Efficient Video Coding in H.264/AVC by using Audio-Visual Information

Jong-Seok Lee & Touradj Ebrahimi
EPFL, Switzerland

MMSP'09
5 October 2009

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
Multimedia Signal Processing Group
Swiss Federal Institute of Technology

Introduction

- **Objective of video coding**
  - Better quality with smaller number of bits

- **How to achieve better video coding efficiency?**
  - Using statistics of signal
  - Using human visual system's characteristics: Focus of attention
    - Only small region around fixation point is captured at high spatial resolution.

Unattended region

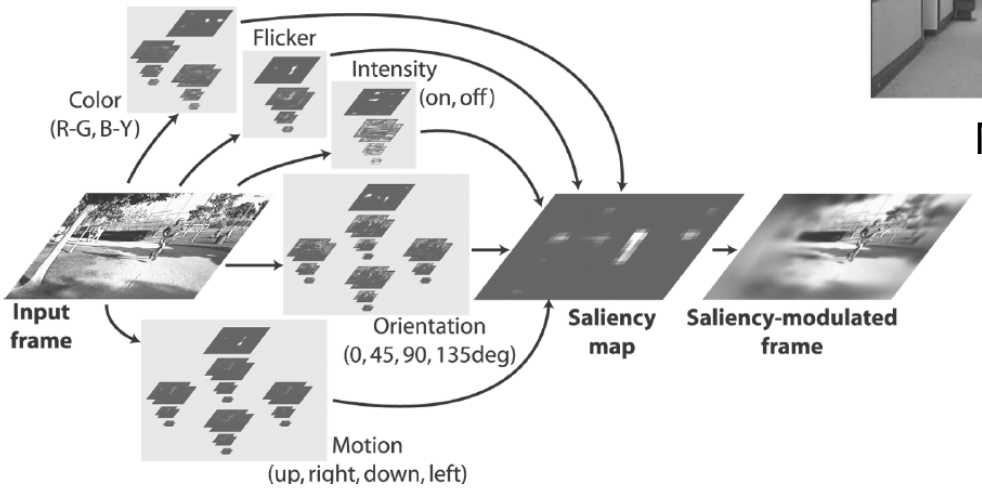
→ more compression

Attended region

→ less compression

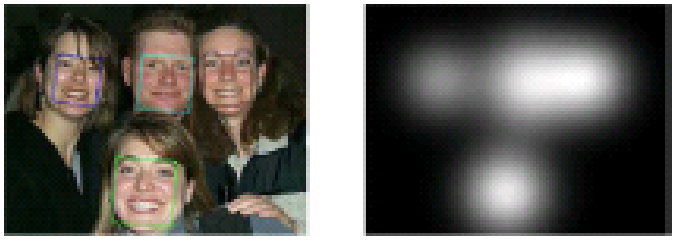
Introduction

- **Which region draws attention?**



Conspicuity-based (Itti, 2004)

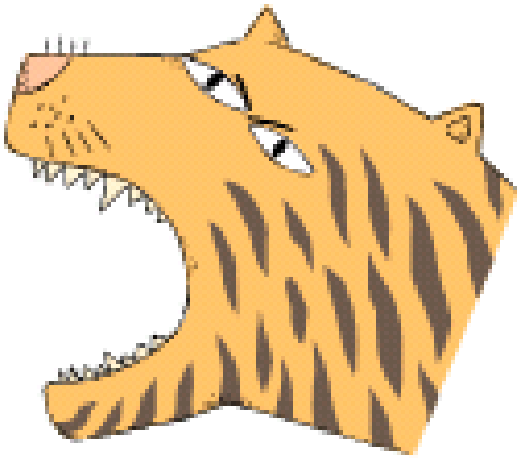
Moving object-based (Cavallaro, 2005)



Face-based (Boccignone, 2008)

No consideration of cross-modal (audio-visual) interaction!

Audio-Visual Focus of Attention

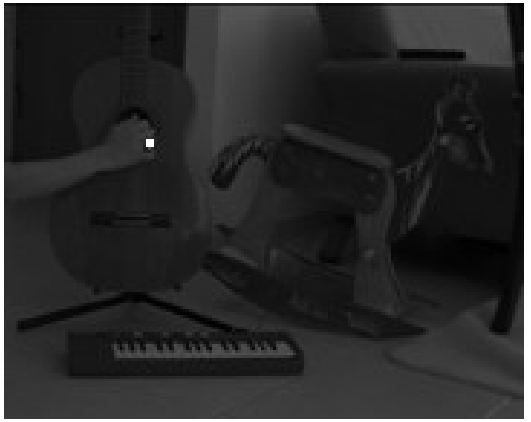


- Abrupt sound draws visual attention to sound source location. (Spence, 1997)
- Attending to auditory stimuli at given location enhances processing of visual stimuli at same location. (Spence, 1996)

