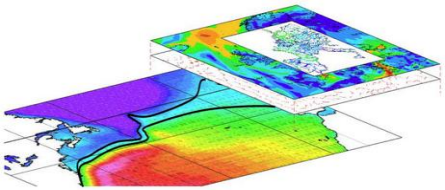


ICLR 2017

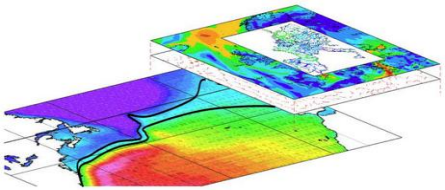
See.4C

European Commission

Spatio-temporal forecasting from short clips of faces: Dataset

Sergio Escalera, UB, CVC, ChaLearn vice-president, IAPR TC-12 chair, head of HuPBA group
Julio Jacques Junior
Xavier Baró

22/4/2017



videos

See.4C

European Commission

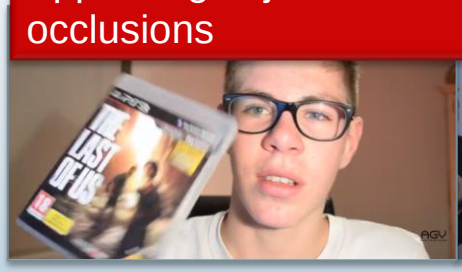
149 *talking-to-camera* videos from different sources

Quality: 720p HD @ 25 FPS

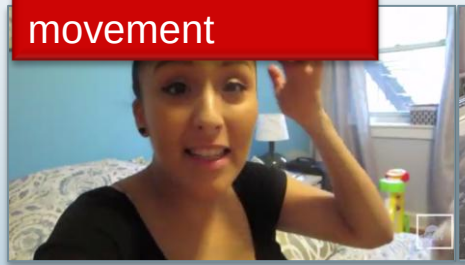
Total duration: 193,510 seconds (4,837,750 frames)

Illumination conditions


Appearing objects & occlusions



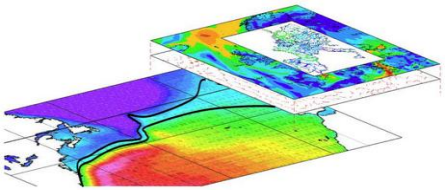
Camera movement




Ethnics & skin color


Upper and full body


Filters and artifacts

Dataset

See.4C

European Commission

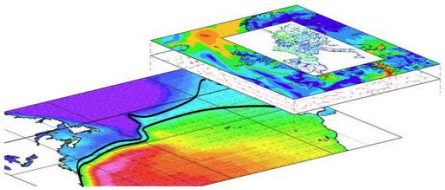
Target dataset:

Single person facing the camera (teleconference scenario)

Grey level low resolution (32x32) images and fast sampling rate (25 fps)

To deliver a realistic task, which can be completed with the computational constraints imposed by a hackathon: training and testing in a reduced period of time.

Image samples of the proposed dataset.



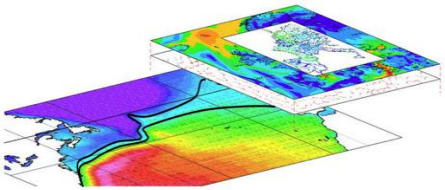
Dataset

See.4C

European Commission

Procedure to obtain valid videos

- We obtained about 48.000 non-overlapped video clips of 5 seconds each.
- Viola & Jones face detector to detect faces
- Clustering of detected regions. We selected 50% of faces nearest to centroid to compute their mean position, width and height
- 2 times height of detected faces is used to define a square region to subtract from initial video



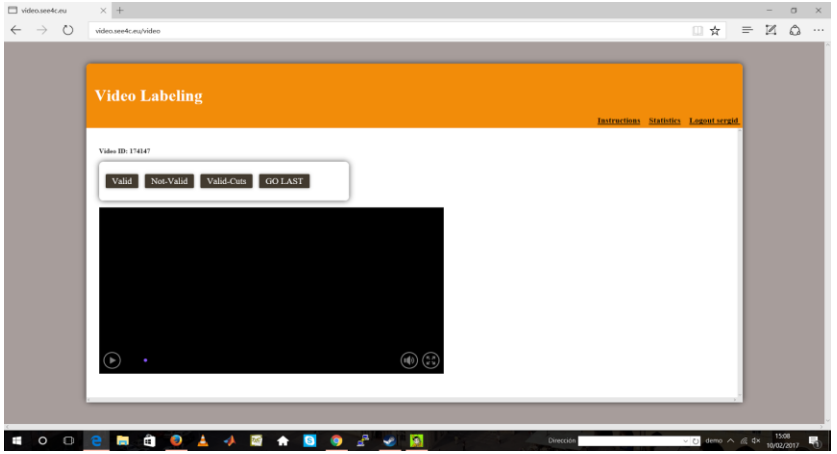
Dataset

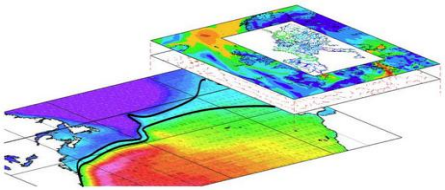
See.4C

European Commission

Final data selection

- We use an app to do the final validation
- **SANITY CHECK!** We check that the clips have **no cuts**, **no camera movement**, and the face remains the 100% of the time in the video (although **may present partial occlusions**)
- Finally, videos are converted to mp4 and, grayscale and 32x32 pixels resolution
- **We selected a set of valid clips for ICLR see4c workshop-hackaton**





