JEG-HYBRID

Video quality experts group meeting, October 24-28, 2016, London

First part

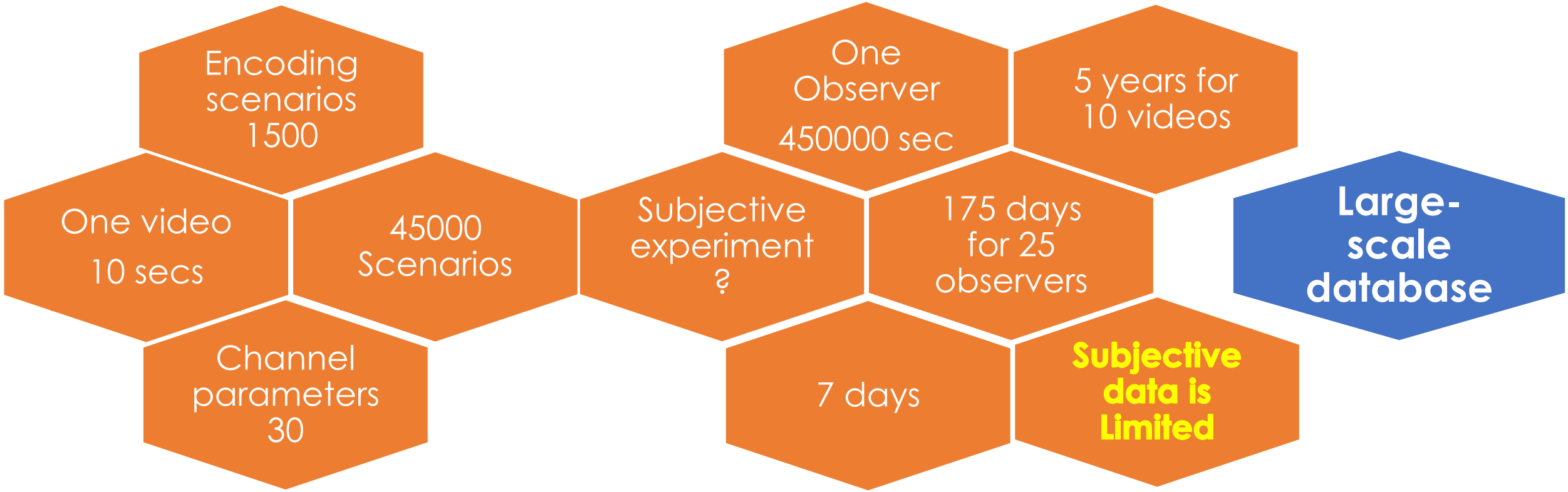
COMPARING SIMPLE VIDEO QUALITY MEASURES FOR LOSS-IMPAIRED VIDEO SEQUENCES ON A LARGE-SCALE DATABASE

Enrico Masala, Ahmed Aldahdooh, Glenn Van Wallendael, Marcus Barkowsky

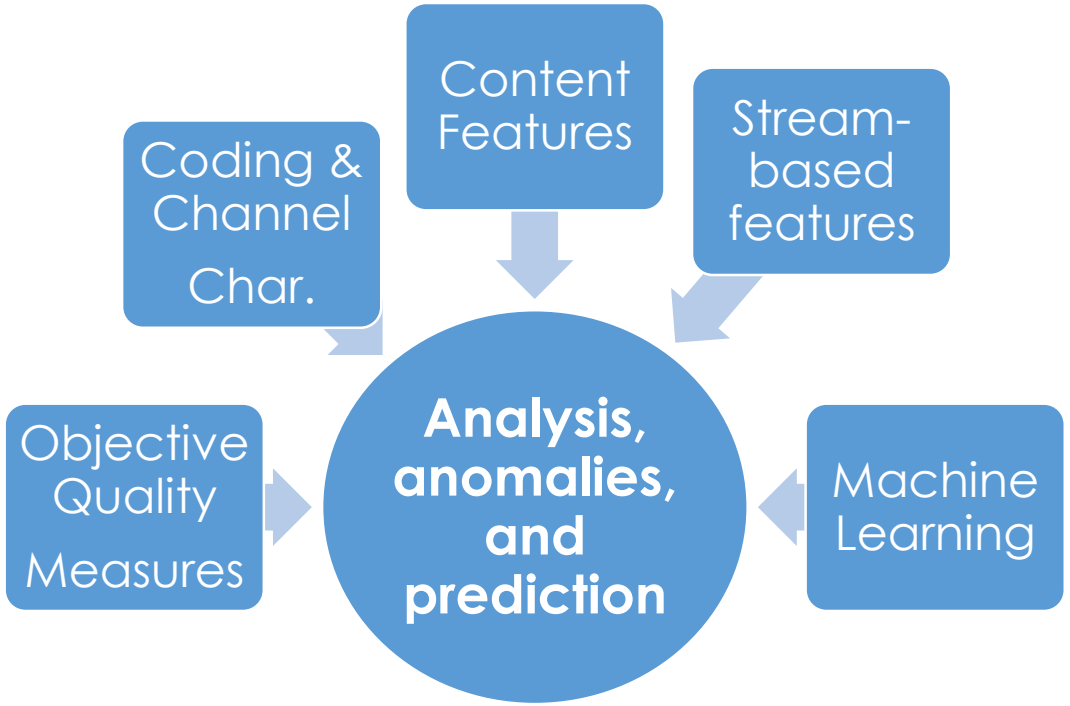
OUTLINE

- Motivation
- Large scale database
  - Loss-impaired videos
  - Content features
- Quality measures disagreement?
- Is full reference behavior predictable?
  - Analysis based on △PSNR Prediction
  - Prediction of △PSNR
- Conclusion

MOTIVATION 1/2



MOTIVATION



Not new Quality Measure

Not Subjective Exp.

LARGE SCALE DATABASE

- Large-scale database
  - SRC:10
  - Compression scenarios: 5952
  - Total: 59520
  - Quality: PSNR, SSIM, VIF, VQM, PVQM
  - Analysis
    - Reasons of disagreement among quality measurements for each sequence.

