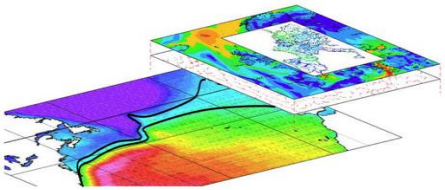


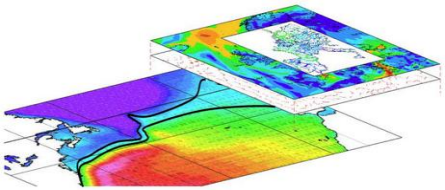
See.4C

European Commission

Faces short clips: data and deep spatio-temporal prediction

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

14/2/2017



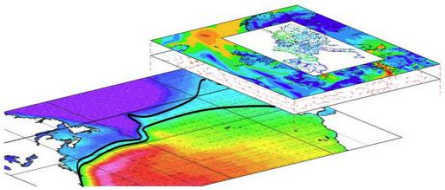
See.4C

European Commission

Index

Dataset

Deep spatio-temporal prediction approaches



videos

See.4C

European Commission

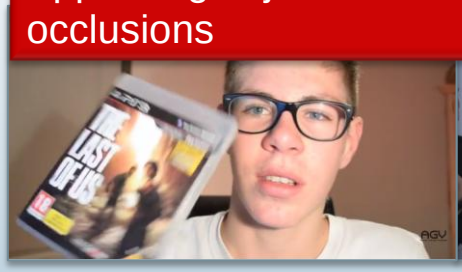
149 *talking-to-camera* videos from different sources

Quality: 720p HD @ 25 FPS

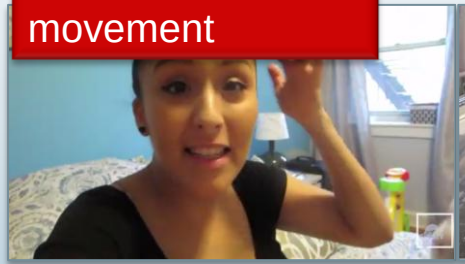
Total duration: 193,510 seconds (4,837,750 frames)

Illumination conditions


Appearing objects & occlusions



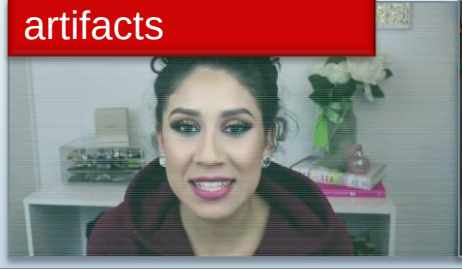
Camera movement

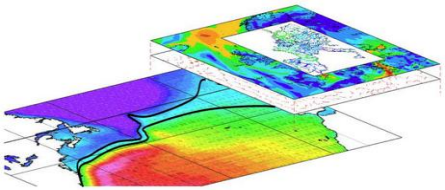



Ethnics & skin color


Upper and full body


Filters and artifacts



Dataset

See.4C

European Commission

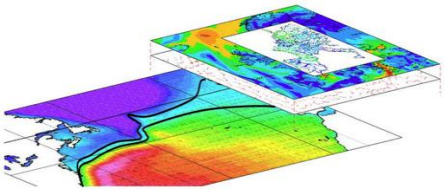
Target dataset:

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 on a large subset of videos in a few minutes.

Image samples of the proposed dataset.



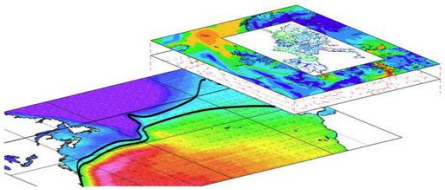
Dataset

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 selected faces nearest to centroid detected faces 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



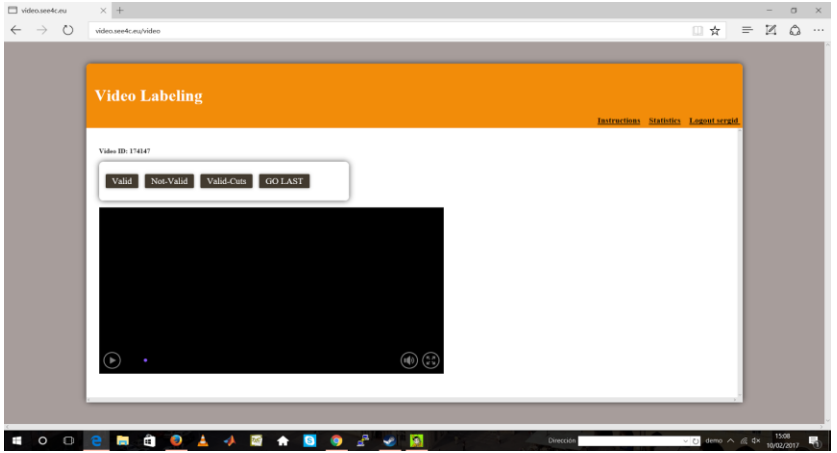
Dataset

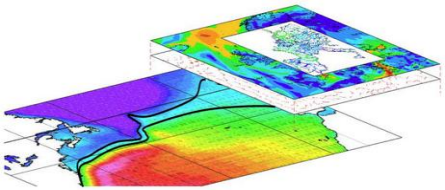
See.4C

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 around 1K clips from valid set for hackaton**