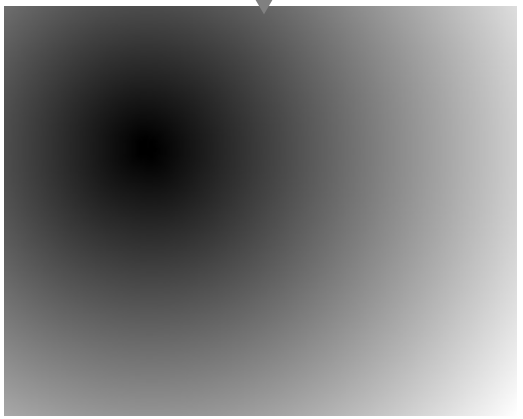
➔ We define sound-emitting region as attended region.

Overall Procedure

Original frame



Source localization



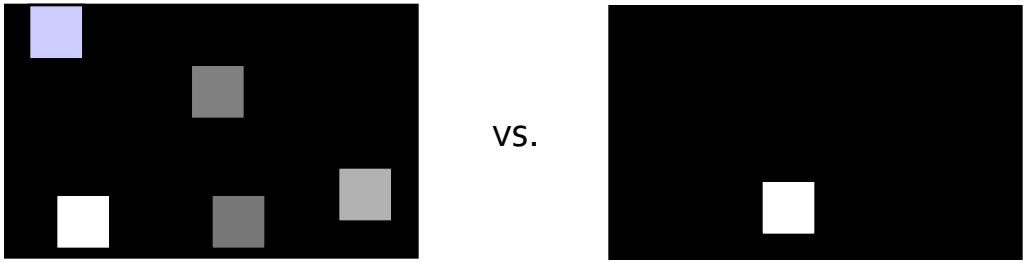
Priority map

Slice grouping

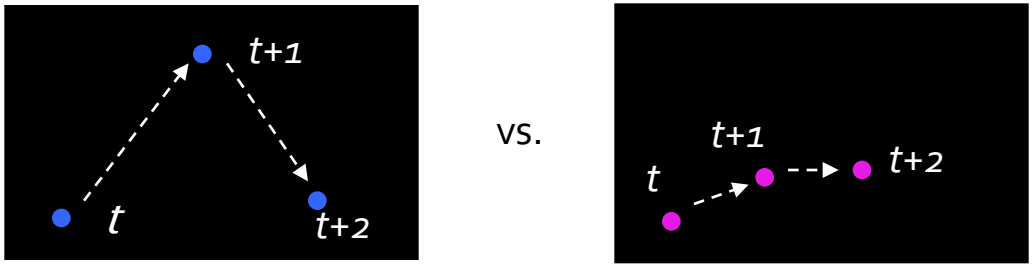
H.264/AVC coding with flexible macroblock ordering (FMO)

Audio-Visual Source Localization

- **To identify spatial location of sound source in scene**
- **Approach**
  - Canonical correlation analysis
    - To find projection vectors of two data for maximizing correlation
  - Sparsity principle



  - Spatio-temporal consistency

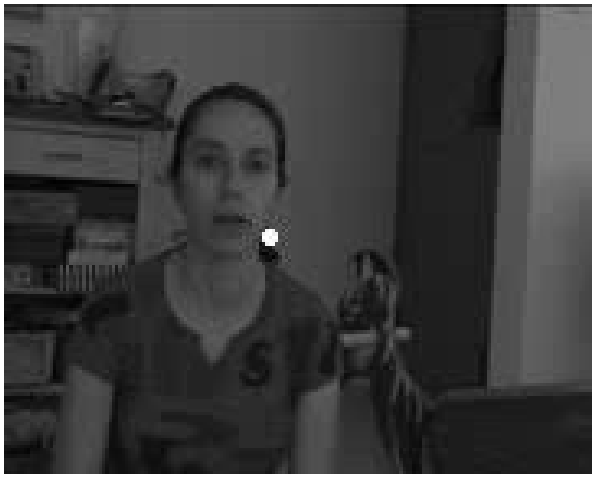


Audio-Visual Source Localization

- **Constraint optimization → linear programming**

- **Advantages**
  - Applicability to normal video with mono audio channel
  - No assumption on sound source
  - No training required