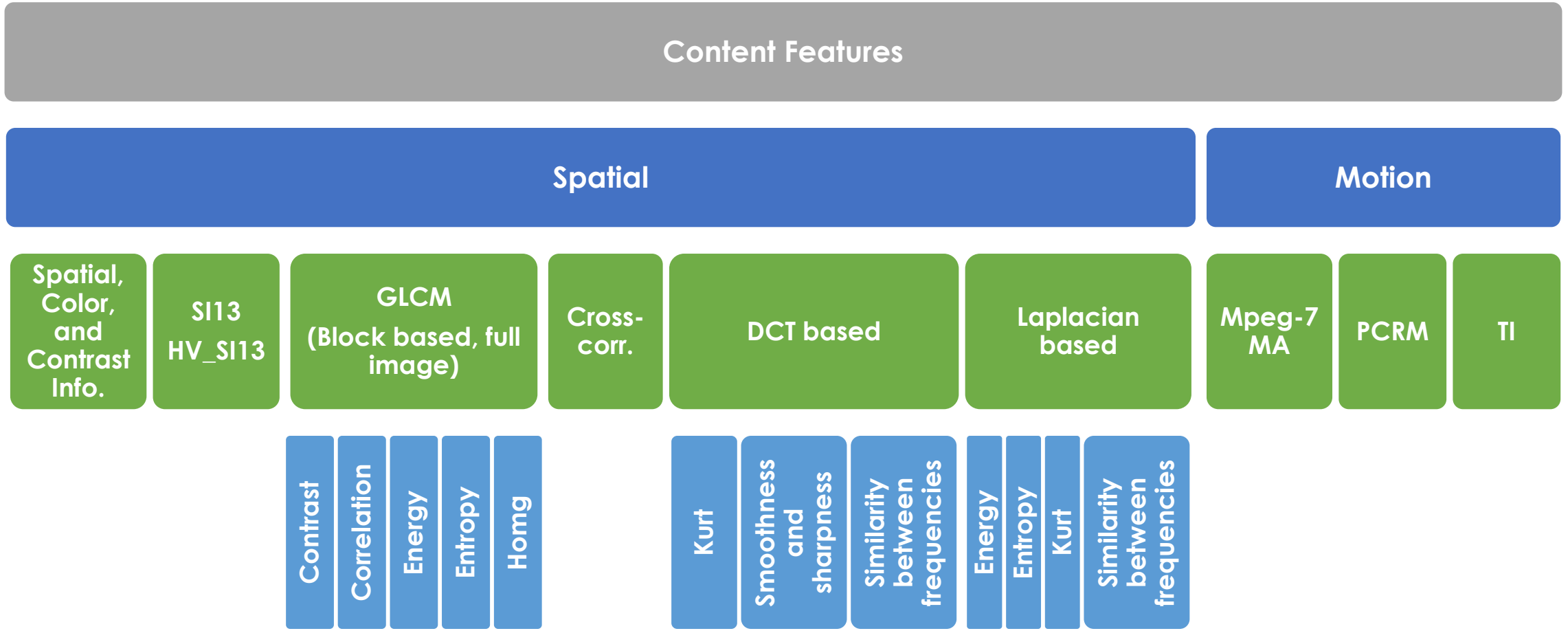
EXTENDED DATABASE 1/4

- 25 loss traces using a 2-state Markov model
  - Packet loss rate: 0.5% and 1%
  - Average burst length: 1, 1.5, 2
- Robust "reference" HM decoder
- Each event affects one slice of the source sequence.
  - The whole frame
  - A slice with a fixed number of macroblocks
  - A slice with a maximum number of bytes

EXTENDED DATABASE 2/4

- Current status:
  - Resolution 960x544
    - 19,840 sequences
    - total 19,840 x 25 = 496,000 scenarios
  - Objective video quality measures: PSNR, SSIM, VIF, VQM, PVQM
    - These measures are not designed for packet loss scenarios
    - But they have been used in literature for such scenarios.
    - The calculation time is feasible.

EXTENDED DATABASE 3/4



EXTENDED DATABASE 4/4

- Extracted from the ten original video sequences.
- Extracted from the luminance frame (Y), and the chrominance frames (Cb and Cr)
- Total features: 209 content features

QUALITY MEASURES DISAGREEMENT 1/4

- Compare each PVS with all the others (within the same sequence)
- Decide about agreement
- $agreement = \begin{cases} 1, & \left|\sum_{Q\in\{PSNR,SSIM,VIF\}} sign(Q(A) - Q(B))\right| = 3 \\ 0, & else \end{cases}$

Agreement

Sequence	PSNR	SSIM	VIF
A	35	0.98	0.93
B	33	0.96	0.92
A-B	+2	+0.02	+0.01

Disagreement

Sequence	PSNR	SSIM	VIF
A	35	0.98	0.92
B	33	0.96	0.93
A-B	+2	+0.02	-0.01

QUALITY MEASURES DISAGREEMENT 2/4

- Reasons of disagreement among quality measurements.

Without loss

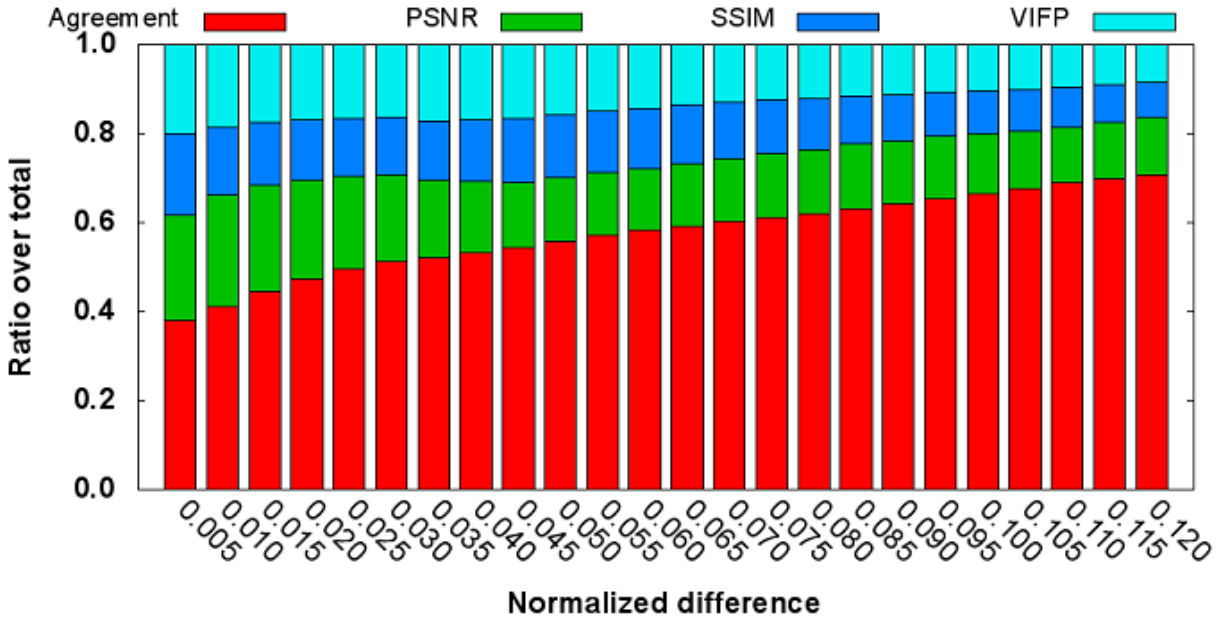
Sequence	Pairs with disagreement	Due to PSNR	Due to SSIM	Due to VIF
src01	3.32%	14.47%	60.72%	24.80%
src02	2.64%	40.74%	45.70%	13.56%
src03	6.27%	61.97%	9.30%	28.73%
src04	4.55%	51.17%	11.76%	37.06%
src05	3.30%	37.89%	18.16%	43.95%
src06	4.99%	28.92%	13.84%	57.24%
src07	6.17%	69.45%	7.41%	23.14%
src08	3.93%	24.58%	59.33%	16.09%
src09	7.65%	20.89%	53.62%	25.49%
src10	3.81%	39.76%	12.55%	47.70%

