Talking Video Heads

Saving Streaming Bitrate by Adaptively Applying Object-based Video Principles to Interview-like Footage

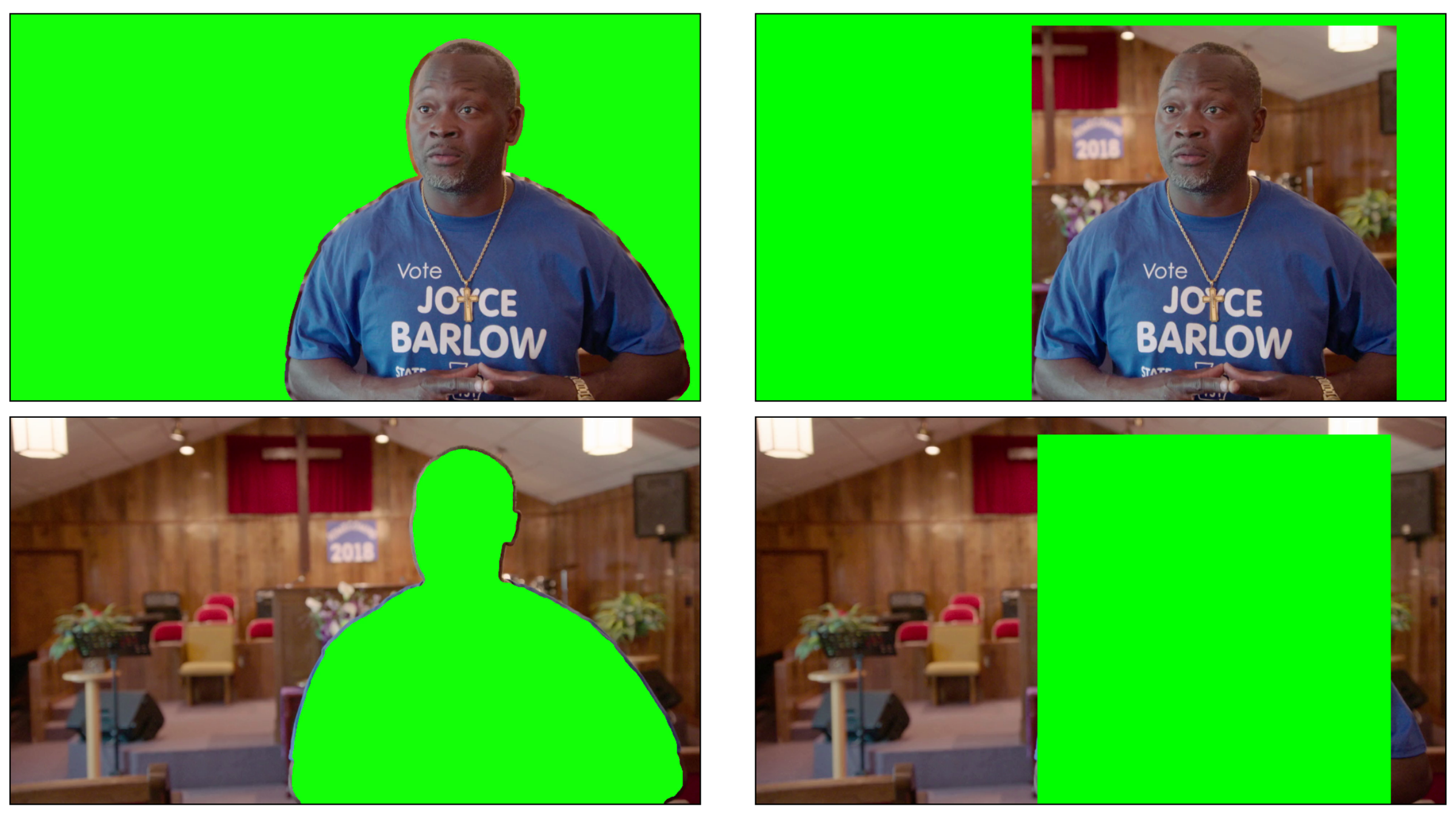
Maarten Wijnants Sven Coppers Gustavo Rovelo Ruiz Peter Quax Wim Lamotte

1. Research Question

For *talking heads* video, **does reducing the quality of the background have a perceptual impact**? If not, **bitrate savings** and **OTT streaming cost reduction** become feasible.

