Subjective and objective evaluation of deepfake videos

Pavel Korshunov and Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland

Motivation

How 'good' the deepfakes are at 'fooling' the humans and machines?

- How realistic are the automatically generated deepfakes?
- Do all deepfakes look the same?
- Lack of comprehensive subjective studies.
- How well the same deepfakes fool algorithms?
- How different algorithms are from humans?
- Lack of comparison between humans and machines.

Dataset and experiments

- Dataset
  - Pre-selected 120 videos from Facebook dataset (from Kaggle competition)
  - 60 deepfakes in five categories of difficulty
  - 60 corresponding real videos
- Subjective study
  - Crowdsourcing scenario (uncontrolled environment)
  - 57 subjects with about 20 answers per video
  - On average, spent 25s on each 10s video
  - ANOVA test: deepfake categories are significantly different
- Objective study
  - Xception and EfficientNet-B4 networks
  - Pre-trained on Google and Celeb-DF
  - The same videos as in subjective study
  - Threshold at FAR=10% on Dev sets

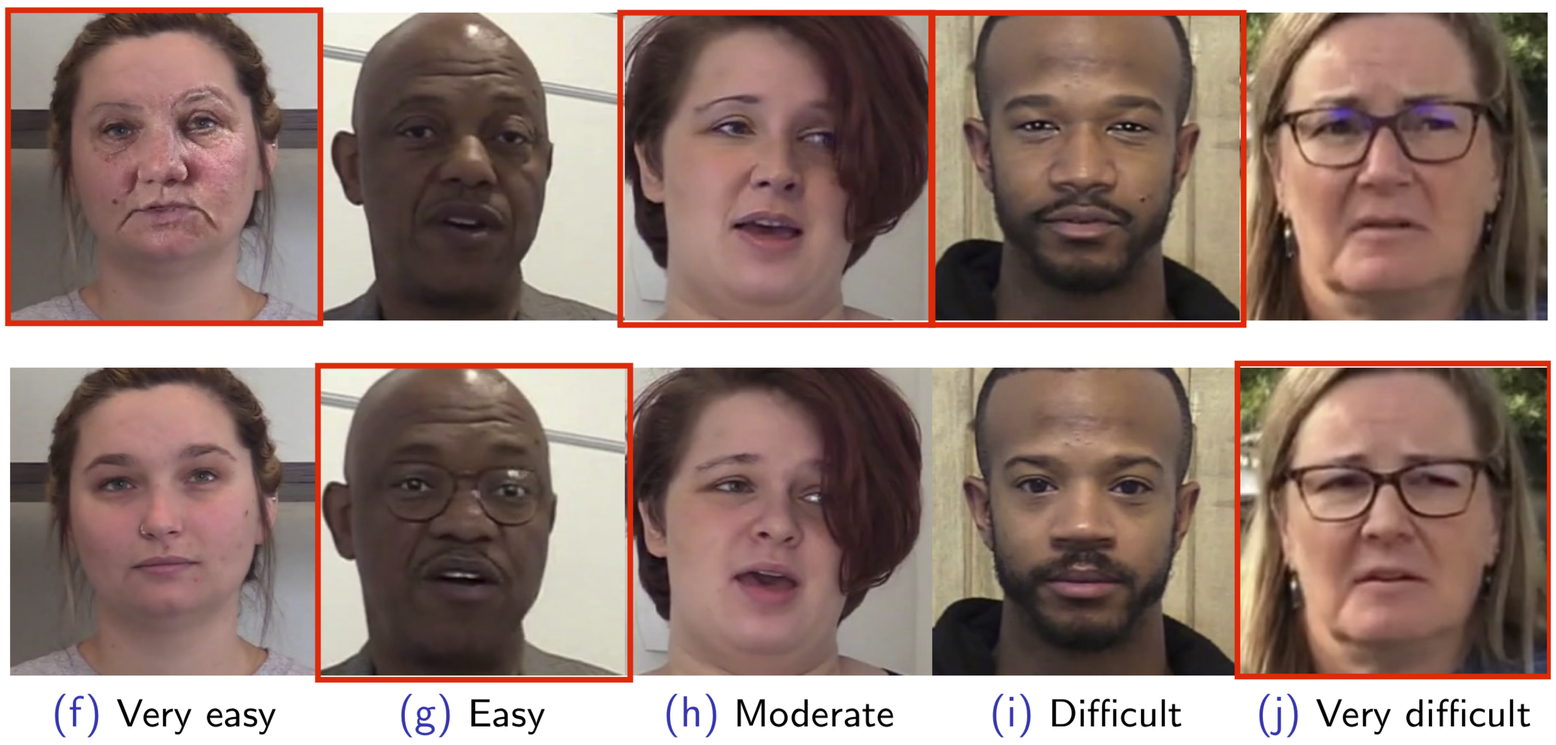


Figure: Real and deepfake videos manually selected from Facebook dataset (Deepfakes are highlighted in red).

Subjective evaluation results

People are confused by good quality deepfakes in 75.5% of cases

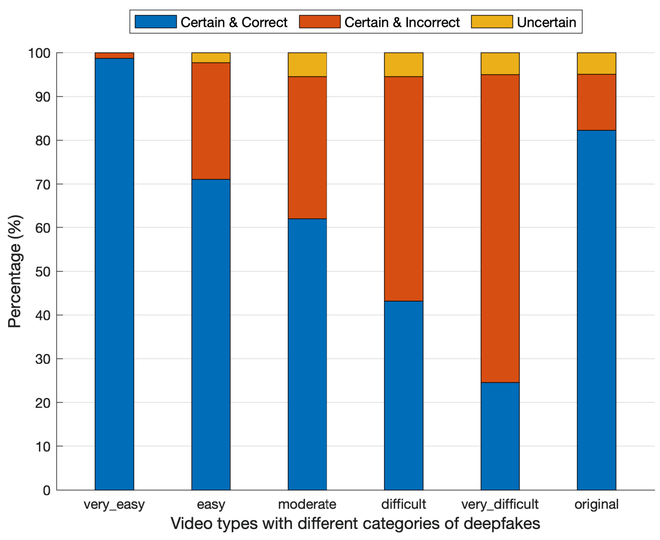


Figure: Average answers per each category.

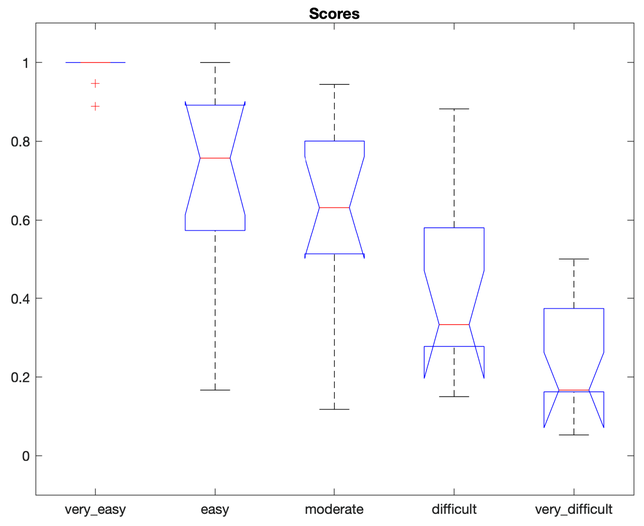


Figure: Median scores with confidence intervals.

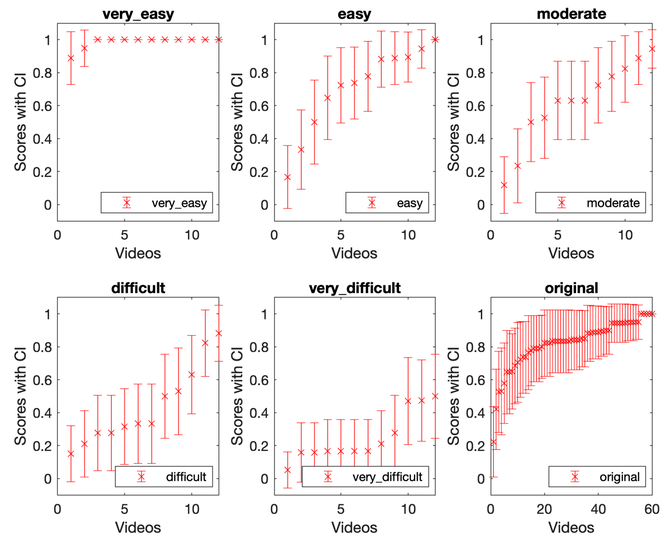
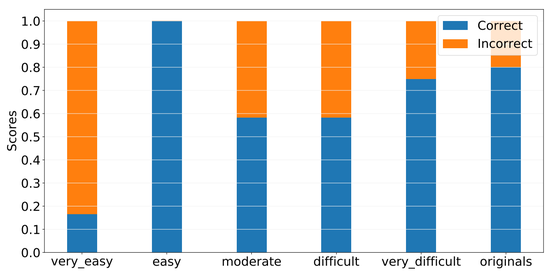