With loss

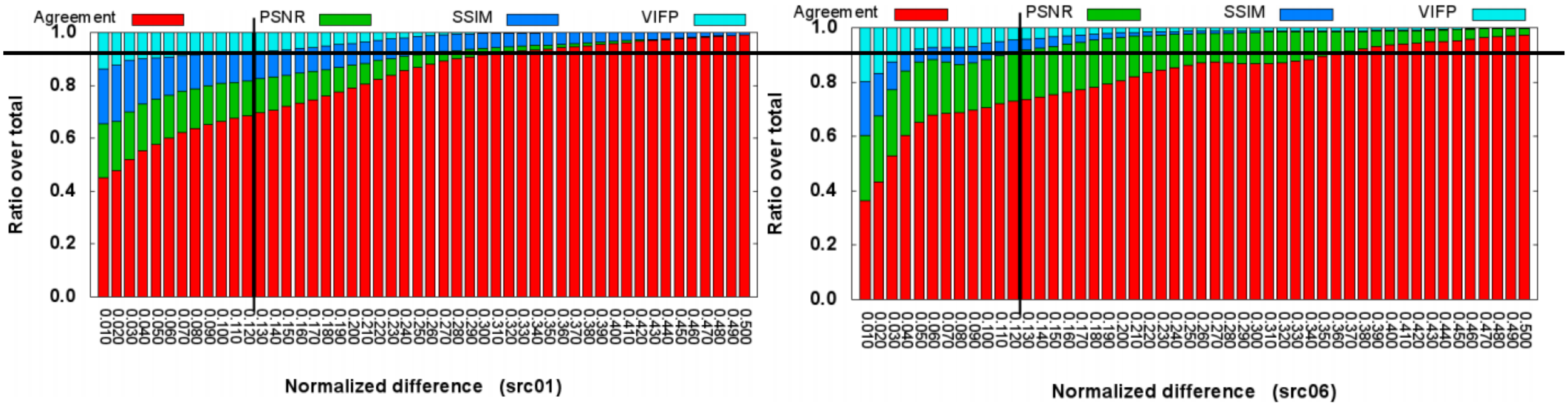
Source	% of disagreement	% due to PSNR	% due to SSIM	% due to VIFP
src01	12.74	38.81	41.60	19.59
src02	4.29	61.37	23.97	14.66
src03	12.07	45.47	26.42	28.11
src04	10.41	57.51	22.55	19.94
src05	4.11	47.26	32.27	20.47
src06	9.98	71.81	12.43	15.76
src07	5.64	65.27	11.89	22.84
src08	5.46	59.19	19.73	21.07
src10	12.44	46.67	32.12	21.21

QUALITY MEASURES DISAGREEMENT 3/4

- Histogram of reason of disagreement (SRC-3) as a function of the normalized difference
- $\hat{d} = \sqrt{\Delta\widehat{PSNR}^2 + \Delta\widehat{SSIM}^2 + \Delta\widehat{VIF}^2}$

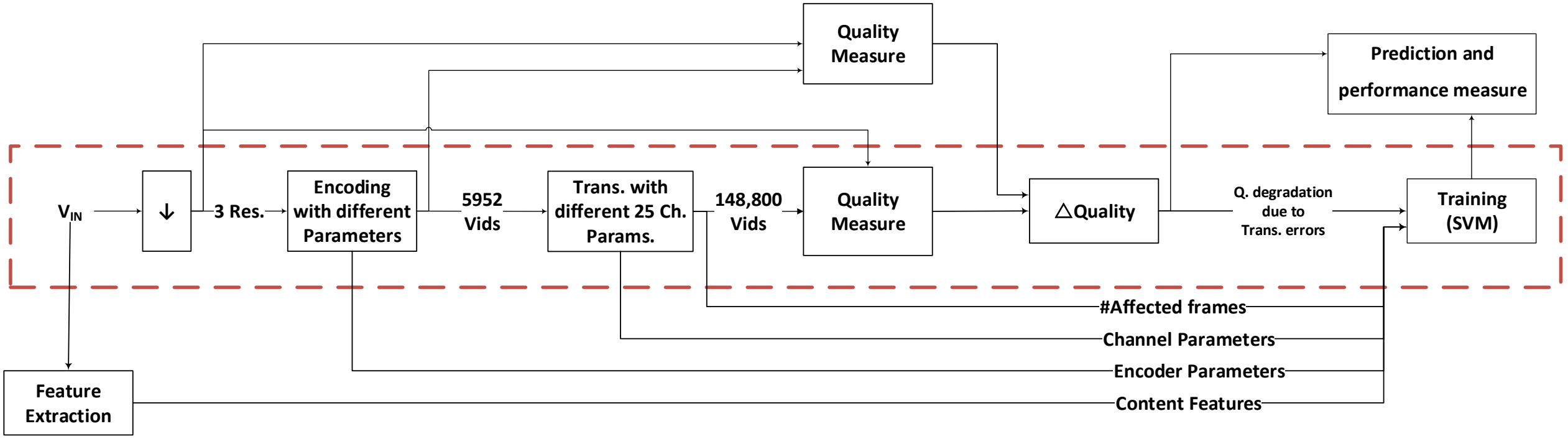


QUALITY MEASURES DISAGREEMENT 4/4



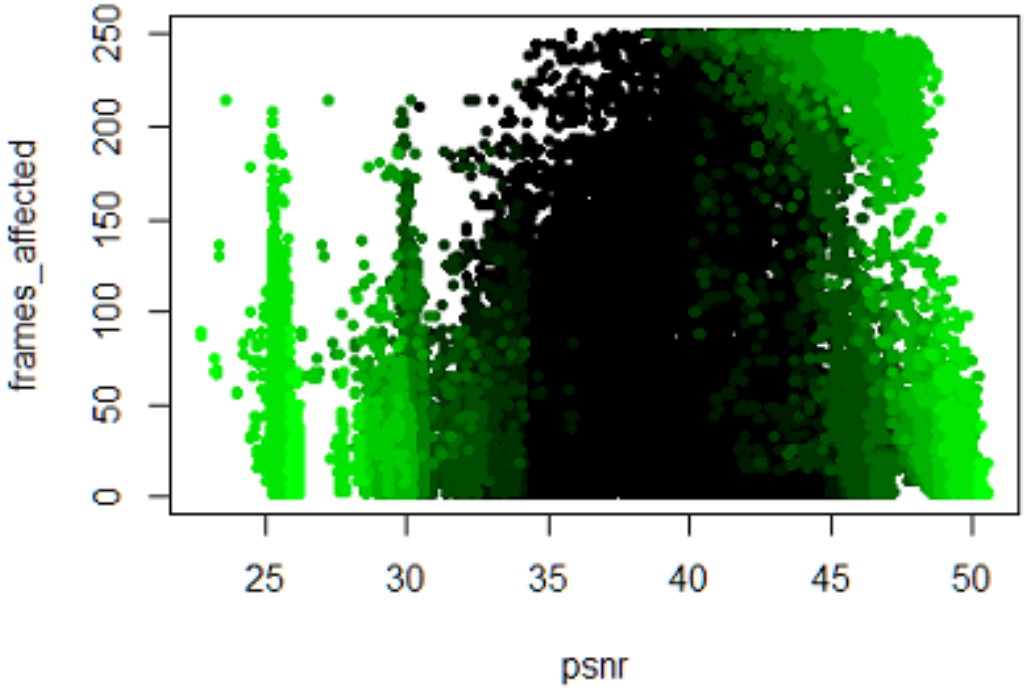
ANALYSIS BASED ON △PSNR PREDICTION 1/4

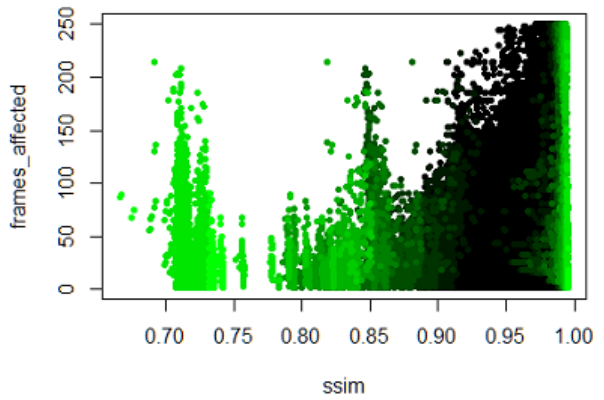
- Is full reference behavior predictable?
- If YES,
  - towards designing a hybrid No-Reference quality metric

