GRADUATE SCHOOL AND RESEARCH CENTER AT THE HEART OF THE DIGITAL SOCIETY

EURECOM
Sophia Antipolis

Affective Multimodal Analysis for the Media Industry

O. Ben-Ahmed and B. Huet
EURECOM
Sophia Antipolis, France

Motivation and Context

- Media Industry and Multimedia Retrieval
  - Indexing and retrieval
  - Annotation
  - Summarization and Trailer creation
  - Enrichments and hyperlinking
- NexGenTV (ANRT National):
  - Multimodal analysis for enriching broadcast content via second screen applications
- ANTRACT (ANR National):
  - Mine video archive to provide information for historians
- MeMAD (H2020 EU):
  - Provide new methods to translate Image and Sounds into Words (for indexing, retrieving, repurposing and accessibility)

NEXGEN TV

ANTRACT

MeMAD
Methods for Managing Audiovisual Data

Predicting Media Interestingness (PMI)

- Automatically analyze media data
- Identify the most attractive content

MediaEval Benchmark

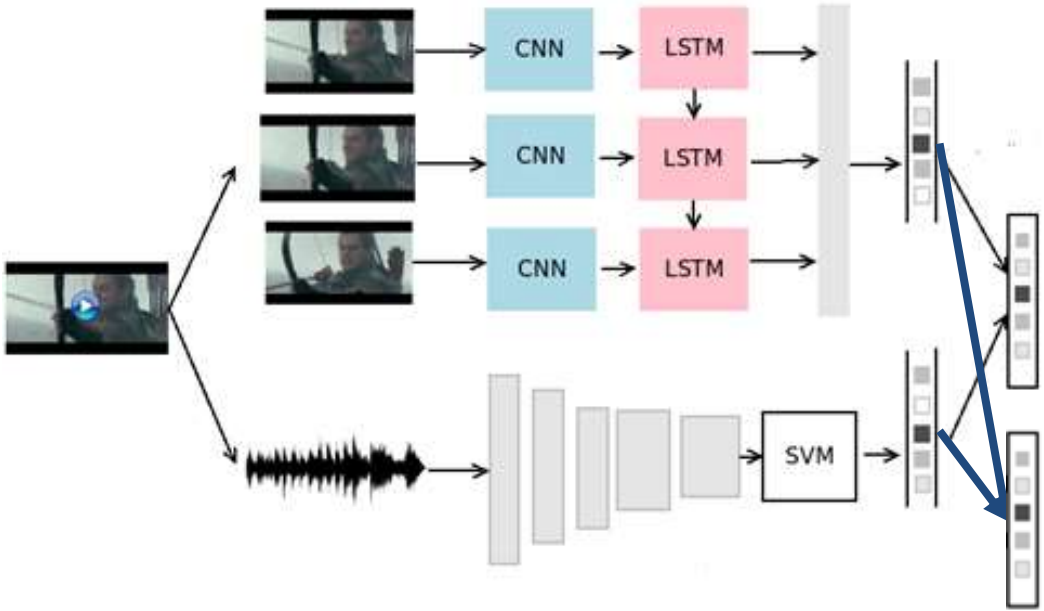
- Data Driven / Content based approaches
  - gap between low-level features and high-level human perception
- Our proposal
  - Address PMI in association with Media Genre

http://www.dailyherald.com/article/20110627/entlife/706279989/

Why Genre Inference for PMI

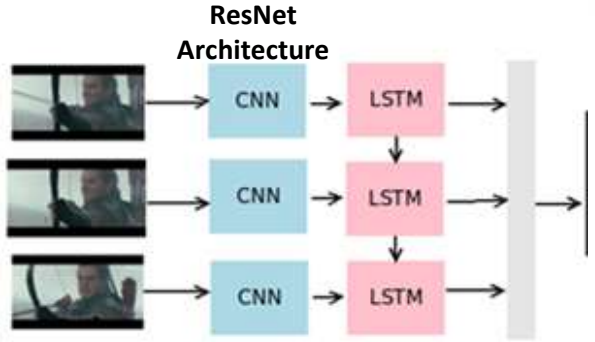
- **Motivation**
  - Interestingness is highly correlated with data emotional content
  - Affective representation of data content
- **Hypothesis**
  - Emotional impact of movie genre can be a factor for interestingness of a video fragment
- **Constraints**
  - Subjectivity of the task – Data collection issues
  - Limited Dataset
- **Method**
  - Mid-level representation based on media genre prediction for PMI
    - Represent each video fragment/image as a distribution of genres
  - Transfer Learning
    - Using genre probability distribution to infer Media Fragment Interestingness

