any repeating comparisons. We built a graphic user interface to compare videos efficiently, a snapshot of which is shown in Fig. 3. Each time it will show two different violent videos randomly and ask the annotator which video is more violent. The observer can only choose one violent video. The choice from the observer will be recorded.

In our experiments, totally 26,262 times different comparisons were collected in about three months by one annotator to maintain the consistency of judgement criteria. After all comparisons are done, we calculated Trueskill violent scores for each video. Figure 4 shows the histogram of violent extent scores after all comparisons. The higher score represents the higher violent extent. These videos are then averagely divided into 3 levels according to their Trueskill scores. In practice, we average Trueskill scores over five runs as the final score, after proving the stability and convergence of the Trueskill method.

3.4 Dataset Evaluation

The violence level calculated through Trueskill will be used as ground-truth label in the following experiments. In this case, we prove the stability and convergence of Trueskill algorithm in our violent video dataset, and provide a more convincing violence level label. Detailed proofs show as follows.
3.4.1 Trueskill Stability in Violent Video Dataset

We first prove the stability of Trueskill on our dataset. We assume that the change of comparison orders will not influence the final violence extent level. In Trueskill algorithm, each time only one set of video pairs \( P_i \) is compared randomly. \( L_t \) represents the violence level calculated by \( P_i \). The sequence of all comparisons can be represented as \( S = (P_1, P_2, \ldots, P_n) \), while in our case, \( n \) equals to 26,262. The final violence level calculated from sequence \( S \) can be represented as \( L \). Then we randomly change the comparison orders to \( S_i = rand(P_1, P_2, \ldots, P_n) \), and calculate the violence level \( L_k \) for each video. We investigate the violence level similarity between \( L \) and \( L_k \).

The accuracy of violent level similarity when changing comparing orders is shown in Table 3. The table includes both the accuracy of 3 separate levels and the total accuracy. We change the comparison order 4 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 independent of comparison order.

| Level | Accuracy | Accuracy | Accuracy | Accuracy |
|-------|----------|----------|----------|----------|
| 1     | 97.33%   | 96.92%   | 96.92%   | 96.92%   |
| 2     | 93.83%   | 93.83%   | 93.42%   | 93.83%   |
| 3     | 96.50%   | 96.91%   | 96.50%   | 96.91%   |
| All   | 95.89%   | 95.89%   | 95.61%   | 95.89%   |

3.4.2 Trueskill Convergence in Violent Video Dataset

We then prove the convergence of Trueskill in Human Violence dataset. We assume that with the increase of comparison times, the violence level of each video will become stable. At present, each video is compared 36 times with other videos. We then decrease the comparison times to \( t \) for each video, and calculate the violence level \( L_t \). We investigate the violence level similarity between consecutive comparison times, i.e., comparing the violence results between \( L_t \) and \( L_{t-1} \). According to previous proof, changing comparison orders will cause disturbance to the final rating. In order to get a more convincing violence level, we repeat the procedure 5 times and calculate Trueskill scores for each time. The average value of 5 times scores are treated as final score for each video.

Figure 5 shows the results. We can see that after about 28 times comparisons for per video, the violence level for each video become stable. By averaging multi-time results, the violence level is more accurate and has higher convergence rate. In this case, we take the violence level got from 5 times average as final ground-truth violence rating for following experiments. Level 1 contains 487 videos with Trueskill scores ranged from -4.7693 to 14.9794. Level 2 contains 486 videos and the corresponding scores are from 14.9841 to 23.9908. Level 3 are 486 videos ranged from 24.0114 to 37.5025.

Figure 2 also shows 6 videos examples from 3 different levels according to Trueskill scores.

4. Violent Rating Based Learning

In this section, we present a violence rating prediction method. Figure 6 illustrates the pipeline of our proposed methods. The input of our method are violent video clips, the output are predicted violence levels. Our method has two main steps: (1) we fine-tuned a two-stream network to extract features for each video; (2) a rank learning machine is trained to predict violence rating for test video.