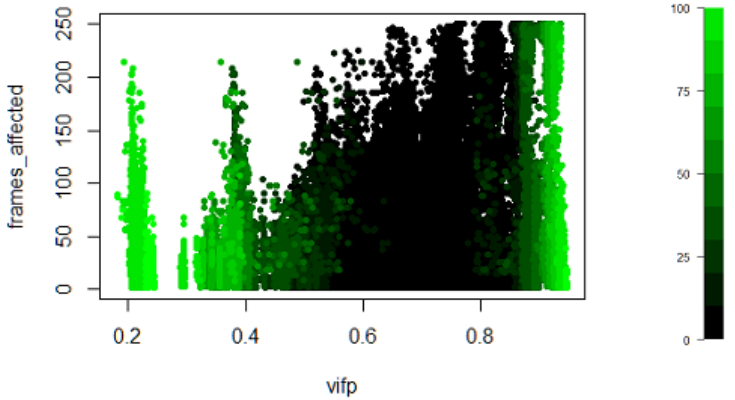


ANALYSIS BASED ON △PSNR PREDICTION 2/4

- Number of affected frame as a feature (SRC-5)
  - The **impact** on the **agreement** between different quality measures







ANALYSIS BASED ON △PSNR PREDICTION 3/4

Machine learning:

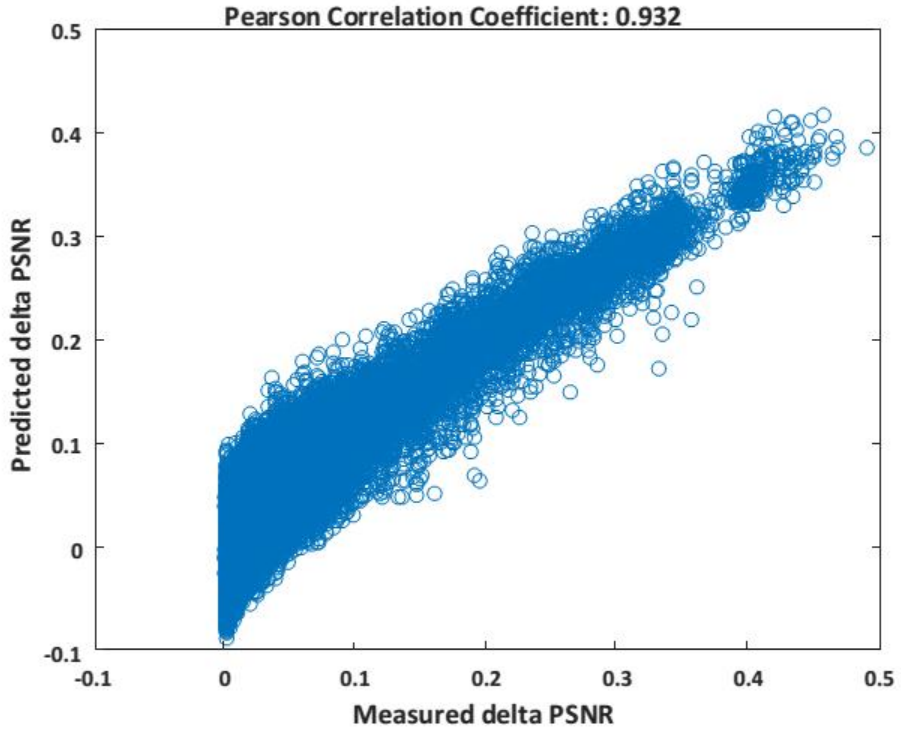
- Epsilon-SVR with radial basis
- 5-fold cross validation

Selected Features

- Encoding Parameters (8)
- Channel Parameters (2)
- Number affected frames (1)
- Content features (7)
  - temporal information
  - correlation of GLCM (2)
  - energy of GLCM
  - entropy of GLCM
  - DCT based smoothness
  - MPEG-7 short length of zero of spatial distribution of the objects

ANALYSIS BASED ON △PSNR PREDICTION 4/4

Performance of predicting △PSNR



Performance of the predicting model

Performance	PCC	SROCC	RMSE
Test data (reduced feature set)	0.9320	0.833	0.0305
Train data (reduced feature set)	0.9310	0.832	0.0305
Test data (All features)	0.9144	0.770	0.0368
Train data (All features)	0.9135	0.769	0.0368

CONCLUSION

- **large scale assessment:** as subjective data is limited.
- Identifying **potential shortcomings** in terms, e.g., **stability and agreement**, of existing video quality metrics.
- **Full reference behavior (PSNR)**, in **loss-impaired** videos, can be **predicted** with high correlation using **content**, **coding**, and **channel** characteristics.

