Violent Scenes Detection: Task Overview and Results

Mats Sjöberg, Bogdan Ionescu, Yu-Gang Jiang,
Vu Lam Quang, Markus Schedl, Claire-Hélène Demarty

MediaEval 2014 Workshop, Barcelona, October 16-17

Task definition

- Fourth year, some changes in organising team
- Related to Technicolor's use case: helping parents select movies suitable for young children by previewing violent scenes
- We focused on “subjective” definition this year:

  The targeted violent segments are those “one would not let an 8 years old child see in a movie because they contain physical violence”.
- No shot-boundaries this year
- Option to use external data in additional runs

Subtasks

- **Training data:** Hollywood-style feature length movies
  - 24 DVDs for training (bought by participants)
  - 10 violence-related concept definitions for part of the data (e.g. blood, firearms, explosions)
- **Main task:** Hollywood-style movies
  - 7 DVDs, equivalent to previous years
- **Generalisation task:** short YouTube videos
  - Same training data
  - 86 Creative Commons licensed YouTube videos
  - Testing generalisation of systems

Training data

- 24 movies (training + test from 2013, minus “Kill Bill 1”)
- 47h 40min
- 14.5 % subjective violence

Léon, Reservoir Dogs, Armageddon, I am Legend, Saving Private Ryan, Eragon, Harry Potter and the Order of the Phoenix, Billy Elliot, Pirates of the Caribbean - The Curse of the Black Pearl, The Sixth Sense, The Wicker Man, Midnight Express, The Wizard of Oz, The Bourne Identity, Dead Poets Society, Independence Day, Fight Club, Fantastic Four 1, Fargo, Forrest Gump, Legally Blond, Pulp Fiction, The God Father 1, The Pianist

Test data

8 Mile,
Braveheart,
Desperado,
Ghost in the Shell,
Jumanji,
Terminator 2,
V for Vendetta

- Main task
  - 13h 53min
  - 17.2% subjective violence
- Generalisation task
  - 2h 37min
  - Length: 6 s - 6 min
  - video games, amateur videos of accidents, sport events
  - 34.9% subjective violence

Ground truth

- Manually created by several human assessors
  - Several regular annotators
  - Merging and difficult cases by master annotators
- Segments containing subjective violent events
- 7 high level video concepts:

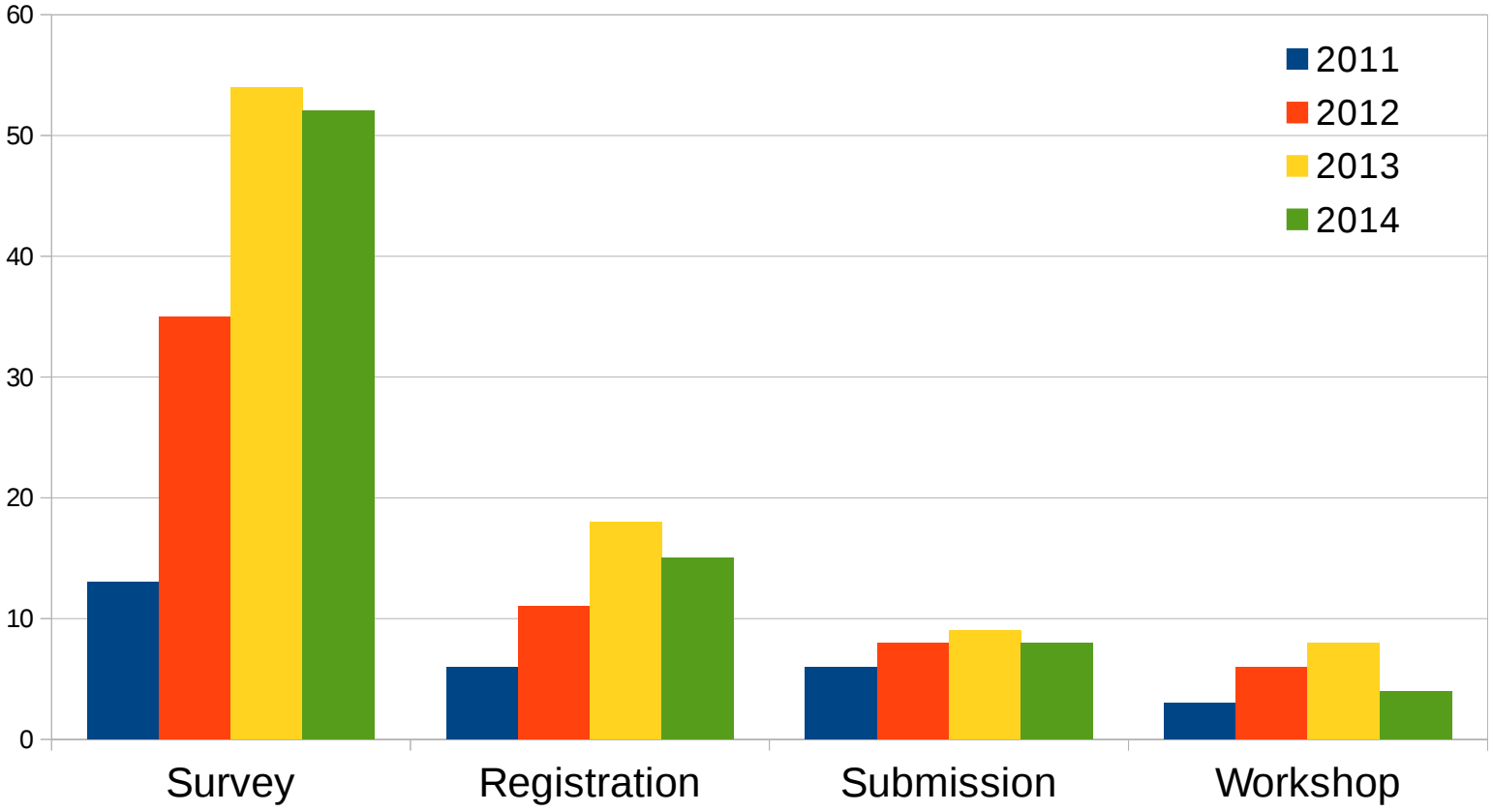
  blood, fire, guns, cold arms (knives etc), fights, car chases, gory scenes
- 3 high level audio concepts:

