VISUAL VIOLENCE RATING WITH PAIRWISE COMPARISON

Ying Ji[1], Yu Wang[2], Jien Kato[2]
[1]Nagoya University, [2]Ritsumeikan University
jiying@nagoya-u.jp, {ywang, jien}@fc.ritsumei.ac.jp

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.

Role of participant
- Attacker
- Victim
- One Vs. One
- One Vs. Group
- Group Vs. Group

Body touch
- Have body touch
- Don't have body touch

Explosion
- Don't have explosion
- Have explosion

Blood status
- No blood
- Static blood
- Flowing blood

Weapon possession
- No weapon
- Hold a weapon
- Use a weapon

Weapon direction
- Act on the opponent
- Act towards the screen
- Other direction

Violence Rating Prediction

- Using two-stream network to extract features for each video
- Rank learning on video violence rating

Learning phase

Data: $D=\{(f_i,l_i)\}_{i=1}^{n}; l_i=\{L_1,L_2,L_3\}; L_1<L_2<L_3$

Ordered pairs: $O=\{(f_i;f_j)\}$, if $l_i>l_j$

Similar pairs: $S=\{(f_i;f_j)\}$, if $l_i=l_j$

Learn w^T to make the maximum number of following constraints satisfied:

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

Solve w^T by solving the following optimization problem:

$$minimize: \left(\frac{1}{2}\|w^T\|^2 + C\left(\sum \varepsilon_{ij}^2 + \sum \gamma_{ij}^2\right)\right)$$
$$s.t.\ \ w^T f_i \geq w^T f_j + 1 - \varepsilon_{ij};\ \forall(i,j)\in O$$
$$\left|w^T f_i - w^T f_j\right| \leq \gamma_{ij}; \forall(i,j)\in S$$
$$\varepsilon_{ij}\geq 0; \gamma_{ij}\geq 0$$

Rating phase

a. Minimum distance prediction

$$S_k = 1/N_k \sum_{l_i=L_k} w^T f_i,\ k\in 1,2,3$$
$$L^* = argmin_{L_k}(w^T f^* - S_k)^2$$

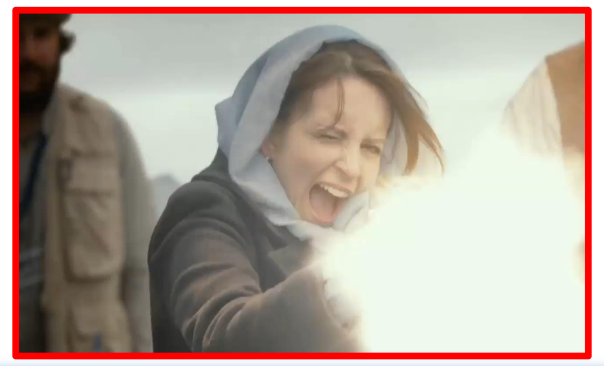
b. Minimum mean distance prediction

$$F_k(w^T f_k) = \mathcal{N}(\mu_k,\sigma_k), k\in 1,2,3$$
$$L^* = argmin_{L_k}(w^T f^* - \mu_k)^2$$

c. Maximum Gaussian likelihood prediction

$$L^* = argmax_{L_k} P(w^T f^* | \mu_k,\sigma_k)^2$$





Experiments

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

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

Methods	Alexnet	VGG16	Resnet-50
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%

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.