4.1 Two-Stream Network Based Feature Extraction

Two-stream network [50] 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 spatial stream is single image frame, while the input of temporal stream is a stack of 10 optical flow frames. We fine-tune a two-stream network using our Human Violence dataset, and use it to extract features. Given a video \( x \) of size \( H \times W \times 3 \times F \), where \( H \) represents the height of the video, \( W \) represents the width of the video, \( F \) is the number of frames. For each video, the number of input frames is set as \( N \) over \( F \), making the input sizes of RGB frame and optical flow become \( H \times W \times 3 \times N \) and \( H \times W \times 10 \times N \) separately. We then extract two-stream features \( f_s \) and \( f_t \) from rectified linear unit (ReLU) layer following fully connected (FC) 7 layer. 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 layers we extracted, \( H' \) is the height of the feature map and \( W' \) is the width. They depend on video size and 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 \( f_t' \). We then continue to normalize the two-stream features and concatenate them as final representation for each video which is denoted as \( f = [N(f'_s); N(f'_t)] \). \( N \) denotes two normalization methods:

\[ S = \begin{bmatrix} f'_s \ f'_t \end{bmatrix} \]
L2 normalization and square root (SR) with L2 normalization method. This concatenated feature \( f \) will be used to predict level in the following step.

### 4.2 Rank Learning on Video Violence Rating

Right now, video \( i \) can be represented as \((f_i, l_i)\), where \( f_i \) is the two-stream feature and \( l_i \) is ground-truth violence rating. In our work, videos are divided into 3 levels: \( L_1, L_2 \) and \( L_3 \), and \( L_1 < L_2 < L_3 \). The purpose of our research is to predict the violence level \( l \) of a test video with feature \( f^* \).

In previous research, multi-class classification is the most commonly used method. However, classification methods take each sample as separate individual 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 relationship, we propose a rank machine to learn the relationship between different violent levels and predict the violent rank.

In the learning stage, we denote the training dataset which contains \( n \) videos as \( D = \{(f_i, l_i)^{n}_{i=1}\} \). 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 \( O = \{(f_i, f_j)\} \), if \( l_i > l_j \), which means video \( i \) have a higher violence level than video \( j \). Similar relationship is defined as \( S = \{(f_i, f_j)\} \), if \( l_i = l_j \), which means video \( i \) and video \( j \) are in the same violence level. Our purpose is to learn a ranking function:

\[
r(f_i) = w^T f_i
\]

\( w^T \) is a coefficient vector. This ranking function should make the maximum number of the following constrains satisfied:

\[
\forall (i, j) \in O : w^T f_i > w^T f_j
\]

\[
\forall (i, j) \in S : w^T f_i = w^T f_j
\]

Solving this problem is an NP hard problem. We follow the work done by Parikh [51] and Joachims [52]. Two non-negative slack variables \( \xi \) and \( \gamma \) which are similar to SVM are introduced to approximate the results. Equation (7) and Eq. (8) are converted into solving the following optimization problem:

\[
\text{minimize } \left( \frac{1}{2} \|w^T\|^2 + C \left( \sum \xi_{ij}^2 + \sum \gamma_{ij}^2 \right) \right)
\]

\[
s.t. \quad w^T f_i + \xi_{ij} \geq 1; \forall (i, j) \in O
\]

\[
|w^T f_i - w^T f_j| \leq \gamma_{ij}; \forall (i, j) \in S
\]

\[
\xi_{ij} \geq 0; \gamma_{ij} \geq 0
\]

where \( C \) is a trade-off constant to maintain the balance between maximizing the margin and meeting the pairwise labels margins. Newton method will be used to calculate \( w^T \).

### 4.3 Violence Rating Prediction

The ranking function \( r(f_i) \) with learned \( w^T \) can be utilized to calculate violence score for each video. Dataset \( D = \{(f_i, l_i)^{n}_{i=1}\} \) can be represented as \( D^* = \{(w^T f_i, l_i)^{n}_{i=1}\} \). In the following, we introduce 3 different methods to predict violence level with violence score. The performance of the methods will be evaluated in Sect. 5.4.

