QoMEX 2015

Quality Evaluation of Computational Models for Movie Summarization

A. Zlatintsi*, P. Koutras*, N. Efthymiou*, P. Maragos*, A. Potamianos* and K. Pastra+

*School of Electr. & Comp. Enginr., National Technical University of Athens, 15773, Athens, Greece

+Cognitive Systems Research Institute, Athens, Greece

1. Outline - Contributions

- Use of state-of-art-techniques for **salient event detection** and **movie summarization** using multicue information (i.e., audio, visual and text information)
- Investigate, through subjective and objective quality evaluation, how such systems can be improved
- The evaluation of computation models for movie summarization is usually based on the comparison of the system-detected observations and ground-truth data

➡ **MovSum Database:** an evolving multimodal video oriented database annotated with:

- **saliency (sensory and semantic events) & cross-media relations**

- Goal: the production of ***user-defined***, ***high-quality*** movie summaries that will heighten the human experience and consist of user-preferred content

2. *MovSum* Database (Movie Summarization Database)

Saliency annotation

7 movies (30 min long):

A Beautiful Mind, Chicago, Crash, The Departed, Gladiator, Lord of the Rings – the Return of the King, Finding Nemo

Sensory information: Audio, Visual and Audio-Visual saliency

- **Semantic information:** conceptually important stand-alone semantic events (e.g., important names, plot elements etc)

"Good to see you again old friend!"

- **Informative-segments:** segments important for understanding the narration-plot of the specific half-hour movie clip
- **Affective information:** intended and experienced emotions [3]

Additionally:

Expert summaries, in relation to the plot (ca. 5min)

Movie structure, incl. scene & shot segmentation

Percentage (%) of Salient Frames (labeled by at least two annotators)								
Layer	BMI	CHI	CRA	DEP	GLA	LOR	FNE	Mean
A	25.4	56.3	55.0	33.4	60.9	58.3	54.6	49.1
V	30.1	46.3	37.9	32.4	39.2	43.3	36.9	38.0
AV	27.4	47.7	43.1	37.8	69.6	50.7	39.7	42.3
AVS	63.2	76.6	64.8	71.8	68.5	72.7	67.6	69.3

Cross-media Relations

Based on the *COSMOROE framework* [4]
Movie: Gone with the wind (duration 1:44:15)

Equivalence: Different modalities express semantically equivalent information.

Type-Token:
Word: "dog", image:"dog"
Acoustic event: barking

Metonymy, part for whole
Word: "land",
image: part of the land

Complementarity: The information expressed in one medium is complement to the information expressed in another medium.

Agent-object Essential
Word: "scarlett, you look"
Image: a piece of paper

Exophora Essential
Word: "That"
Image: pointing at a dress

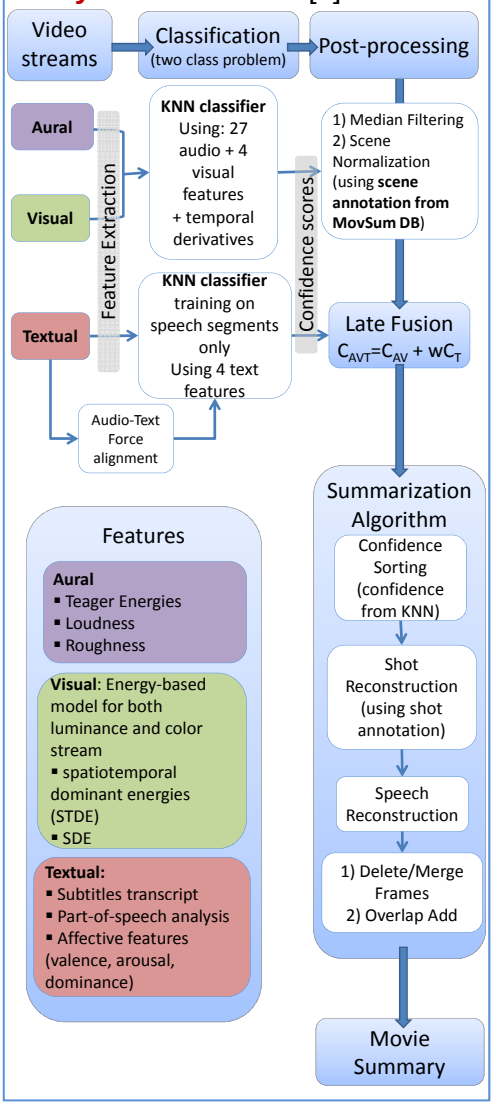
Independence: Each medium carries an independent message and their combination creates a coherent multimedia message
incl. *Contradiction, Symbiosis, Meta-Information*

Acoustic events annotation: incl. a) animal sounds, human sounds, natural/environmental sounds, machine sounds, general background sounds (music etc.)