2. Object-Based Video (OBV)

Contour (CT) Bounding Box (BB)



3. Implementation

We applied H.264 video compression (x264 encoder, **high profile**, `veryslow` encoding preset) and used **Constant Rate Factor (CRF)** compression mode. Quality degrades as CRF rises. H.264's default CRF value is 23.

4. Method

Two-step experimental design combining **(bespoke) pre-study** with standard **Absolute Category Rating** test:

Experiment 1:

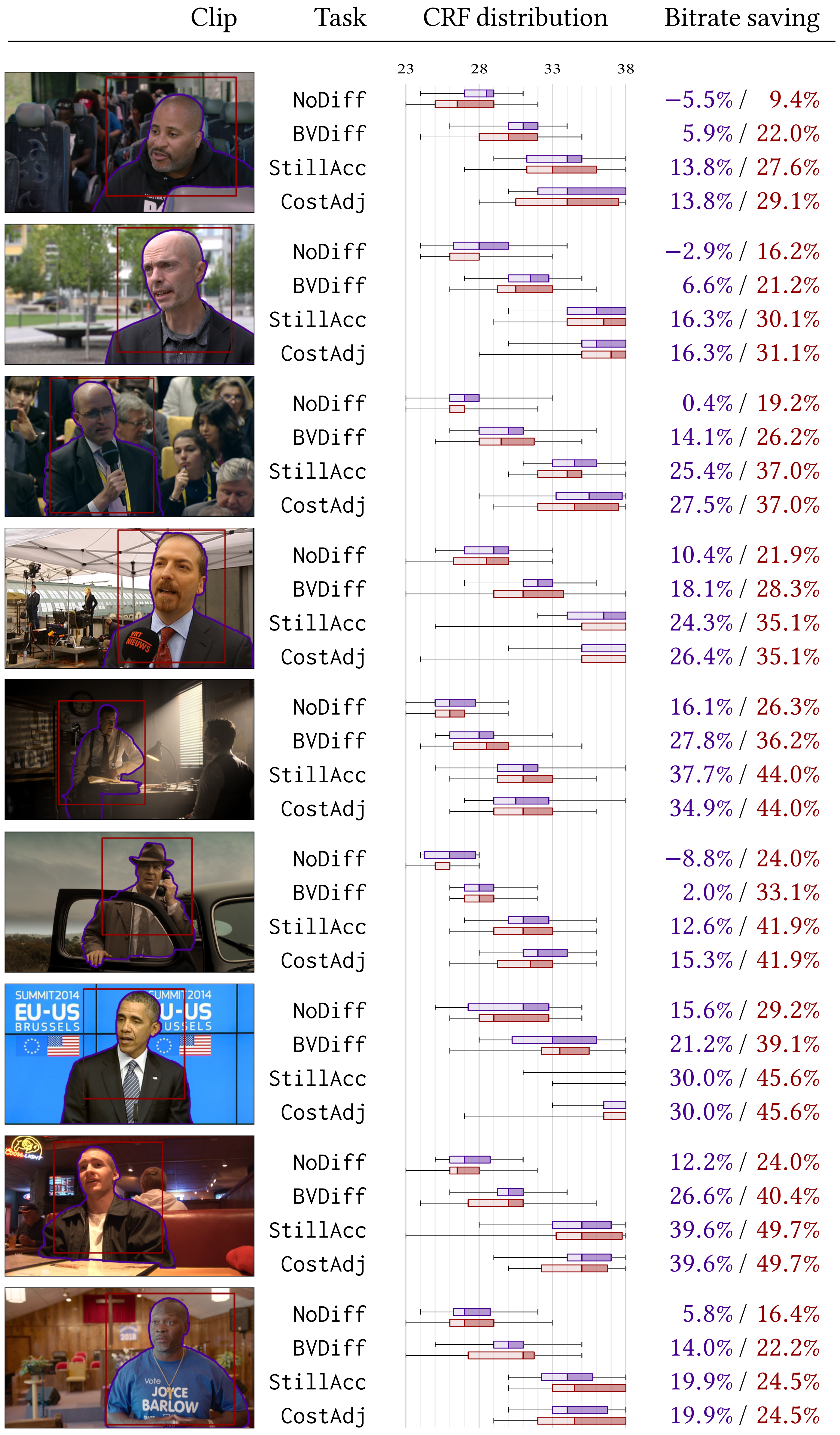
- **Dual screen setup** to perceptually compare traditional (at CRF 23) versus object-based video coding
- OBV foreground always streamed at maximal quality (CRF 23)
- **OBV background quality could be freely adjusted** (CRF 23 ⟶ CRF 38)
- Step-by-step decrease background quality to identify thresholds: **No Difference** (`NoDiff`), **Barely Visible Differences** (`BVDiff`), **Still Acceptable** (`StillAcc`), **Cost Adjusted** (`CostAdj`)
- Implemented with **OBV-familiar participants** (n=18)