Dataset

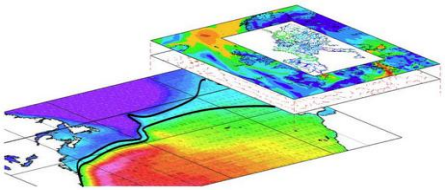
See.4C

European Commission

How sample videos look?

- Focused on face region
- Different inner face (expressions) and head movements





Dataset

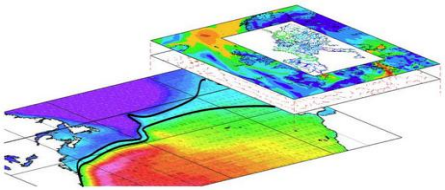
See.4C

European Commission

Alternative databases

Prediction of future frames in a video sequence, employed in: “Lotter, W.; Kreiman, G.; Cox, D. *Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning*. ArXiv, 2016.”

Daset	Source	Description	Annotation
KITTI	A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: *The kitti dataset.* International Journal of Robotics Research (IJRR), 2013.	Captured by a roof-mounted camera on a car driving around an urban environment in Germany. Sequences of 10 frames were sampled from the “City”, “Residential”, and “Road” categories.	Object annotations (3D bounding-box tracklets)
Pedestrian dataset	P. Dollár, C. Wojek, B. Schiele, and P. Perona. *Pedestrian detection: A benchmark.* In CVPR, 2009.	10 hours of 640x480 30Hz video taken from a vehicle driving through regular traffic in an urban environment.	Pedestrian bounding boxes (with temporal annotations)



Dataset

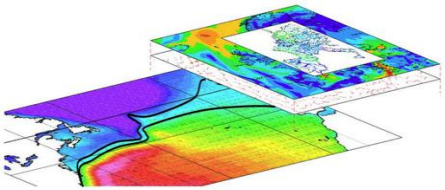
See.4C

European Commission

Alternative databases

Prediction of future frames in a video sequence, employed in: Mathieu, M.; Couprie, C.; LeCun, Y. *Deep multi-scale video prediction beyond mean square error.* ICLR, 2016.

Daset	Source	Description	Annotation
UFC101	Soomro, K.; Roshan, A.; Shah, M. *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild.* ArXiv, 2012	13320 videos (from youtube) from 101 action categories (different clip durations).	Action categories are grouped into 25 groups
Sports1m	Karpathy A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. *Large-scale Video Classification with Convolutional Neural Networks.* CVPR 2014.	1 million (1,133,158) YouTube videos belonging to 487 classes	Sport labels



Dataset

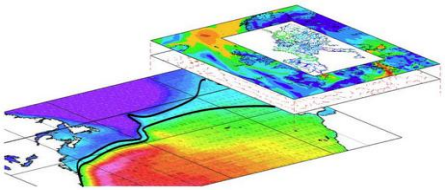
See.4C

European Commission

Alternative databases

Forecasting Actions and Objects, employed in: Vondrick, C.; Pirsiavash, H.; Torralba, A. *Anticipating Visual Representations from Unlabeled Video.* CVPR 2016.

Daset	Source	Description	Annotation
TV Human Interaction Dataset	Patron-Perez, A., Marszalek, M., Zisserman, A. and Reid, I. High Five: Recognising human interactions in TV shows. BMVC, 2010	300 video clips collected from over 20 different TV shows and containing 4 interactions: handshakes, high fives, hugs and kisses, and no interaction.	Upper body of people, head orientation and interaction label of each person.
ADL	Pirsiavash, H.; Ramanan, D. Detecting Activities of Daily Living in First-person Camera Views. CVPR, 2012	One million frames of dozens of people performing unscripted, everyday activities.	Activities, object tracks, hand positions, and interaction events



Dataset

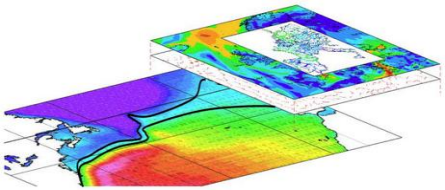
See.4C

European Commission

Alternative databases - **http://chalearnlap.cvc.uab.es/**

Apparent Personality Analysis, employed in: Ponce-López, V.; Chen, B.; Oliu, M.; Cornearu, C.; Clapés, A.; Guyon, I.; Baró, X.; Escalante, H.; Escalera, S. ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. ECCV 2016

Daset	Source	Description	Annotation
First impressions challenges	2016 Looking at People ECCV Challenge 2017 Looking at People CVPR Challenge	10,000 15-second videos collected from YouTube.	Personality traits Personality traits and job candidate screening label (audio, video, transcriptions available)



ICLR 2017

See.4C

European Commission

Thank you!