  gunshots, screams, explosions

Evaluation metrics

- **Official metric:** MAP2014
  - All returned positive-marked segments ordered according to supplied confidence score
  - Hit if >50% overlap in either hypothesis or ground truth segment
  - Several hits on same ground truth segment only counted as one true positive (others discarded, not counted as false positives)
  - Otherwise same as typical average precision definition

Task participation



Task participation

- Grand total of 67 runs submitted
  - **Main task**: 37 runs
  - **Generalisation task**: 30 runs
  - Most teams submitted to both tasks (75%)
  - Only one team used external data (for DNN training)

Results – main task

TEAM	BEST RUN	PREC.	RECALL	MAP@100	MAP2014
Fudan*	4	41.1%	72.1%	72.7%	63.0%
NII-UIT*	1	17.1%	100.0%	77.3%	55.9%
FAR*	1	28.0%	71.3%	57.0%	45.1%
MIC-TJU	3	17.0%	98.4%	63.6%	44.6%
Recod	audio3+text +visual3_ext	33.0%	69.7%	49.3%	37.6%
VIVOLAB	4	38.1%	58.4%	38.2%	17.8%
TUB-IRML	4	31.7%	17.3%	40.9%	17.2%
MTMDCC	dbnbow	15.8%	24.6%	16.5%	2.6%

“Random run” has MAP2014 of 6.1%

* = at least one participant member of organising team

Movie-specific results – main task

	Ghost in the Shell	Brave heart	Jumanji	Desperado	V for Vendetta	Terminator 2	8 Mile
Fudan*	90%	49%	71%	54%	51%	62%	65%
NII-UIT*	94%	51%	22%	58%	50%	58%	59%
FAR*	83%	29%	29%	38%	48%	56%	32%
MIC-TJU	75%	33%	19%	58%	30%	42%	55%
Recod	57%	48%	19%	34%	38%	48%	20%
VIVOLAB	30%	21%	24%	15%	13%	19%	4%
TUB-IRML	28%	18%	6%	25%	17%	21%	5%
MTMDCC	4%	1%	2%	2%	8%	1%	0%
MEAN	**65%**	**35%**	**27%**	**40%**	**35%**	**44%**	**34%**

Results – generalisation task

TEAM	RUN	PREC	RECALL	MAP@100	MAP2014
FAR*	3	49.7%	85.8%	86.0%	66.4%
Recod	audio3+visual3_ext	48.1%	88.4%	86.8%	61.8%
Fudan*	2	59.0%	43.4%	71.9%	60.4%
MIC-TJU	4	44.4%	97.3%	55.5%	56.6%
TUB-IRML	1	63.3%	25.2%	58.2%	51.7%
VIVOLAB	3	51.3%	33.6%	56.5%	43.0%

“Random run” has MAP2014 of 36.4% !

* = at least one participant member of organising team

Trends this year

- Deep neural networks emerging topic
  - generating visual features
  - directly (classification or fusion)
- Features
  - Trajectory-based features
  - GMM Fisher vector encoding
  - Multimodality: almost all teams using audio + video, one team used text as well
- Concepts not very popular this year (only two teams)

Thank you!

What's next?

- 15:45: VSD poster session with coffee
- 16:45: VSD technical retreat session in the “Tower” room
  - Discussion of methods
  - Lessons learned
  - Violent Scenes Detection 2015?