**Minimum distance prediction**

For each violence level, the ranking score with feature \( f \) can be represented as:

\[
S_k = \frac{1}{N_k} \sum_{l_i=L_k} w^T f_i, \quad k = 1, 2, 3
\]

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 \( f^* \) can be calculated as:

\[
L^* = \text{argmin}_{L_k} (w^T f^* - S_k)^2
\]

**Minimum mean distance prediction**

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

\[
F_k(w^T f_i) = N(\mu_k, \sigma_k), \quad k = 1, 2, 3
\]
where $\mu_k$ is the mean value of the Gaussian distribution, and $\sigma_k$ is the standard deviation. Given a new video with feature $f^*$, the violence level can be calculated as:

$$
L^* = \arg\min_{L_k}(w^T f^* - \mu_k)^2
$$

(16)

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

$$
L^* = \arg\max_{L_k} P(w^T f^* | \mu_k, \sigma_k)
$$

(17)

5. Experiment

5.1 Implementation

Our experiments are implemented in MATLAB. Totally 1,459 videos are used. 75%, which are 1,095 videos, are used as training data, while the rest 364 videos are used as test data. We use three networks: Alexnet [53], VGG16 [54] and ResNet-50 [55]. These networks are pre-trained on ImageNet following by fine-tuning on Human Violence Dataset. Alexnet has 5 convolutional layers and 3 fully-connected layers. ReLU is applied after all convolutional and fully-connected layers. VGG16 has 13 convolutional layers and 3 fully-connected layers. Max-pooling is performed over a 2×2 pixel window with stride 2. All hidden layers use ReLU as 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 [50].

All images are resized to 256 × 342 beforehand. When implementing VGG16 and ResNet-50, a 224 × 224 sub-image is cropped from the selected image randomly. When using Alexnet, a 227 × 227 sub-image is cropped. In the training stage, for spatial network, we randomly choose one frame from each video and resize it to the required size. For temporal network, 10 continuous optical flow frames are randomly chosen from each video. In the test stage, for spatial network, the middle frame of video is used. For temporal network, the middle 10 continuous optical flow frames are used. We start the learning rate for two streams from $10^{-3}$, and reduce it by a factor of 10 every 30 epochs until 90th epoch. When using Alexnet and VGG16, a dropout layer is added after fully connected layer for two streams. According to the good practices in the work [56], [57], dropout ratio for spatial network is set as 0.8. For temporal network, dropout ratio is set as 0.9.

5.2 Evaluation of Two-Stream Network

Recent years, deep neural network shows good performance in object classification with the development of large-scale datasets and powerful GPUs. In general, deep learning methods perform better than traditional classification methods. And with the increase of network depth, the accuracy becomes higher. We first conduct two experiments: (1) improved Dense Trajectories [40] is used to extract trajectory features; (2) a two-stream network is used for violence classification.

In improved Dense Trajectories, 4 different descriptors are computed: histogram of oriented gradient (HOG), histogram of flow (HOF), motion boundary histogram (MBH) and trajectory. In our experiment, trajectory length is set as 3 frames, and totally 402 dimensional features are calculated for each video. The dimensions are decreased to 201 by using PCA. We then use fisher vector to encode the features. Finally, we employ a linear SVM for classification. The accuracy of iDT is 49.17% and shown in Table 4.

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

The end-to-end results of two-stream network are also included in Table 4. We can see that a deeper network can better predict violence rating. Resnet-50 shows the best performance. iDT performs worse than VGG16 and Resnet-50. 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 7 visualizes the feature maps of two streams using VGG16. For each convolutional layer, we calculate the mean value of each channel and visualize the most active channel with a pseudocolor image. It can be seen that the shallower layers mainly provide edge or texture patterns, while the deeper layers can provide more discriminative information such as gun shoot.

5.3 Evaluation of Deep Features

CNN representation has been proved to be a powerful descriptor in previous researches [58], [59]. As mentioned above, deeper layers can produce more characteristic features than the shallower. The activations extracted from
fully connected layer with a linear SVM usually show outstanding performance [60]. Normally the extracted features are normalized to small scale such as 0 to 1, in order to maintain the balance and prevent numerical difficulties before classification.