Our Framework

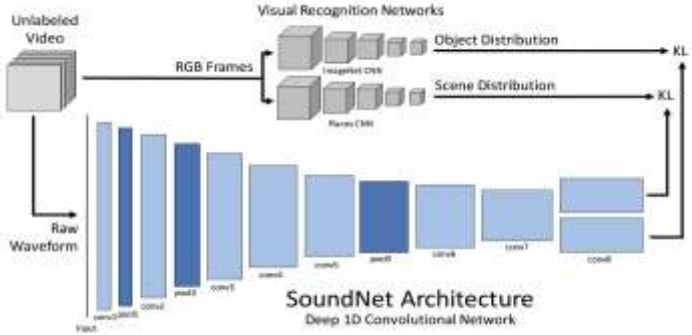


Media Genre Prediction

- **Visual Branch**
  - Deep CNN for features extraction
  - LSTM for modelling temporal cue



- **Audio Branch**
  - Deep features extraction : Soundnet
  - SVM classifier

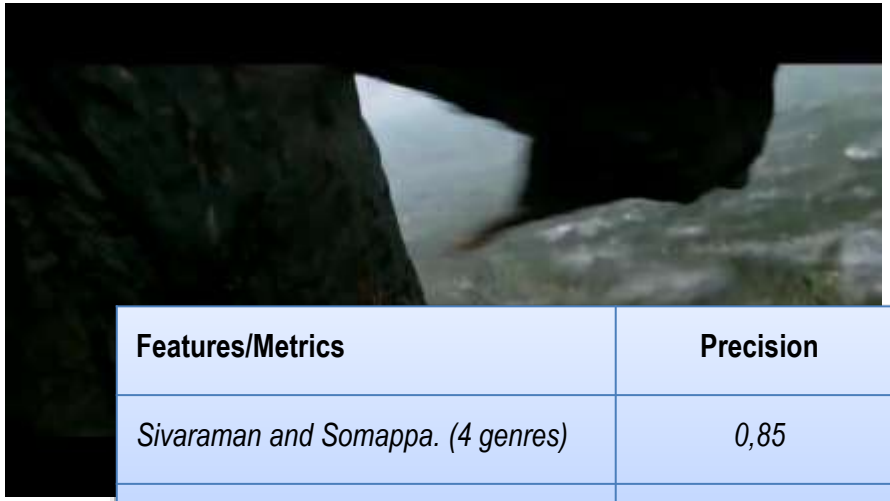


Media Genre Dataset

- Movie trailers dataset
  - K S Sivaraman and G. Somappa. MovieScope
    - 525 YouTube Videos over 4 Genres (available on IMDB)
  - Extended with a 5th Genre
    - Sci-Fi

Trailer Genre	Training	Test	Total
Action	69	44	114
Drama	95	39	134
Horror	99	59	158
Romance	80	39	119
Sci-fi	72	35	107
Total	**415**	**216**	**632**

Media Genre Prediction Example

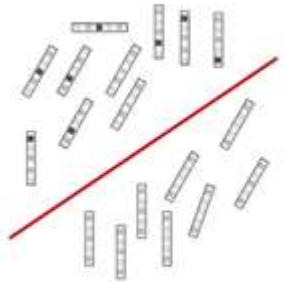


Features/Metrics	Precision
Sivaraman and Somappa. (4 genres)	*0,85*
One Frame-VGG + Soundnet	0,85
ResNet-LSTM + Soundnet	**0,87**

	Visual	Audio	Audio-Visual
Action	2.5%	33,34%	17.92%
Drama	0%	17,78%	8.89%
Horror	0.37%	2.61%	1.49%
Romance	0%	3.87%	1.93%
Sci-fi	97.12%	42,40%	69,76%