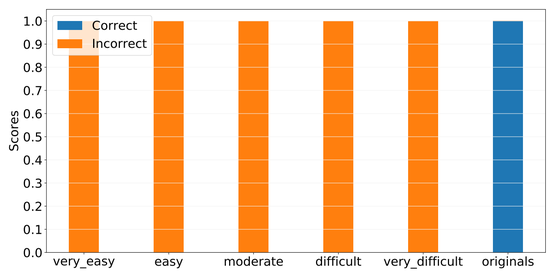
Figure: Average scores for each video and category.

The results for algorithms

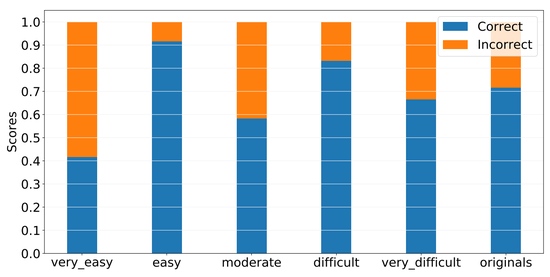
The algorithms struggle to detect many videos that look obviously fake to humans



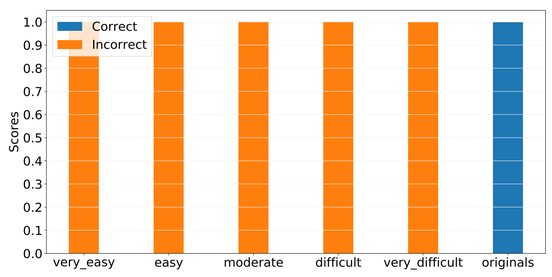
(a) EfficientNet trained on Google



(b) EfficientNet trained on Celeb-DF



(c) Xception trained on Google



(d) Xception trained on Celeb-DF

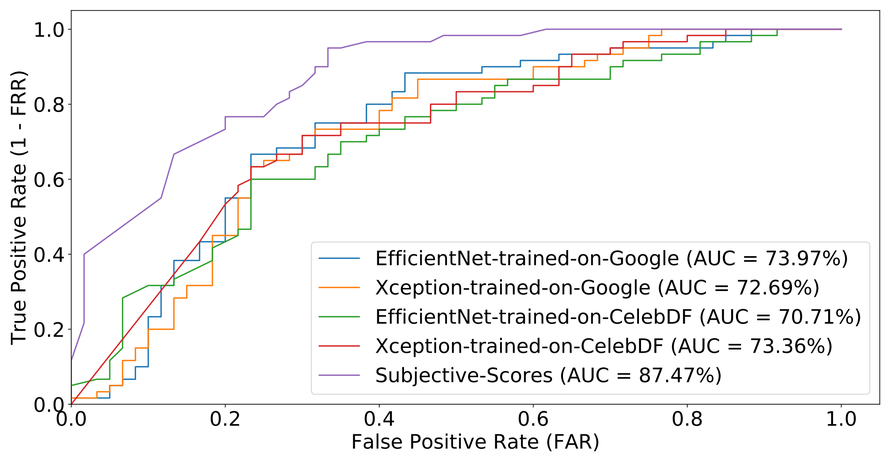


Figure: ROC curves for humans and algorithms.

- Deepfakes fool both human and machines
- Machine vision is very different from human vision

Contact: Pavel.Korshunov@idiap.ch
Source code: https://gitlab.idiap.ch/bob/bob.paper.subjective-deepfakes