In this case, we use the fine-tuned two-stream network as a feature extractor and extract features from ReLU7 layer. As mentioned in Sect. 4.1, we set \( N \) as the frame number for each video. Two pooling operations and two normalization operations are conducted on extracted features. Since violent activity is a continuous action, violent extent judgement depends on the information from all frames in a single video. Average pooling will better utilize all violence features than max pooling in our case. Finally, violence level is predicted by feeding the features into a linear SVM.

Table 4 shows the evaluation results of different pooling and normalization methods for spatial network and temporal network. For each pooling method, we concatenate the best performed normalization method in each single stream as two-stream features. The concatenated two-stream features are fed into a linear SVM. Table 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.

5.4 Evaluation of Proposed Method

We now evaluate our proposed method. The best performed two-stream features in Table 4 are used to train the rank learning machine proposed in Sect. 4.2. Three different methods are conducted to predict violence rating. The comparison results between our methods and existing methods are summarized in Table 5. It can be seen 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 and estimated as Gaussian distribution. So Gaussian distribution can better match the relationship of rating difference. However, the minimum distance prediction method performs worse. We hypothesize that this is because learning rank relationship is complicated, and directly using the rank score difference can not reflect the inner relationship.

Figure 8 shows three confusion matrix of using VGG16 network. The first is two-stream network end-to-end result. The second is the confusion matrix of extracted two-stream features with SVM. Both the spatial network and temporal network use the average pooling with L2 normalization, while the final one is our proposed maximum likelihood prediction method. The figures prove that our method can better predict violence rating in all 3 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. And videos in level 1 have a high possibility to predicted as level 2, while videos in level 3 are more likely to be judged as level 2 than level 1. We hypothesize this is because 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 9 shows some prediction examples using our proposed method. In each level, videos in the box of the same color with ground-truth level color are the successful examples. However, there are still some failure predictions. 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. We believe this is because our method does not detect exact objects in the video, so it may fuse the kettle with some other aggressive weapons, leading the attacking behaviour to become a high violence action. By observing the remaining failure examples and their Trueskill scores, we can see that the boundaries between videos which have closing scores are very unclear. Sometimes the actions are very similar, only the direction of the weapon or the frequency of using 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 matrix.

6. Conclusion

This paper proposes a novel visual violence rating prediction method. It is mainly composed of 2 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; (2) a violence rating prediction machine is learned by utilizing deep features and pairwise relationship between videos.

To evaluate our proposed method, we establish Human Violence Dataset that consists of 1,930 violent videos. Besides, 6 objective violence attributes and one subjective violence rating level are well labelled for each video. The dataset is evaluated in both stability and convergence. The experiment results on violence rating prediction show that our proposed method performs better than existing classification methods. This indicates our method better reflects the strength relationship between different violence levels and can produce more representations for violent video.

The main advantage of our proposed method over the previous studies is that we focus 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 single score evaluation. Finally, the strong interactions
References