Interestingness Classification

- **Features vectors**
  - Probability vector for the genre distribution for video fragments
- **Classifier**
  - Binary SVM,
  - Including assessors confidence scores
- **Modality Analysis**
  - Audio genre vector
  - Visual genre vector
  - Audio-Visual genre vector
    - Concatenated visual- and audio-based genre vectors probabilities



Media Interestingness Dataset

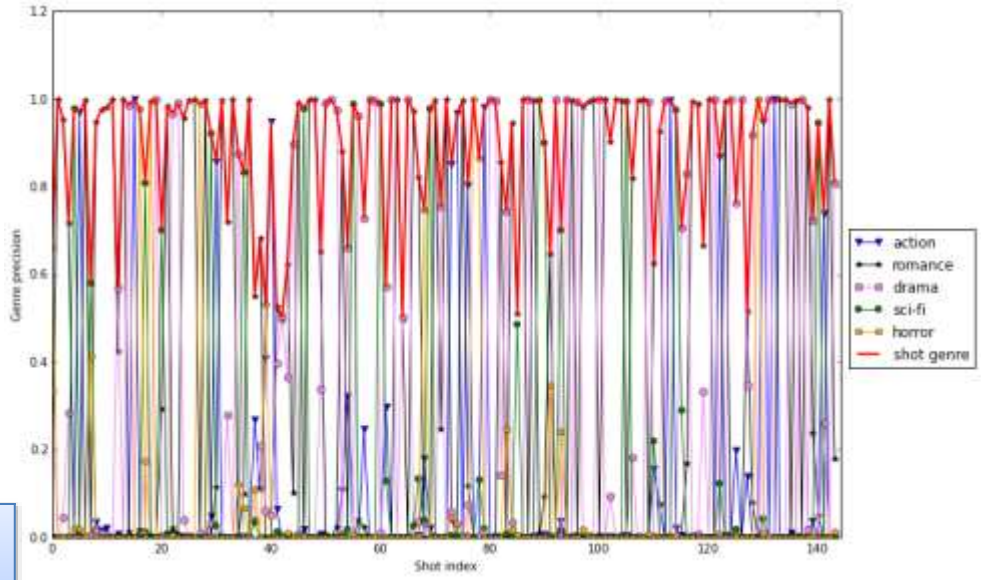
MediaEval Benchmark

- Creative Commons licensed “Hollywood” Video
  - 103 movie trailers and 4 continuous extracts
  - Shot segmentation
  - 7396 video segments for training and 2435 video segments for testing
  - Low level and mid level features
  - Annotation (GT) performed by human assessors
    - Interesting (1) or Not Interesting (2)

Example 1: The longest ride

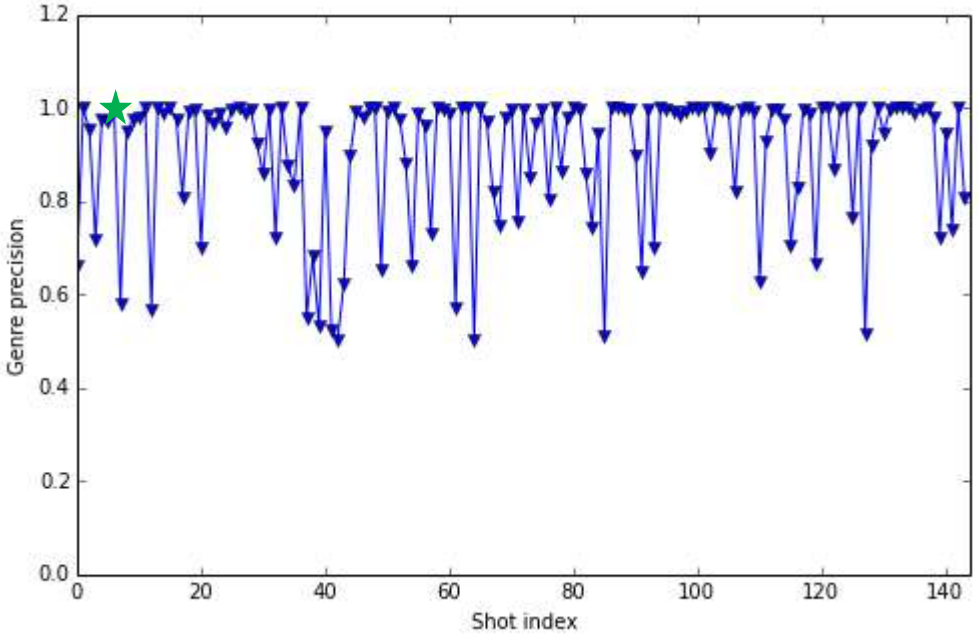
GENRE	Action	Drama	Horror	**Romance**	Sci-fi
COVERAGE	9.02%	29.86%	6.25%	**43.75%**	11.11%