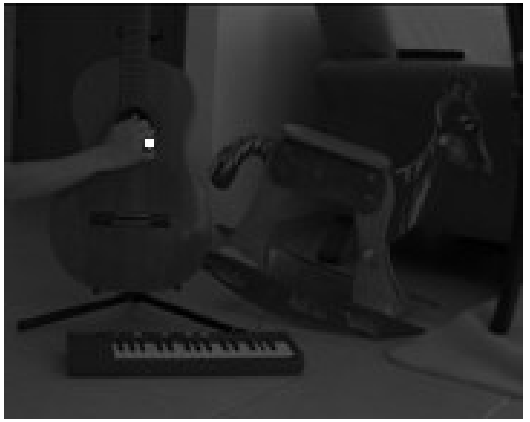
- **Example**



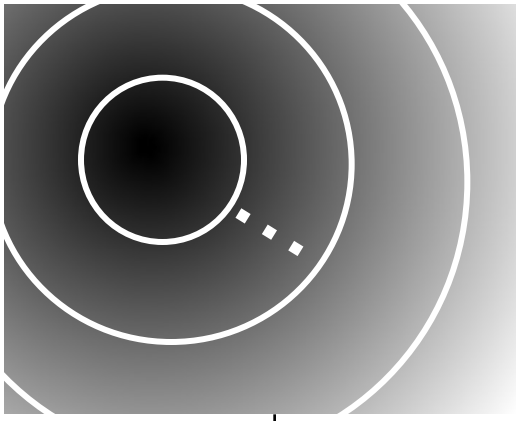
J.-S. Lee, F. De Simone, T. Ebrahimi
"Video coding based on audio-visual attention,"
ICME'09

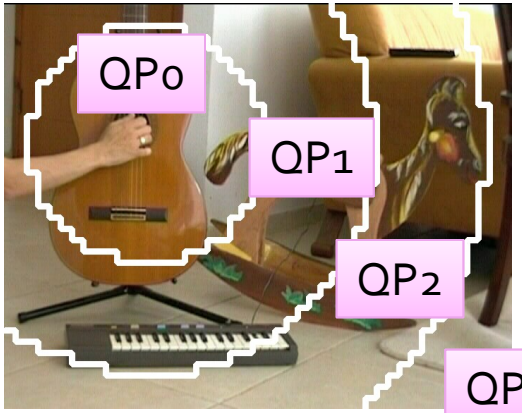
Video Coding

Localization result



Priority map





$QP_1=QP_0+\Delta QP$

$QP_2=QP_1+\Delta QP$

$QP_3=QP_2+\Delta QP$

...

Slice grouping

H.264/AVC coding with FMO (Type 6)

* QP=quantization parameter

Experiments

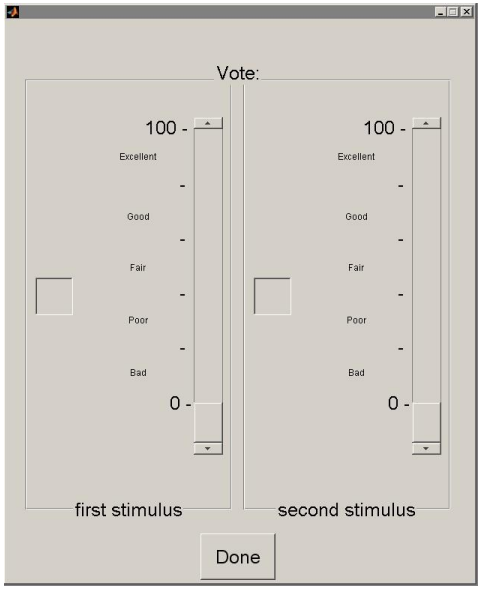
- **2 test sequences including multiple moving objects in scene**

- **Audio-visual source localization**
  - Visual features: differential grayscale pixel value
  - Audio features: differential frame energy

- **H.264/AVC coding: JM reference software**
  - Constant QP mode
  - Rate control (adaptive QP) mode
  - Proposed method (FMO enabled)

Experiments

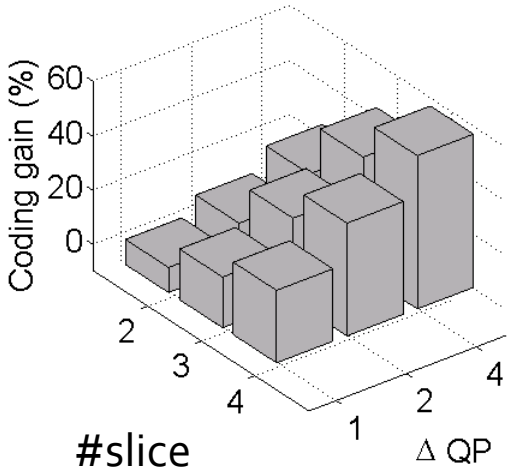
- **Subjective test**
  - Is quality degradation acceptable?
  - ITU-R BT.500-11
  - Double stimulus continuous quality scale (DSCQS)