[1] V. Rideout, “The common sense census: Media use by kids age zero to eight.” San Francisco, CA: Common Sense Media, pp.263–283, 2017.
[2] V.J. Rideout, The common sense census: Media use by tweens and teens, Common Sense Media Incorporated, 2015.
[3] C.A. Anderson, B.J. Bushman, B.D. Barthow, J. Cantor, D. Christakis, S.M. Coane, E. Donnerstein, J.F. Brockmyer, D.A. Gentile, C.S. Green, R. Huesmann, T. Hummer, B. Krahe, V.C. Strasburger, W. Warburton, B.J. Wilson, and M. Ybarra, “Screen violence and youth behavior,” Pediatrics, vol.140, no.Supplement 2, pp.S142–S147, 2017.
[4] L.D. Eron, L.R. Huesmann, M.M. Lefkowitz, and L.O. Waldner, “Does television violence cause aggression?,” American Psychologist, vol.27, no.4, pp.253–263, 1972.
[5] C.A. Anderson, A. Sakamoto, D.A. Gentile, N. Ihori, A. Shibuya, S. Yukawa, M. Naito, and K. Kobayashi, “Longitudinal effects of violent video games on aggression in Japan and the united states,” Pediatrics, vol.122, no.5, pp.e1067–e1072, 2008.
[6] L.R. Huesmann, J. Moise-Titus, C.L. Podolski, and L.D. Eron, “Longitudinal relations between children’s exposure to tv violence and their aggressive and violent behavior in young adulthood: 1977-1992,” Developmental psychology, vol.39, no.2, pp.201–221, 2003.
[7] S. Benini, L. Camini, and R. Leonardi, “A connotative space for supporting movie affective recommendation,” IEEE Trans. Multimedia, vol.13, no.6, pp.1356–1370, 2011.
[8] Y. Zhou, J. Wu, T.H. Chan, S.-W. Ho, D.-M. Chiu, and D. Wu, “In-terpreting video recommendation mechanisms by mining view count traces,” IEEE Trans. Multimedia, vol.20, no.8, pp.2153–2165, 2017.
[9] A.B. Jelodar, D. Paulius, and Y. Sun, “Long activity video un-derstanding using functional object-oriented network.” IEEE Trans. Multimedia, vol.21, no.7, pp.1813–1824, 2018.
[10] M.Y. Chen and A. Hauptmann, “Mosift: Recognizing human actions in surveillance videos;” Tech. Rep., CMU-CS-09-161, Carnegie Mellon University, 2009.
[11] F.D. De Souza, G.C. Chávez, E.A. da Valle Jr., and A.A. Araújo, “Violence detection in video using spatio-temporal features,” 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp.224–230, IEEE, 2010.
[12] J. Laptev, “On space-time interest points,” International journal of computer vision, vol.64, no.3, pp.107–123, 2005.
[13] T. Hassner, Y. Icher, and O. Kliper-Gross, “Violent flows: Real-time detection of violent crowd behavior,” 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.1–6, IEEE, 2012.
[14] J. Yang, Y.-G. Jiang, A.G. Hauptmann, and C.-W. Ngo, “Evaluating bag-of-visual-words representations in scene classification;” Proc. international workshop on Workshop on multimedia information retrieval, pp.197–206, ACM, 2007.
[15] J. Lin and W. Wang, “Weakly-supervised violence detection in movies with audio and video based co-training,” Pacific-Rim Conference on Multimedia, pp.930–935, Springer, 2009.
[16] N. Derbas and G. Quenot, “Joint audio-visual words for violent scenes detection in movies,” Proc. International Conference on Multimedia Retrieval, pp.483–486, ACM, 2014.
[17] W. Hu, X. Ding, B. Li, J. Wang, Y. Gao, F. Wang, and S. Maybank, “Multi-perspective cost-sensitive context-aware multi-instance sparse coding and its application to sensitive video recognition,” IEEE Trans. Multimedia, vol.18, no.1, pp.76–89, 2015.
[18] E.B. Nievas, O.D. Suarez, G.B. García, and R. Sukthankar, “Violence detection in video using computer vision techniques,” International conference on Computer analysis of images and patterns, pp.332–339, Springer, 2011.
[19] Y. Ji, Y. Wang, and J. Katoy, “Visual violence rating with pairwise comparison,” 2019 IEEE International Conference on Image Processing (ICIP), pp.3332–3336, IEEE, 2019.