Example 1: The longest ride



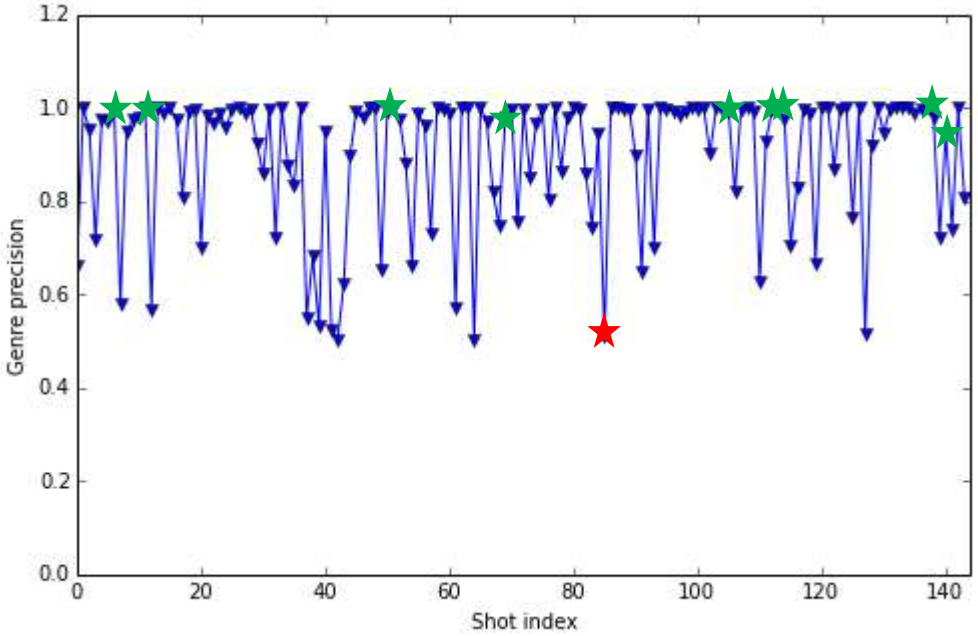
GENRE	Action	Drama	Horror	**Romance**	Sci-fi
COVERAGE	9.02%	29.86%	6.25%	**43.75%**	11.11%