Experiment 2:

- Classic ACR study involving five pre-rendered versions of each video: a **traditional encode at CRF 23** (`TR`), complemented with **two contour-based** (`CT`) and **two bounding boxed** (`BB`) OBV versions
- The four OBV versions combined a **CRF 23 foreground** quality with the (rounded down) **median background CRF value** corresponding to respectively the **`NoDiff` and `BVDiff` task as elicited in Experiment 1**
- Implemented with **OBV-agnostic participants** (n=30)

5. Experiment 1: Results

Contour versus **bounding box** CRF values plus median bitrate savings relative to `TR` at CRF 23:



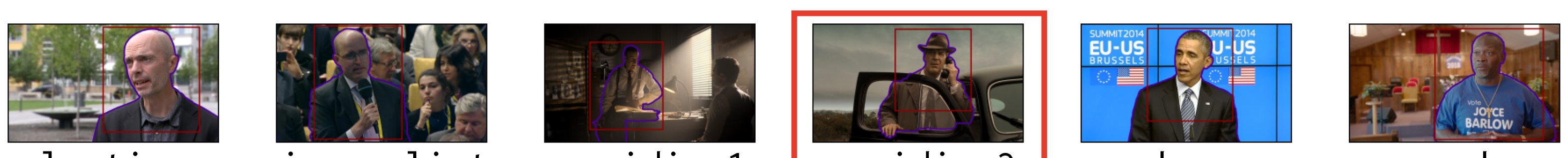
Task	Bitrate saving
NoDiff	−5.5% / 9.4%
BVDiff	5.9% / 22.0%
StillAcc	13.8% / 27.6%
CostAdj	13.8% / 29.1%
NoDiff	−2.9% / 16.2%
BVDiff	6.6% / 21.2%
StillAcc	16.3% / 30.1%
CostAdj	16.3% / 31.1%
NoDiff	0.4% / 19.2%
BVDiff	14.1% / 26.2%
StillAcc	25.4% / 37.0%
CostAdj	27.5% / 37.0%
NoDiff	10.4% / 21.9%
BVDiff	18.1% / 28.3%
StillAcc	24.3% / 35.1%
CostAdj	26.4% / 35.1%
NoDiff	16.1% / 26.3%
BVDiff	27.8% / 36.2%
StillAcc	37.7% / 44.0%
CostAdj	34.9% / 44.0%
NoDiff	−8.8% / 24.0%
BVDiff	2.0% / 33.1%
StillAcc	12.6% / 41.9%
CostAdj	15.3% / 41.9%
NoDiff	15.6% / 29.2%
BVDiff	21.2% / 39.1%
StillAcc	30.0% / 45.6%
CostAdj	30.0% / 45.6%
NoDiff	12.2% / 24.0%
BVDiff	26.6% / 40.4%
StillAcc	39.6% / 49.7%
CostAdj	39.6% / 49.7%
NoDiff	5.8% / 16.4%
BVDiff	14.0% / 22.2%
StillAcc	19.9% / 24.5%
CostAdj	19.9% / 24.5%

6. Experiment 2: Results

In terms of **objective quality metrics** (averaged over content corpus), traditional encoding at CRF 23 outperforms its four OBV competitors:

	Y-PSNR	SSIM	VMAF
TR_CRF23	41.993 ± 1.148	0.981 ± 0.007	93.808 ± 2.291
CT_NoDiff	41.196 ± 1.297	0.978 ± 0.007	90.740 ± 1.910
CT_BVDiff	40.671 ± 1.414	0.976 ± 0.007	88.932 ± 2.558
BB_NoDiff	38.508 ± 0.684	0.977 ± 0.007	91.055 ± 1.523
BB_BVDiff	38.212 ± 0.795	0.976 ± 0.007	89.476 ± 1.972

However, the ACR **Mean Opinion Scores (MOS)** and Standard deviation of Opinion Scores (SOS) prove that these **objective quality differences are not necessarily perceived** by human viewers:



	elections	journalist	meridian1	meridian3	obama	preacher	average
TR_CRF23	3.90 ± 0.88	3.03 ± 0.76	4.00 ± 0.91	4.03 ± 0.76	3.27 ± 0.78	3.17 ± 0.83	3.57 ± 0.92
CT_NoDiff	3.97 ± 0.85	2.90 ± 0.92	3.67 ± 0.99	3.40 ± 0.97	3.33 ± 0.84	3.20 ± 1.00	3.41 ± 0.98
CT_BVDiff	4.00 ± 0.87	2.73 ± 0.78	3.70 ± 0.95	3.57 ± 1.04	3.43 ± 0.77	3.13 ± 1.01	3.43 ± 0.99
BB_NoDiff	3.73 ± 0.91	2.77 ± 0.86	3.70 ± 1.06	3.47 ± 0.94	3.13 ± 1.07	3.17 ± 0.83	3.33 ± 1.00
BB_BVDiff	3.73 ± 0.91	2.70 ± 0.79	3.57 ± 1.07	3.10 ± 1.12	3.00 ± 0.98	3.17 ± 1.02	3.21 ± 1.04

Only `meridian3` showed statistically significant differences (non-parametric Friedman test, Bonferroni corrections): `TR` vs `BB_BVDiff` ($p < 0.005$, $r = 0.44$), `TR` vs `BB_NoDiff` ($p < 0.01$, $r = 0.33$), and `TR` vs `CT_NoDiff` ($p < 0.01$, $r = 0.34$). The `meridian` clips were the only **filmic videos** in our corpus; movie content poses high subjective quality requirements*. Without `meridian`, the MOS averaging becomes:

	average w/o meridian3	average w/o meridian1/3
TR_CRF23	3.47 ± 0.92	3.34 ± 0.87
CT_NoDiff	3.41 ± 0.98	3.35 ± 0.98
CT_BVDiff	3.40 ± 0.98	3.33 ± 0.97
BB_NoDiff	3.30 ± 1.01	3.20 ± 0.98
BB_BVDiff	3.23 ± 1.02	3.15 ± 0.99

* Song et al., "Saving Bitrate vs. Pleasing Users: Where is the Break-even Point in Mobile Video Quality?", 2011.

7. Conclusions

1. For the non-movie content in our corpus, **contour-based OBV lowers bitrate requirements by 14% on average** (compared to frame-based H.264 video coding at CRF 23) **without incurring statistically significant penalties w.r.t. perceived quality**; average MOS difference is as small as 0.01 on a 5-point categorical scale
2. OBV-aware viewers can incur quite extensive background quality reductions (cf. `StillAcc` in Exp. 1)
3. **Bounding boxed OBV** is economically attractive (w.r.t. production cost) plus **yields substantial bitrate bonuses**
4. OBV works well with classic *talking heads* footage, is less compatible with movie-like content
5. **Spatiotemporal compression artifacts** like time-varying blockiness were found to be **extremely detrimental and frustrating** in terms of perceived video quality

Acknowlegdments: At the time this research was conducted, Maarten Wijnants was funded by a VLAIO InnovationMandate (project number HBC.2016.0625), co-sponsored by Androme. Sven Coppers is funded by the Special Research Fund (BOF) of Hasselt University (R-8150). We thank our colleague dr. Davy Vanacken for his methodological advice and our shepherd prof. Wei Tsang Ooi for his help and support.