[20] M. Schedl, M. Sjöberg, I. Mironica, B. Ionescu, V.L. Quang, and Y.G. Jiang, “Vsd2014: a dataset for violent scenes detection in hollywood movies and web videos,” Sixth Sense, vol.6, no.20, pp.12–40, 2015.
[21] L. Grealy, C. Driscoll, and K. Cather, “A history of age-based film classification in japan,” Japan Forum, pp.1–26, Taylor & Francis, 2020.
[22] D.A. Gentile, “The rating systems for media products,” Handbook of children, media, and development, pp.527–551, 2008.
[23] L. Jenkins, T. Webb, N. Browne, A.A. Aflit, and J. Kraus, “An evaluation of the motion picture association of america’s treatment of violence in pg-, pg-13-, and r-rated films,” Pediatrics, vol.115, no.5, pp.e512–e517, 2005.
[24] Z. Tan and Y. Zhang, “Predicting the top-n popular videos via a cross-domain hybrid model,” IEEE Trans. Multimedia, vol.21, no.1, pp.147–156, 2018.
[25] J. Lee and J.-S. Lee, “Music popularity: Metrics, characteristics, and audio-based prediction,” IEEE Trans. Multimedia, vol.20, no.11, pp.3173–3182, 2018.
[26] T. Trzciński and P. Rokita, “Predicting popularity of online videos using support vector regression,” IEEE Trans. Multimedia, vol.19, no.11, pp.2564–2570, 2017.
[27] X. Zhang, X. Gao, W. Li, and L. He, “A gated peripheral-foveal con-volutional neural network for unified image aesthetic prediction,” IEEE Transactions on Multimedia, vol.21, no.11, pp.2815–2826, 2019.
[28] J. Donahue and K. Grauman, “Annotator rationales for visual recognition,” 2011 International Conference on Computer Vision, pp.1395–1402, IEEE, 2011.
[29] R. Datta, D. Joshi, J. Li, and J.Z. Wang, “Studying aesthetics in pho-tographic images using a computational approach,” European con-ference on computer vision, pp.288–301, Springer, 2006.
[30] N. Murray, L. Marchesotti, and F. Perronnin, “Ava: A large-scale database for aesthetic visual analysis,” 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp.2408–2415, IEEE, 2012.
[31] R. Herbrich, T. Minka, and T. Graepel, “Trueskill: a bayesian skill rating system,” Advances in neural information processing systems, pp.569–576, 2007.
[32] D.R. Hunter, “Mm algorithms for generalized bradley-terry models,” The annals of statistics, vol.32, no.1, pp.384–406, 2004.
[33] Y.G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang, “Under-standing and predicting interestingness of videos;” AAAI, 2013.
[34] A. Dubey, N. Naïk, D. Parikh, R. Raskar, and C.A. Hidalgo, “Deep learning the city: Quantifying urban perception at a global scale,” European Conference on Computer Vision, pp.196–212, Springer, 2016.
[35] M.H. Kiapour, K. Yamaguchi, A.C. Berg, and T.L. Berg, “Hipster wars: Discovering elements of fashion styles,” European conference on computer vision, pp.472–488, Springer, 2014.
[36] L. Palmer, A. Bialkowski, G.J. Brostow, J. Ambeck-Madsen, and N. Lavie, “Predicting the perceptual demands of urban driving with video regression,” 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp.409–417, IEEE, 2017.
[37] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su, “Violence de-tection in movies,” 2011 Eighth International Conference Computer Graphics, Imaging and Visualization, pp.119–124, IEEE, 2011.
[38] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, “Audio-visual fusion for detecting violent scenes in videos,” Hellenic conference on artificial intelligence, pp.91–100, Springer, 2010.
[39] A. Datta, M. Shah, and N.D.V. Lobo, “Person-on-person violence detection in video data;” Object recognition supported by user inter-
action for service robots, pp.433–438, IEEE, 2002.