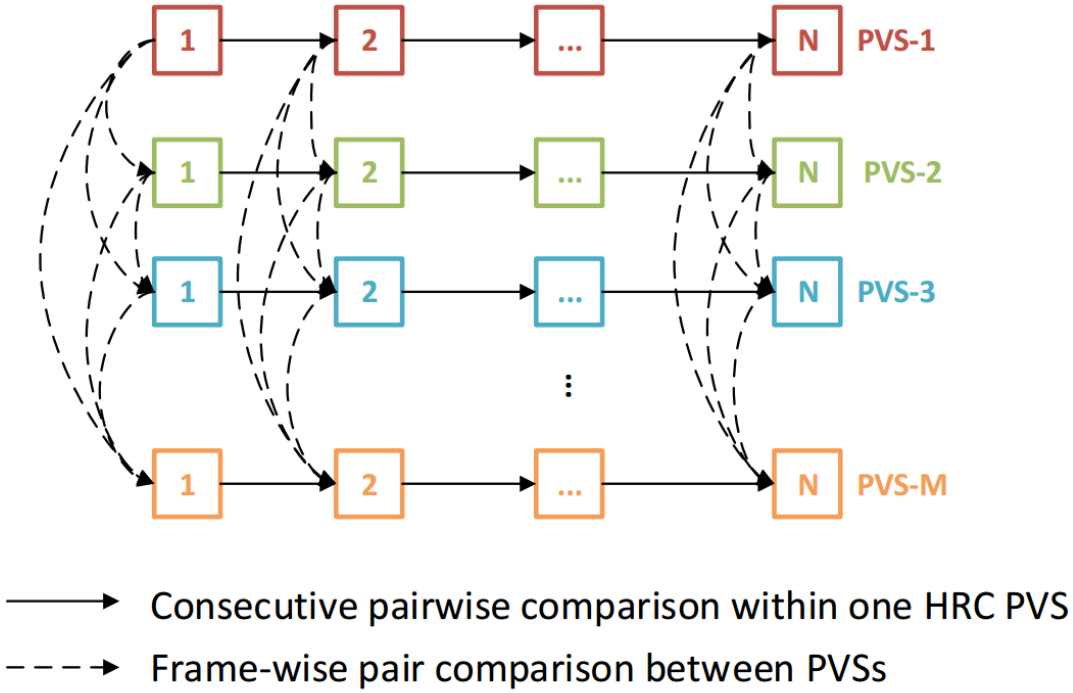
Second part

COMPARING TEMPORAL BEHAVIOR OF FAST OBJECTIVE VIDEO QUALITY MEASURES ON A LARGE-SCALE DATABASE

Enrico Masala, Ahmed Aldahdooh, Glenn Van Wallendael, Marcus Barkowsky

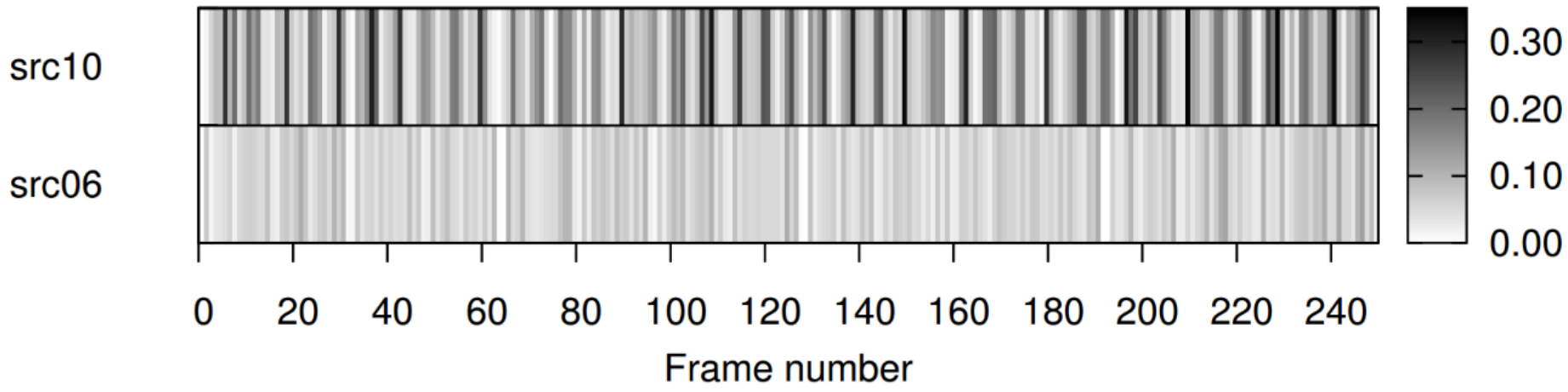
MOTIVATION

- Evaluation of objective measurements that is difficult to achieve in subjective assessment:
  - Consistency within a video sequence is analyzed as well as across video sequences.
    - Characterization of single frame prediction performance in the context of a video sequence.



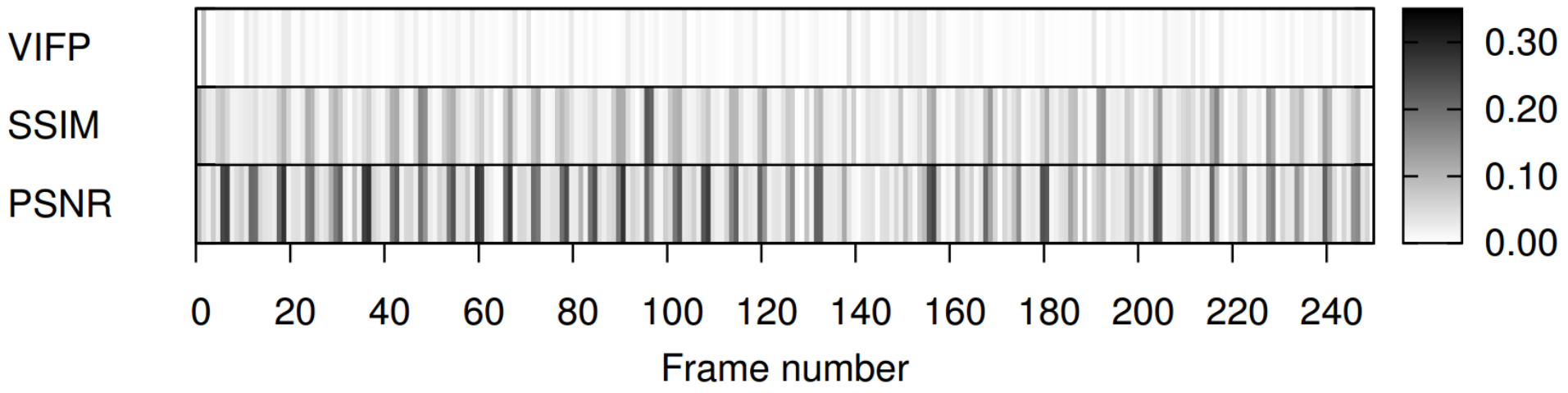
CONSISTENCY MEASURE ON CONSECUTIVE FRAMES

- The darker the bar for a particular frame, the higher the fraction of disagreement between the frame represented on the X-axis and its previous frame.



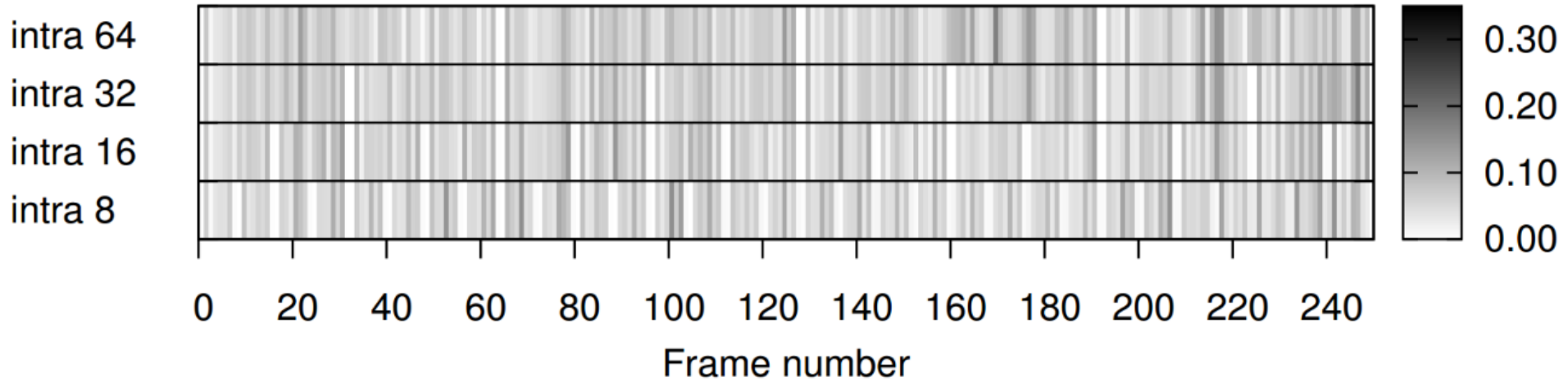
THE IMPACT OF CONTENT

- The majority of disagreement in SRC3 is due to PSNR, while the majority of disagreements in SRC10 is due to SSIM
- depending on the type of the source content, PSNR, SSIM, and VIF can act differently