Example 1: The longest ride



Predicted Interestingness= 1★
Romance = 97,45%

Example 1: The longest ride



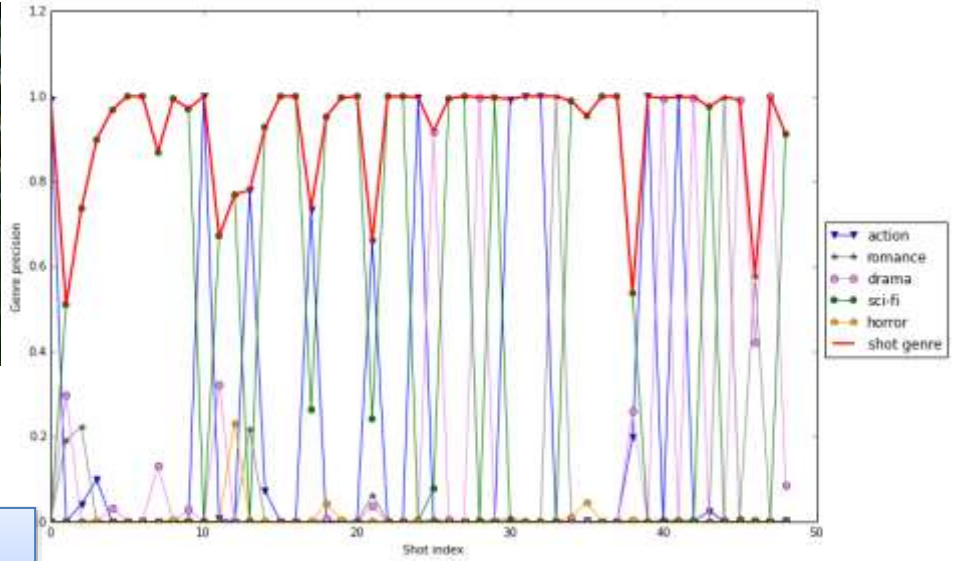
Predicted Interestingness= 0★

Romance = 48,49%

Example 2: Tears of Steel

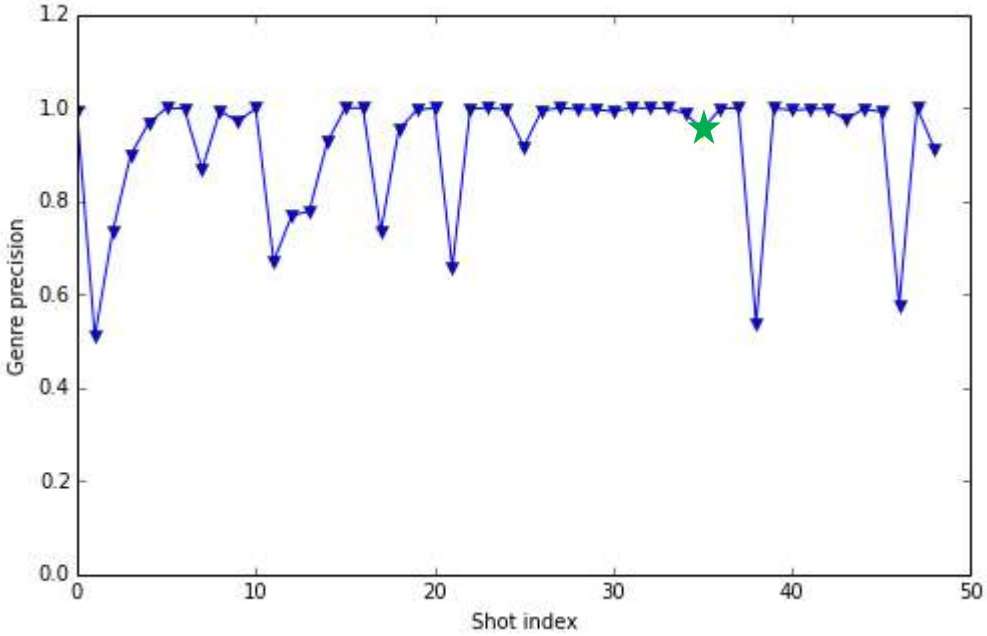
GENRE	Action	Drama	Horror	Romance	**Sci-fi**
COVERAGE	18.37%	16.32%	6.12%	6.12%	**53.06%**