Bell

Gallop

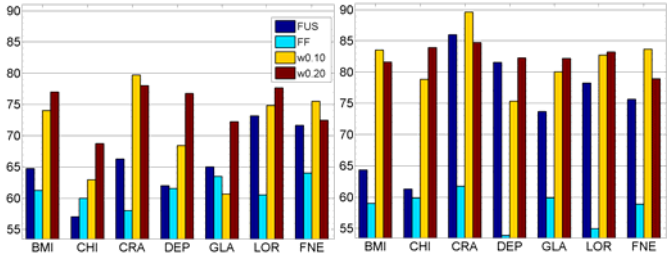
3. Movie Summarization System Overview [1]



4. Subjective Qualitative and Objective Experimental Results

Subjective Qualitative Evaluation setup

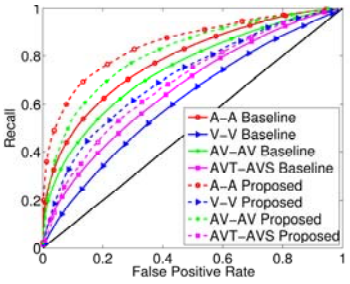
Summaries x5, ca. 6 min.
20 users
Evaluation on:
T_W0.1: text weight $T_W = 0.1$
T_W0.2: text weight $T_W = 0.2$
FUS: fusion method based on [2]
FF: fast-forward (sub-sampling 2 sec. every 10 sec.).



Results

- ➢ up to 80% for informativeness
- ➢ up to 90% for enjoyability
- ➢Different T_W is important and related to the movie genre
- ➢ Action movies need higher T_W
- ➢Boundary correction contributed to enjoyability:

a) smoother transitions and
b) semantically coherent events



Objective Machine Learning Results

- ➢ Proposed system outperforms the baseline [2] for all evaluation setups
- ➢Greater improvement for A-A

Improvements due to:

- ➢Advanced monomodal frontends, in all modalities
- ➢Carefully designed movie summarization algorithm that

a) corrects the boundaries and
b) Results in smoother transiotions

References

[1] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, A. Potamianos, "Predicting Audio-Visual Salient Events Based on Visual, Audio and Text Modalities for Movie Summarization", acc. ICIP-2015.
[2] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention", IEEE Trans. on Multimedia, vol. 15(7), pp. 1553–1568, 2013.
[3] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking", in Proc. ICASSP-2011.
[4] K. Pastra, "COSMOROE: a cross-media relations framework for modelling multimedia dialectics", Multimedia Systems, vol. 14(5), 2008
[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20(11), pp. 1254–1259, 1998.
[6] K. Maninis, P. Koutras, and P. Maragos, "Advances on action recognition in videos using and interest point detector based on multiband spatio-temporal energies", in Proc. ICIP-2014.
[7] J. Kaiser, "On a simple algorithm to calculate the energy of a signal", in Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Process., 1990.
[8] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis", IEEE Trans. Signal Process, vol. 41, p. 3024-3051, 1993.
[9] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Distributional semantic models for affective text analysis", IEEE Trans. Audio, Speech, and Language Process., vol. 21(11),pp. 2379–92,2013.
[10] M. Bradley and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Tech. report C-1". Center for Research in Psychophysiology, Univ. Florida, 1999.
[11] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, "A saliency-based approach to audio event detection and summarization", in Proc. European Signal Process. Conf., 2012.
[12] P. Maragos, The Image and Video Processing Handbook, 2nd ed. Elsevier Acad. Press, 2005, ch. Morphological Filtering for Image Enhancement and Feature Detection, pp. 135–156.

Acknowledgments

This research work was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources.
For more information: Please contact: *[nzlat,pkoutras,maragos]@cs.ntua.gr*

European Union
European Social Fund
OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING
MANAGING AUTHORITY
NSRF 2007-2013
Co- financed by Greece and the European Union