[40] H. Wang and C. Schmid, “Action recognition with improved trajectories,” Proc. IEEE international conference on computer vision, pp.3551–3558, 2013.

[41] D. Castán, M. Rodríguez, A. Ortega, C. Orrite, and E. Lleida, “Vivolab and cvlab-mediaeval 2014: Violent scenes detection affect task,” MediaEval, 2014.

[42] V. Lam, S. Phan, D.-D. Le, D.A. Duong, and S. Satoh, “Evaluation of multiple features for violent scenes detection,” Multimedia Tools and Applications, vol.76, no.5, pp.7041–7065, 2017.

[43] Q. Dai, R.W. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, and Y.G. Jiang, “Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning,” MediaEval, 2015.

[44] Q. Dai, Z. Wu, Y.G. Jiang, X. Xue, and J. Tang, “Fudan-njusat at mediaeval 2014: Violent scenes detection using deep neural networks,” MediaEval, 2014.

[45] O. Seddati, E. Kulah, G. Pironkov, S. Dupont, S. Mahmoudi, and T. Dutoit, “Umons at mediaeval 2015 affective impact of movies task including violent scenes detection,” MediaEval, 2015.

[46] X. Li, Y. Huo, Q. Jin, and J. Xu, “Detecting violence in video using subclasses,” Proc. 24th ACM international conference on Multimedia, pp.586–590, ACM, 2016.

[47] Z. Dong, J. Qin, and Y. Wang, “Multi-stream deep networks for person to person violence detection in videos,” Chinese Conference on Pattern Recognition, pp.517–531, Springer, 2016.

[48] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, “Bidirectional convolutional lstm for the detection of violence in videos,” Proc. European Conference on Computer Vision (ECCV), pp.280–295, 2018.

[49] E. Apostolidis and V. Mezaris, “Fast shot segmentation combining global and local visual descriptors,” 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6583–6587, IEEE, 2014.

[50] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” Advances in neural information processing systems, pp.568–576, 2014.

[51] D. Parikh and K. Grauman, “Relative attributes,” 2011 IEEE International Conference on Computer Vision (ICCV), pp.503–510, IEEE, 2011.

[52] T. Joachims, “Optimizing search engines using clickthrough data,” Proc. eighth ACM SIGKDD international conference on Knowledge Discovery and data mining, pp.133–142, ACM, 2002.

[53] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, pp.1097–1105, 2012.

[54] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[55] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE conference on computer vision and pattern recognition, pp.770–778, 2016.

[56] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” European conference on computer vision, pp.20–36, Springer, 2016.

[57] Y. Wang, J. Song, L. Wang, L. Van Gool, and O. Hilliges, “Two-stream sr-cnns for action recognition in videos,” BMVC, 2016.

[58] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” International conference on machine learning, pp.647–655, 2014.

[59] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” Advances in neural information processing systems, pp.487–495, 2014.

[60] M.D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” European conference on computer vision, pp.818–833, Springer, 2014.

Ying Ji received the M.S. degree in Informatics, from Nagoya University in 2019. She is currently a Ph.D. Candidate with the Graduate School of Informatics, Nagoya University.

Yu Wang received the M.S. degree in Information Science and Ph.D. degree in Engineering, from Nagoya University, in 2010 and 2013, respectively. He is currently an assistant professor with the College of Information Science and Engineering, Ritsumeikan University.

Jien Kato received the M.E. and Ph.D. degrees in Information Engineering from Nagoya University in 1990 and 1993, respectively. Then she became a research associate at Toyama University. She was a visiting researcher at the University of Oxford from 1999 for one year. She became an associate professor at the Graduate School of Engineering of Nagoya University in 2000. She has been a professor at the College of Information Science and Engineering of Ritsumeikan University since 2018. Her research interests include object recognition, visual event recognition and machine learning. She is a member of IEICE, IPSJ and JSAI, and also a senior member of IEEE.

Kensaku Mori is a professor of Information and Communications and Graduate School of Information Science, Nagoya University and the director of Information Technology Center of Nagoya University. He is also MICCAI Fellow. He is the President of International Society for Computer-aided Surgery. He received MEng (Information Engineering) and PhD (Information Engineering) from Nagoya University in 1994 and 1996. He started his academic carrier in 1997 as an assistant professor of Nagoya University. Then he was promoted as associate professor in 2001 in Nagoya University. In 2009, he has obtained full professorship in Nagoya University. In 2016, he was appointed as the director of Information Technology Center of Nagoya University. In 2017, he was appointed as a professor of Department of Intelligent Systems, Graduate School of Informatics. He is involved in many international conference organizations, including SPIE Medical Imaging, CARS, and MICCAI, as a general chair or program committee members. He was a conference co-chair of SPIE Medical Imaging Computer Aided Diagnosis conference in 2018 and 2019. He is also an editorial board member of Journal of Medical Imaging of SPIE.