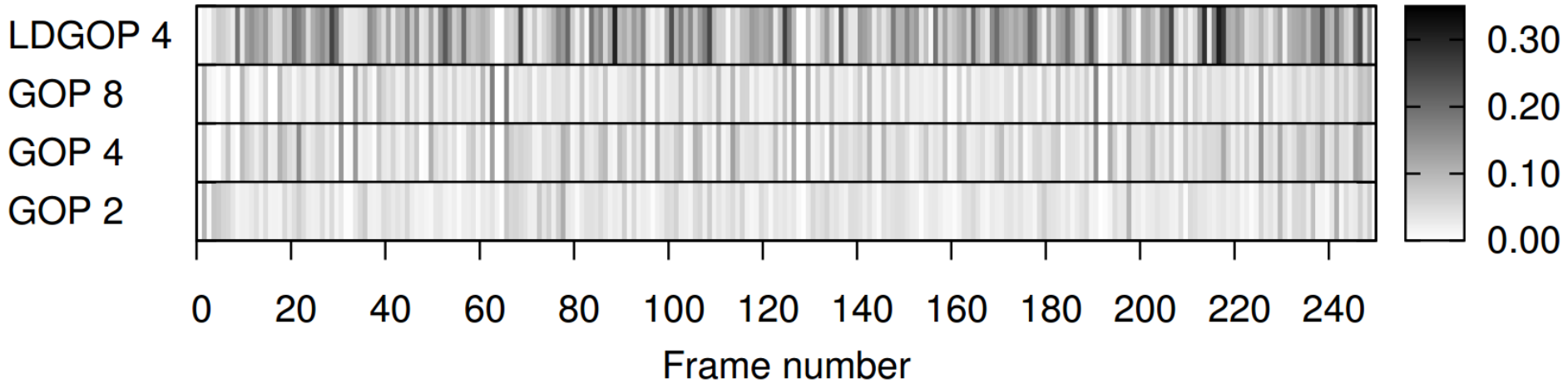
THE IMPACT OF INTRA-PERIOD

- the disagreement fractions between Intra-frames and next/previous frames is very low compared to other frames.
- Similar observations can be made for all Intra-periods (8, 16, 32, and 64) and also for the other source contents. (RSC-6 below)



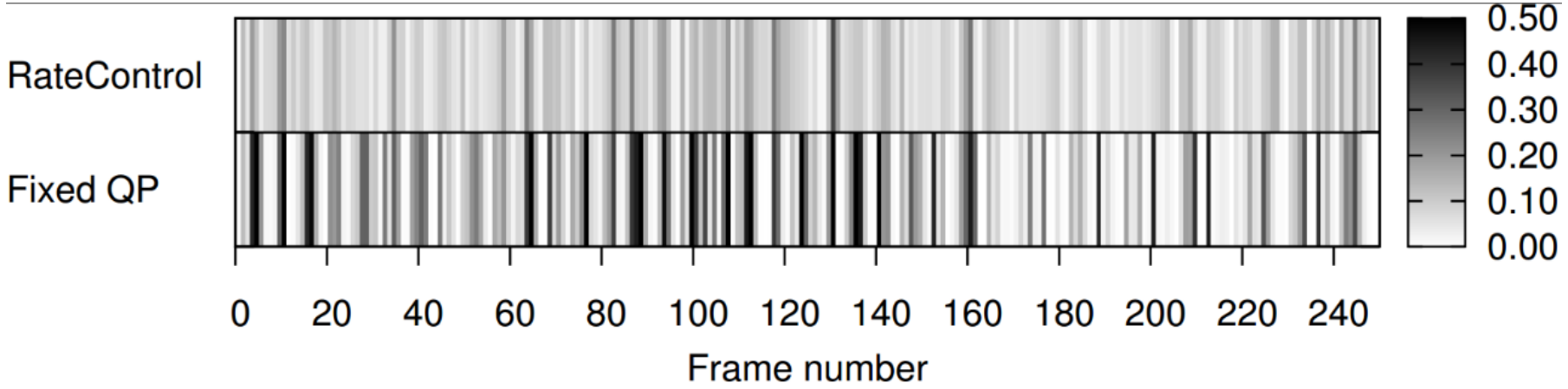
THE IMPACT OF GOP STRUCTURE

- The number of disagreement in the low-delay configuration is higher than the number in the hierarchical coding structures.



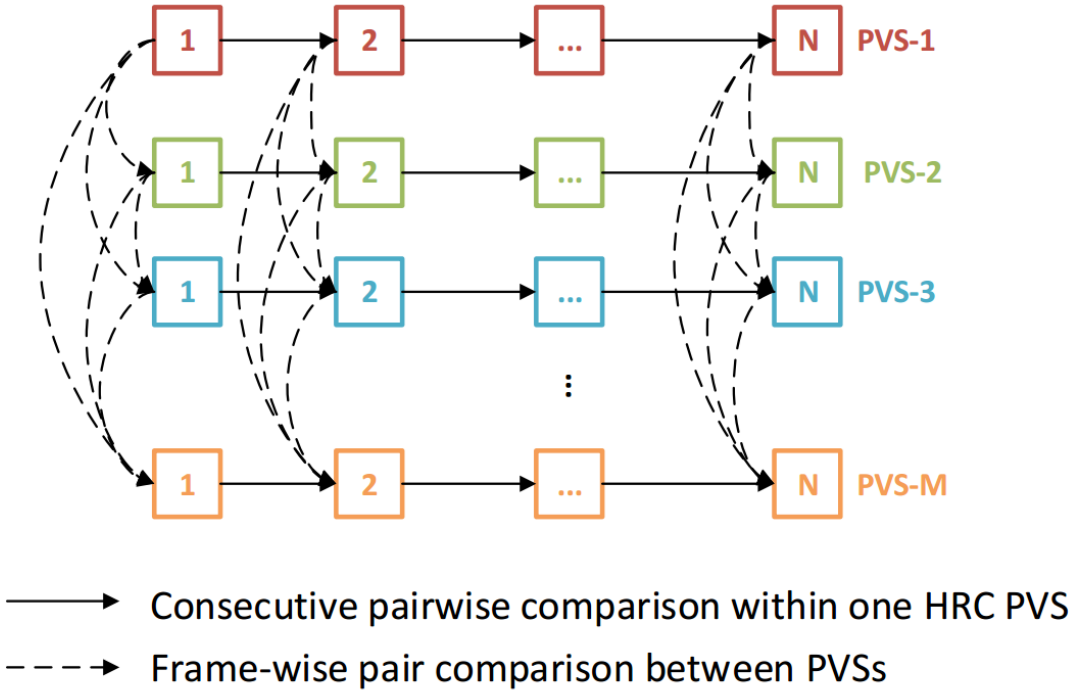
THE IMPACT OF QP AND RATE CONTROL

- Fixed QP:
  - Very low and very high disagreement fractions periodically alternate at the beginning and the middle of the GOP respectively.
- Rate control: not observed.



