

Motivation & Contributions

Motivation

- Children's exposure to violence has become a severe problem with the rapid development of Internet.
- Recognizing violent video and estimating violence extent become crucial.
- Existing researches focus on violent scene or violent action detection, lacking overall violence extent information.
- There is no dataset includes violence rating labels.

Contributions

- We build a dataset for video violent extent analysis.
- Each video is labelled with 6 objective violent labels and one subjective violence rating label.
- We propose a violence rating prediction approach.

Violent Video Dataset

- 1,930 violent video clips collected from 1,020 action movie promotion videos.
- Each video is manually annotated with 6 objective violent attributes that influence violence extent.
- We employ Trueskill pairwise comparison method to provide ground-truth violence rating for each video.



VISUAL VIOLENCE RATING WITH PAIRWISE COMPARISON

Ying Ji¹, Yu Wang², Jien Kato²

¹Nagoya University, ²Ritsumeikan University

jiying@nagoya-u.jp, {ywang, jien}@fc.ritsumei.ac.jp

Violence Rating Prediction

- Rank learning on video violence rating Learning phase

Data: $D = \{(f_i, l_i)_{i=1}^n\}; l_i = \{L_1, L_2\}$ Ordered pairs: $O = \{(f_i; f_j)\}$, if l_i Similar pairs: $S = \{(f_i; f_j)\}$, if $l_i =$ Learn w^T to make the maximum number of following constraints satisfied:

Solve w^T

$$\forall (i,j) \in 0: w^T f_i > w^T f_j$$

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

by solving the following optimization problem:

$$minimize: \left(\frac{1}{2} ||w^T||^2 + C\left(\sum \varepsilon_{ij}^2 + \sum \gamma_{ij}^2\right)\right)$$

s t $w^T f_i > w^T f_i + 1 - s_i : \forall (i, i) \in O$

$$\begin{aligned} & \left\langle \begin{array}{c} \mathcal{L} & \left\langle \begin{array}{c} \mathcal{L} & \mathcal{L} \\ \end{array} \right\rangle \\ s.t. & w^{T} f_{i} \geq w^{T} f_{j} + 1 - \varepsilon_{ij}; \ \forall (i,j) \in \mathcal{I} \\ & \left| w^{T} f_{i} - w^{T} f_{j} \right| \leq \gamma_{ij}; \forall (i,j) \in \mathcal{S} \\ & \varepsilon_{ij} \geq 0; \gamma_{ij} \geq 0 \end{aligned}$$

Rating phase a. Minimum distance prediction $S_k = 1/N_k$ $L^* = argmin_L$

b. Minimum mean distance predic $F_k(w^T f_k) = \mathcal{N}(w^T f_k)$ $L^* = argmin_L$

c. Maximum Gaussian likelihood p $L^* = argmax_{L_k}$



Using two-stream network to extract features for each video

$$\{L_{2}, L_{3}\}; L_{1} < L_{2} < L_{3}$$

> l_{j}
= l_{j}

$$w^{T} f_{i}, k \in 1, 2, 3$$

$$=L_{k}$$

$$(w^{T} f^{*} - S_{k})^{2}$$

$$(\mu_{k}, \sigma_{k}), k \in 1, 2, 3$$

$$(\mu_{k}, \sigma_{k}), k \in 1, 2, 3$$

$$(w^{T} f^{*} - \mu_{k})^{2}$$
orediction
$$P(w^{T} f^{*} | \mu_{k}, \sigma_{k})^{2}$$