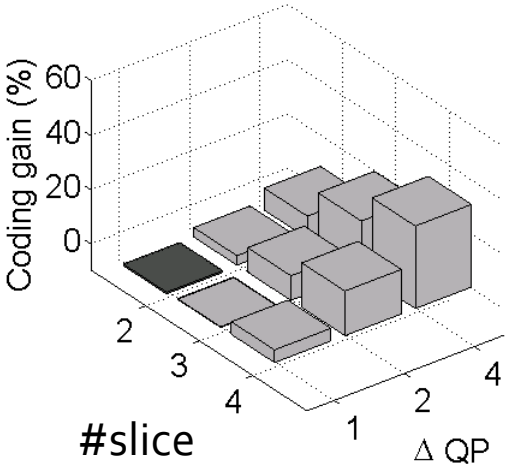
Result

- **Coding gain by proposed method over constant QP mode**

QP0=22



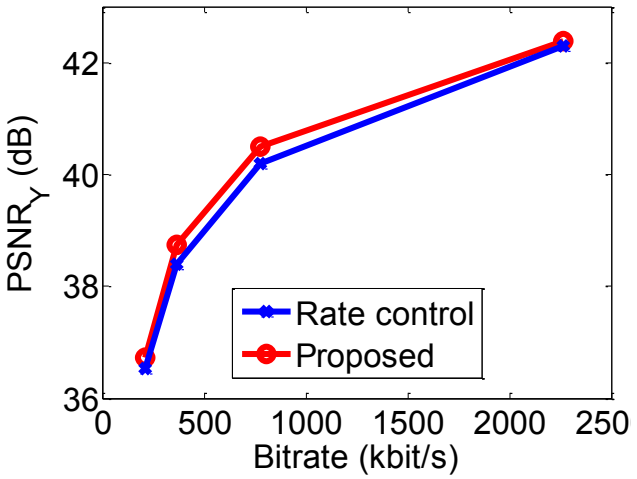
QP0=30



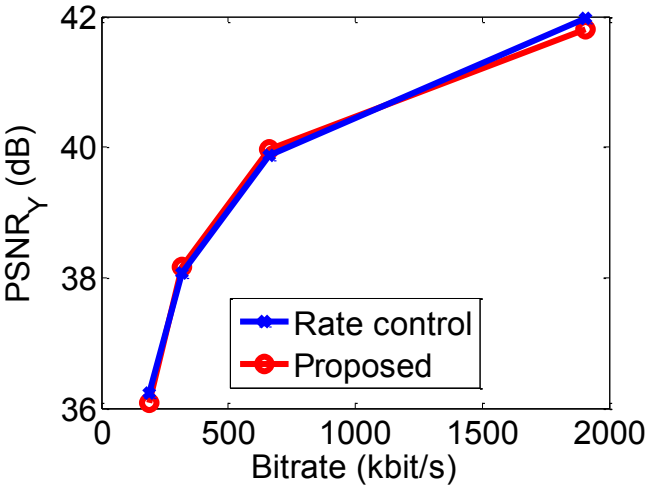
Result

- **Rate-distortion curves**
  - Proposed method (#slice=2) vs. rate control

ΔQP=1

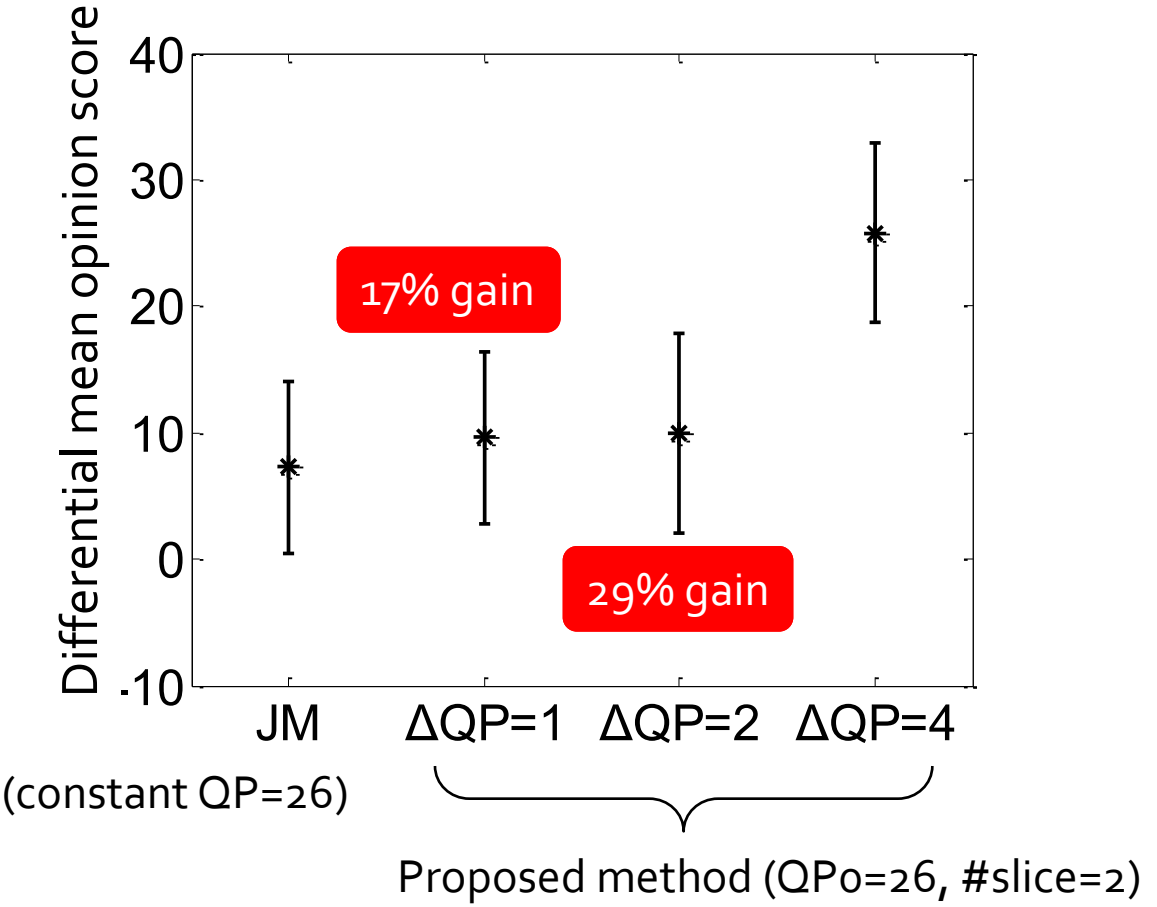


ΔQP=4



Result

- **Subjective quality comparison**



Conclusion & Discussion

- **Audio-visual focus of attention (AV FoA) influences perceived quality. And, it can be used for efficient video coding by H.264/AVC.**

- **Discarding information outside focus of attention does not degrade perceived quality significantly.**

- **AV FoA does not explain everything. It should be combined with other attention mechanisms.**

Questions/comments are welcome!

Contact

jong-seok.lee@epfl.ch

http://mmspg.epfl.ch

References

- L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," IEEE Trans. Image Process., 2004
- A. Cavallaro, O. Steiger, T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," IEEE Trans. Circuits Syst. Video Technol., 2005
- G. Boccignone, A. Marcelli, P. Napoletano, G. D. Fiore, G. Iacovoni, S. Morsa, "Bayesian integration of face and low-level cues for foveated video coding," IEEE Trans. Circuits Syst. Video Technol., 2008
- B. Stein, M. Meredith, "The merging of Senses," MIT Press, 1993
- R. Sharma, V. I. Pavlovic, T. S. Huang, "Toward multimodal human-computer interface," Proc. IEEE, 1998
- H. McGurk, J. MacDonald, "Hearing lips and seeing voices," Nature, 1976
- J.-S. Lee, C. H. Park, "Robust audio-visual speech recognition based on late integration," IEEE Trans. Multimedia, 2008
- M. Sargin, Y. Yemez, E. Erzin, A. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," IEEE Trans. Multimedia, 2007
- P. Perez, J. Vermaak, A. Blake, "Data fusion for visual tracking with particles," Proc. IEEE, 2004
- B. Rivet, L. Girin, C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signal from convolutive mixtures," IEEE Trans. Multimedia, 2007
- C. Spence, J. Driver, "Audiovisual links in exogenous covert spatial orienting," Perception & Psychophysics, 1997
- C. Spence, J. Driver, "Audiovisual links in endogenous covert spatial attention," J. Experimental Psychology: Human Perception & Performance, 1996
- E. Kidron, Y. Schechner, M. Eland, "Cross-modal localization via sparsity," IEEE Trans. Signal Process., 2007

jong-seok.lee@epfl.ch