Example 2: Tears of Steel



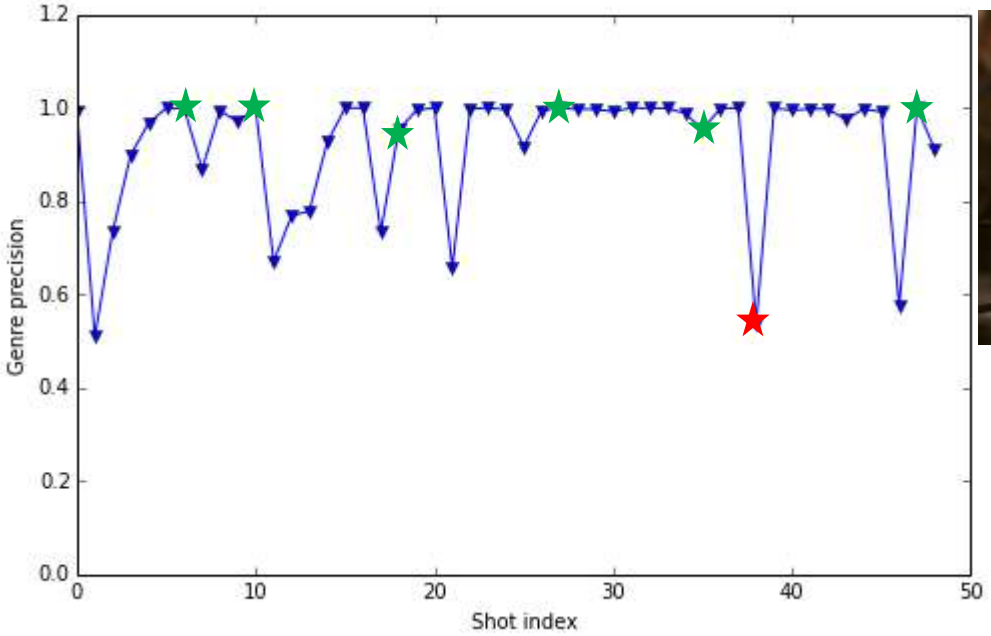
GENRE	Action	Drama	Horror	Romance	**Sci-fi**
COVERAGE	18.37%	16.32%	6.12%	6.12%	**53.06%**

Example 2: Tears of Steel



Predicted Interestingness= 1★
Sci-fi =95,38%

Example 2: Tears of Steel



Predicted Interestingness= 0 ★
Sci-fi= 53,67%

Experiments and Results

	MAP	MAP@10
Audio (Soundnet)	0.1806	0.0511
Visual key frame-VGG (MediaEval'17)	0.1960	0.0732
Visual (LSTM+Resnet)	0.1991	0.0815
VGG+Audio (MediaEval'17)	0.2094	0.0827
Resnet-LSTM+Audio	**0.2122**	**0.0841**
Baseline [ref]	0.1716	0.0564

Conclusions and Future Work

- We proposed to train a **Genre Recognition System** as a mid-level representation for **Predicting Media Interestingness**

- **Deep Audio and Visual Features** for Genre Recognition
- LSTM used to predict Genre over Video Shot duration

- Transfer Learning for Predicting Media Interestingness
- Best Results:
  - **0.21** MAP (on test set) [Previous State of the Art 0.20 MAP]

- Audio brings limited additional information for PMI
- End-to-End joint learning (fine-tuning) of audio-visual features
- Extent mid-level representation with other features (Emotion, Valence/Arousal, etc.)

Recent Related Publications

- **Ben-Ahmed, O., J. Wacker, A. Gaballo, B. Huet, *EURECOM @MediaEval 2017: Media genre inference for predicting media interestingness*,** MEDIAEVAL 2017, MediaEval Benchmarking Initiative for Multimedia Evaluation, 13-15 September 2017, Dublin, Ireland
- **Pini S., O. Ben-Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, Rita; B. Huet, *Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild,*** ICMI 2017, 19th ACM International Conference on Multimodal Interaction, November 13-17th, 2017, Glasgow, United Kingdom
- **Smith, J. R., D. Joshi, B. Huet, W. Hsu, J. Cota, *Harnessing A.I. for augmenting creativity: Application to movie trailer creation*,** ACMMM 2017, 25th ACM Multimedia Conference, October 23-27, 2017, Mountain View, CA, USA
- **Tiwari S. N., N. Q. K. Duong, F. Lefebvre, C.-H. Demarty, B. Huet, L. Chevallier, *Deep features for multimodal emotion classification*, on HAL**
- **Paleari, M., R. Chellali, B. Huet, *Bimodal emotion recognition,*** ICSR 2010, International Conference on Social Robotics, November 23-24, 2010, Singapore / Also published as LNCS Volume 6414/2010
- **Mérialdo B. and B. Huet, *Automatic video summarization***, Book chapter in "Interactive Video, Algorithms and Technologies" by Hammoud, Riad (Ed.), 2006, XVI, 250 p, ISBN: 3-540-33214-6

Questions?

Thank you,
Benoit Huet.