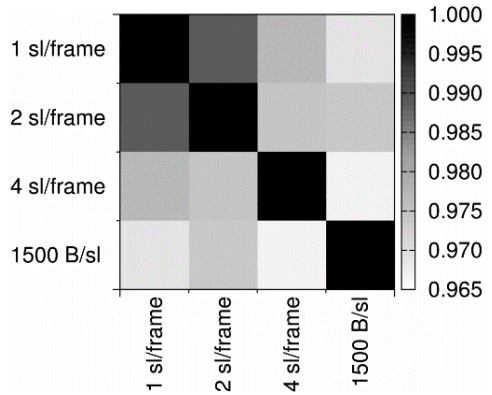
CONSISTENCY WITH RESPECT TO SOURCE CONTENT AND CODING PARAMETERS

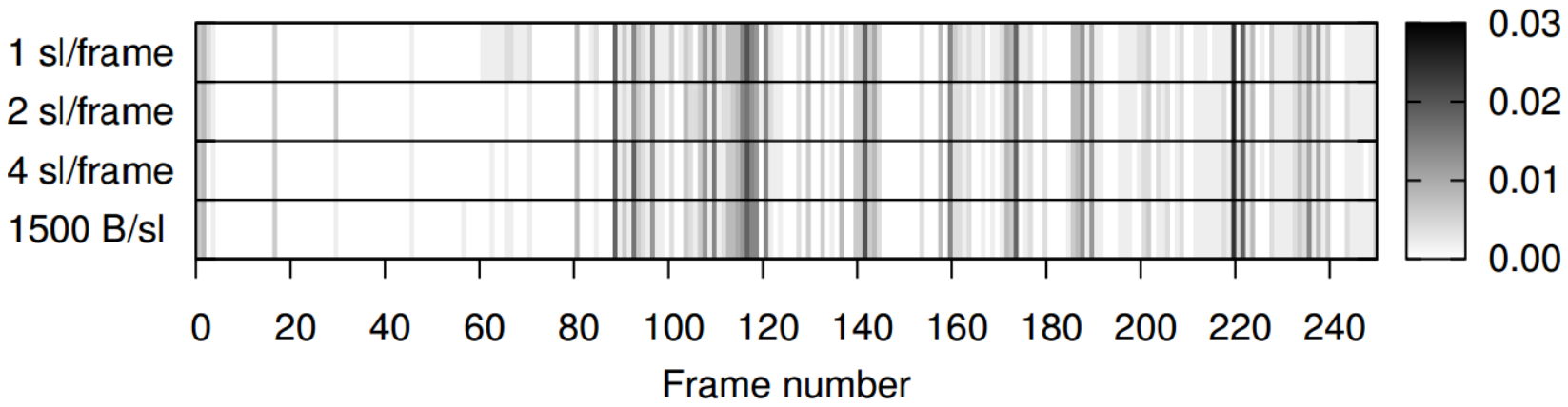
- the agreement of the measures is analyzed across PVS
- we only consider Case Agree, and we investigate if such an agreement at the overall sequence level corresponds to agreement for single frames as well.
- only sequences for which the agreement holds for more than 90% of the frames (Case Agree90).



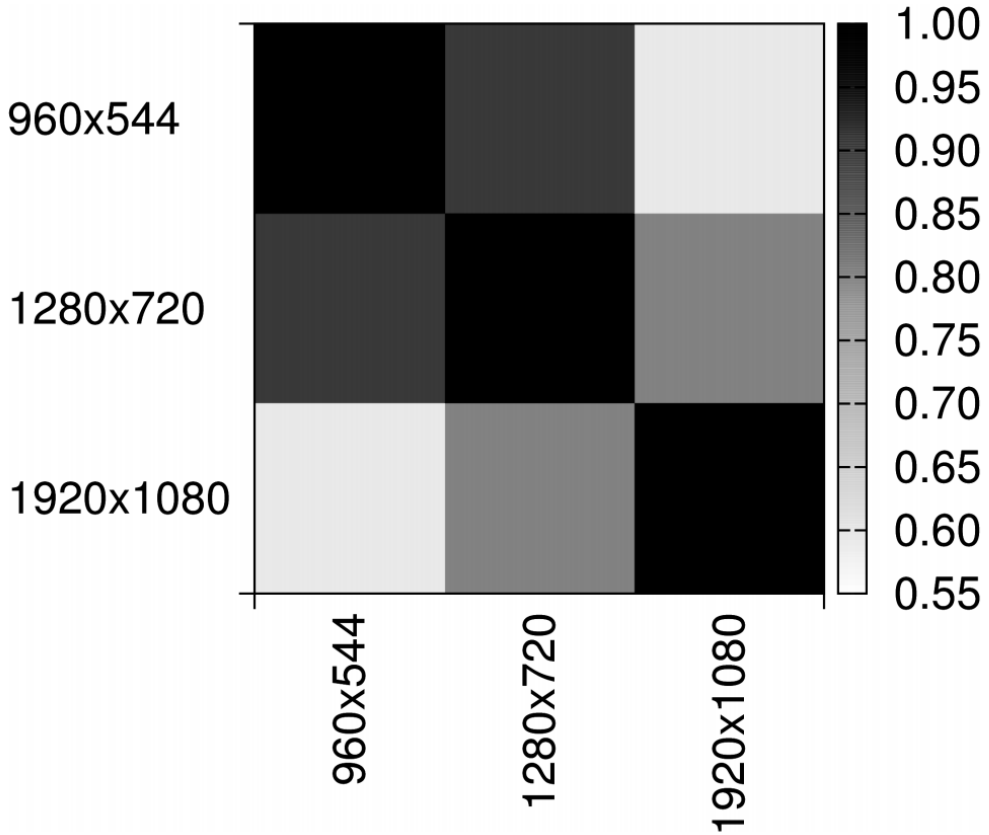
NUMBER OF SLICES PER FRAME.

- the number of slices does not significantly affect the number of disagreement.