Experiments

- Results

	Method End-to-end			Feature					
Method End-to			Pooling	Raw	L2-norm		SR + L2-nrom		
	20 0 / 0/		Average	39.29%	39.84%		40.93%		
Spatial	3	9.84%	Max	40.11%	41.21%		38.46%		
Alexnet Temporal	11 7C 0/		Average	40.48%	41.23%		42.03%		
Temporal	4	1./5%	Max	42.31%	44.78%		42.03%		
Two strange	Л	C 400/	Average	-	-		46.70%		
iwo-stream	40.40%		Max	-	45.33%		_		
Spatial	Л	J 0C0/	Average	45.60%	47.80%		46.98%		
Spatial	42.80%		Max	45.05%	45.88%		46.70%		
Tomporal	16 120/		Average	47.53%	49.18%		47.53%		
Temporal	4	0.45%	Max	42.31%	47.80%		48.90%		
Two stroom	E0 200/		Average	-	51.65%		-		
		0.20%	Max	_	51.10%		_		
Spatial	11 220/		Average	41.48%	43.96%		48.63%		
Spatial	4	4.2370	Max	42.03%	46.70%		48.08%		
Tomporal	Л	0 000/	Average	46.70%	47.53%		50.27%		
		0.90%	Max	49.18%	49.73%		49.45%		
Two stroom	F	0 0 0 0/	Average	_	_		53.02%		
IWO-Stream	5	0.0270	Max	_	50.8	82%	_		
Methods			Alexnet		VGG16		Resnet-50		
Two-stream End-to-end			46.40%		50.28%		50.82%		
Two-stream feature + SVM			70%	51.65%		53.02%			
Minimum distance prediction			38%	45.60%		49.45%			
Mean distance prediction			18%	53.37%			57.69%		
Maximum Gaussian likelihood			10%	53.85%		57.97%			
	n End-to-end feature + SVN ance prediction	Temporal4Temporal4Spatial4Temporal4Temporal4Spatial4Spatial4Temporal4Temporal4Spatial5Spatial4Temporal5Spatial5Spatial5Spatial5Temporal5Spatial5	Temporal 41.75% Two-stream 46.40% Spatial 42.86% Temporal 46.43% Two-stream 50.28% Spatial 44.23% Temporal 48.90% Temporal 48.90% Two-stream 50.82% Two-stream 40.82% Two-stream 40.82% Two-stream 40.82% Two-stream 40.82% Two-stream 40.82% Two-stream 40.90%	Spatial 39.84% Average MaxTemporal 41.75% Average MaxTemporal 46.40% Average MaxFwo-stream 46.40% Average MaxSpatial 42.86% MaxTemporal 46.43% Average MaxTemporal 46.43% Average MaxTwo-stream 50.28% MaxFwo-stream 50.28% MaxFwo-stream 50.28% MaxSpatial 44.23% Average 	Spatial 39.84% Average 39.29% Temporal 41.75% Average 40.11% Temporal 41.75% Average 40.48% Two-stream 46.40% Average $-$ Spatial 42.86% Average $-$ Spatial 42.86% Average 45.60% Temporal 46.43% Average 47.53% Temporal 46.43% Average 47.53% Temporal 46.43% Average $-$ Max 42.31% Average $-$ Temporal 46.43% Average $-$ Max 42.31% Average $-$ Two-stream 50.28% Average $-$ Spatial 44.23% Average 46.70% Temporal 48.90% Average $-$ Max 49.18% Average $-$ Temporal 50.82% Average $-$ Max	Spatial 39.84% Average 39.29% 39.84% Temporal 41.75% Average 40.48% 41.2 Temporal 41.75% Average 40.48% 41.2 Two-stream 46.40% Average $ -$ Spatial 42.86% Average $ -$ Spatial 42.86% Average 45.60% 47.8 Temporal 46.43% Average 47.53% 49.1 Temporal 46.43% Average 47.53% 49.1 Temporal 46.43% Average 47.53% 49.1 Temporal 46.43% Average $ 51.6$ Two-stream 50.28% Average $ 51.6$ Spatial 44.23% Average 41.48% 43.9 Temporal 48.90% Average $ -$ Two-stream 50.82% Max 49.18% 49.7 Max	Spatial 39.84% Average 39.29% 39.84% Max 40.11% 41.21% Temporal 41.75% Average 40.48% 41.23% Temporal 41.75% Average 40.48% 41.23% Temporal 41.75% Average 40.48% 41.23% Temporal 46.40% Average $ -$ Spatial 42.86% Average 45.60% 47.80% Spatial 42.86% Average 47.53% 49.18% Temporal 46.43% Average 47.53% 49.18% Temporal 46.43% Average $ 51.65\%$ Two-stream 50.28% Average $ 51.10\%$ Spatial 44.23% Average 41.48% 43.96% Temporal 48.90% Average 40.70% 47.53% Temporal 48.90% Average $ -$ Temporal 50.82%		

Conclusion

- We provide a novel violent video dataset with 6 objective attributes and one subject violence level.
- We propose a violence rating prediction method. It can fully utilize the pairwise relationship between different videos.

Dataset: weapon possession attribute • Training data: 1,095; Test data: 364 Network: Alexnet, VGG16, Resnet-50