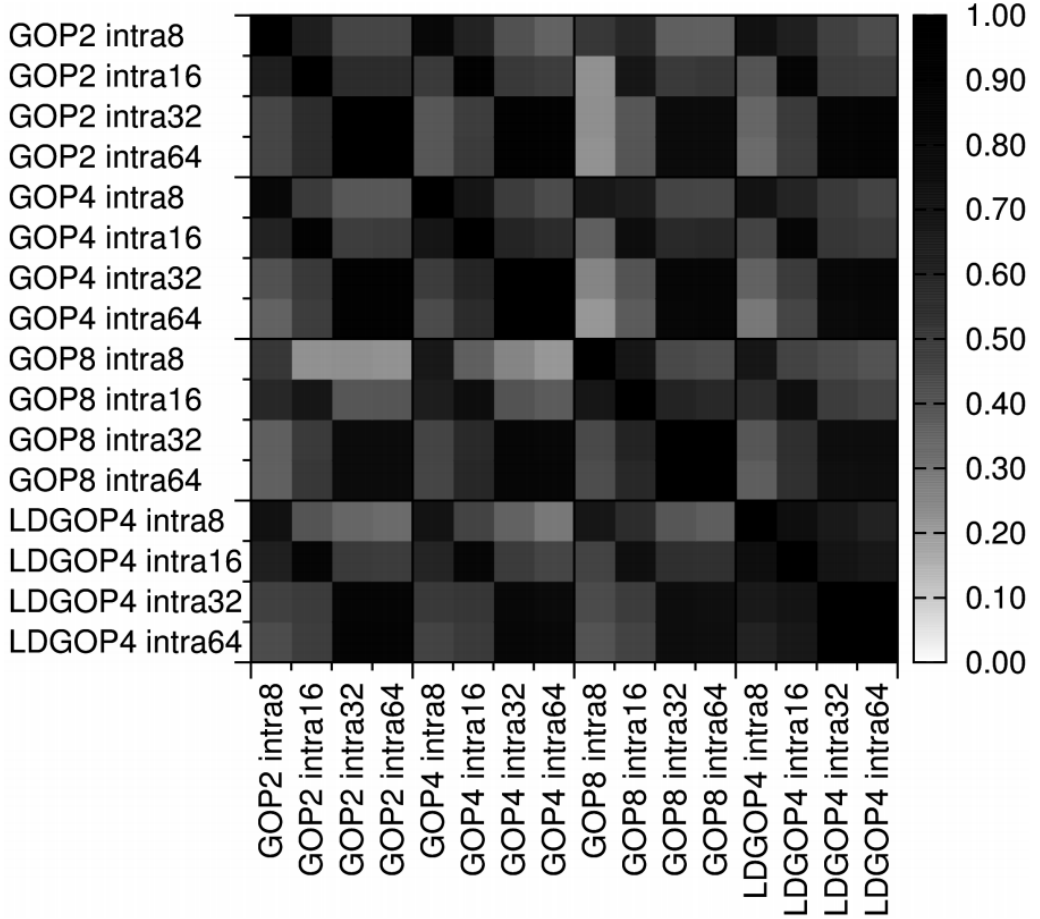


IMPACT OF THE RESOLUTION



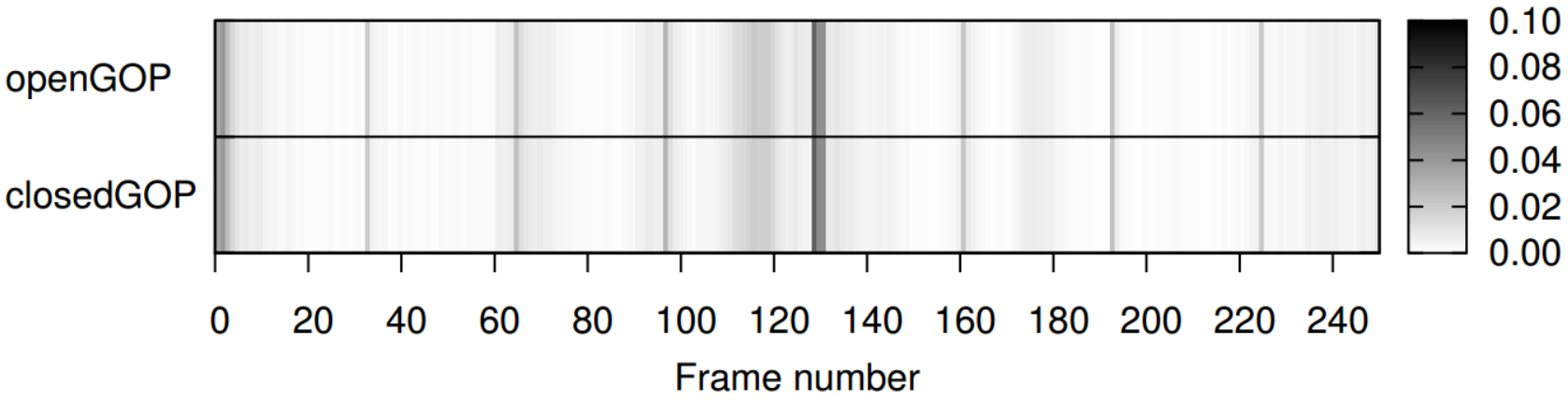
THE INTRA-PERIOD AND THE GOP SIZE

- GOP size is small and the intra period is large, there might be a strong impact on the position of disagreements,
- whereas with the largest GOP size the effect is reduced.
- low-delay GOP configuration (LDGOP) correlation is very high, meaning that the influence of the intra refresh rate is much more reduced



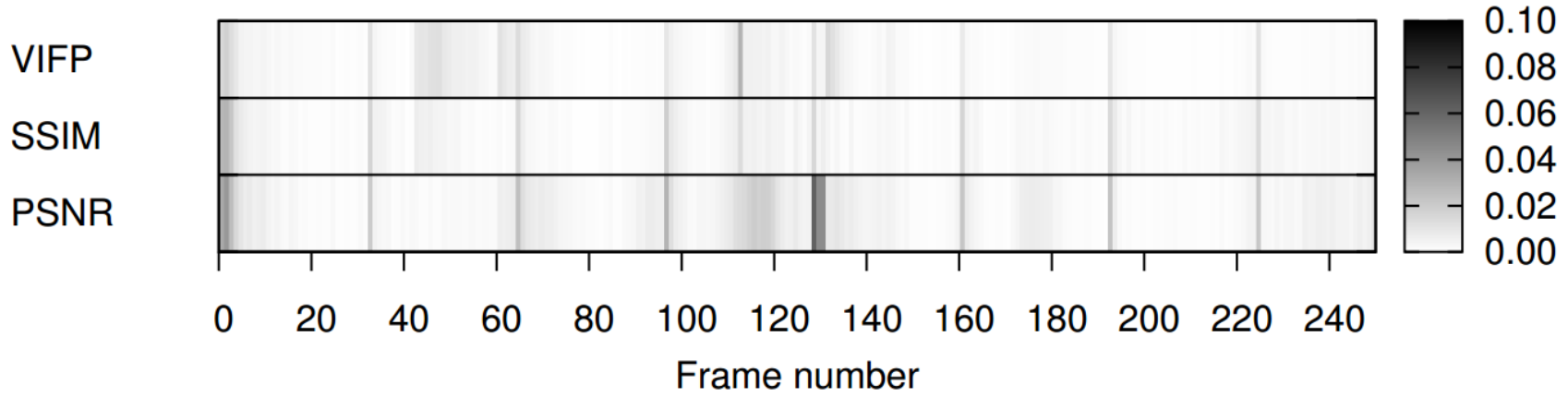
RATE CONTROL VS QP

- in the first part a high fraction of disagreement is visible.
- This can be ascribed to the fact that an initial, fixed, QP is used by the HM rate control algorithm, which then quickly adapts to the requested bitrate.
- Peaks at Intra refresh rate



CAUSE OF DISAGREEMENT

- Behaviors are common for all metrics, e.g., the initial frames and the periodicity of the peaks, others seem to be peculiar of the measures.



CONCLUSION

- disagreement between several objective measures exists on a frame-level even if the measures agree on a sequence level.
- the usage of one single measure may not be sufficient
- pronounced correlation between content characteristics and encoder parameter selection encourages further analysis

FUTURE WORK

- Further work on the largescale database approach requires
  - a significant extension of the samples (Call for contributions)
    - both sequences and algorithm results

- Investigating different temporal pooling strategies for FR measurements
- Can large-scale database be represented with less HRCs?
- Towards NR measurement: using content characteristics (encoding para., stream-based, features/indicators) in estimating the quality.

REFERENCES

- A.Aldahdooh, E. Masala, O. Janssens, G. Van Wallendael and M. Barkowsky, "Comparing simple video quality measures for loss-impaired video sequences on a large-scale database," 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, 2016, pp. 1-6. b.
- A.Aldahdooh, E. Masala, G. Van Wallendael and M. Barkowsky, "Comparing temporal behavior of fast objective video quality measures on a large-scale database," accepted to be presented in picture coding symposium 2016 conference.

THANKS

- Questions