Dataset

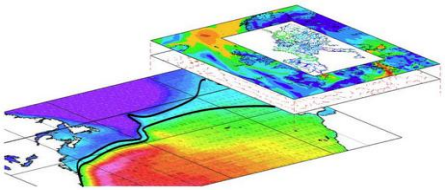
See.4C

European Commission

How sample videos look?

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





Dataset

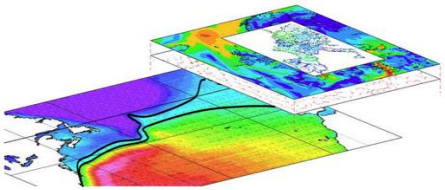
See.4C

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)

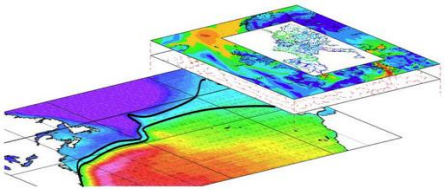


Dataset

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

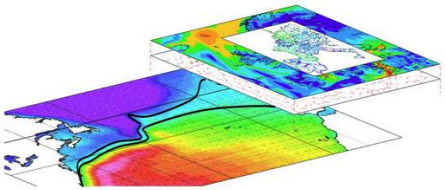
See.4C

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

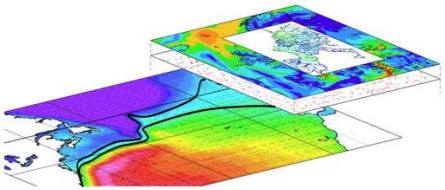
See.4C

European Commission

Alternative databases

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 challenge	2016 Looking at People ECCV Challenge	10,000 15-second videos collected from YouTube.	Personality traits



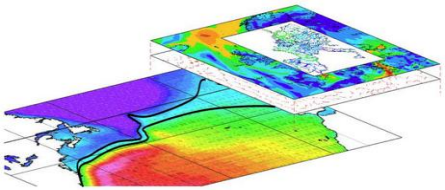
See.4C

European Commission

Index

Dataset

Deep spatio-temporal prediction approaches



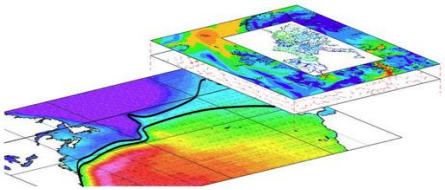
Deep approaches

See.4C

European Commission

- Deep models are widely considered for many CV (CNN) and ML tasks
- Deep temporal models, specially those based on CNN-RNN/LSTM are current state of the art in CV for behavior analysis in videos
- Deep learning generative approaches are also a current trend for spatio-temporal prediction of videos (future frames generation)

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative Adversarial Networks, arXiv:1406.2661



Dataset

Baseline method

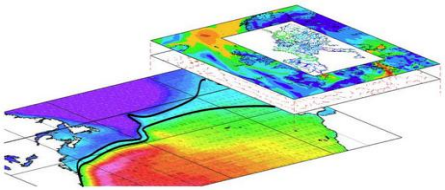
See.4C

European Commission

•**Prednet** - A predictive neural network architecture (**recurrent convolutional network**), inspired by the concept of "**predictive coding**" from neuroscience.

•**Predictive coding** posits that the brain is continually making predictions of incoming sensory stimuli.

Lotter, W., Kreiman, G., & Cox, D. (2016). **Deep predictive coding networks for video prediction and unsupervised learning**. *arXiv preprint arXiv:1605.08104*.

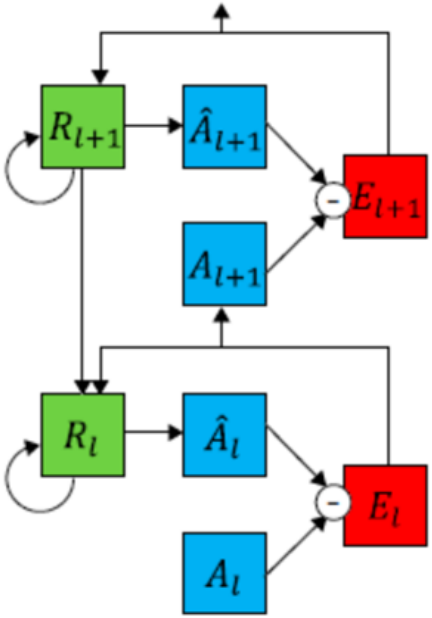


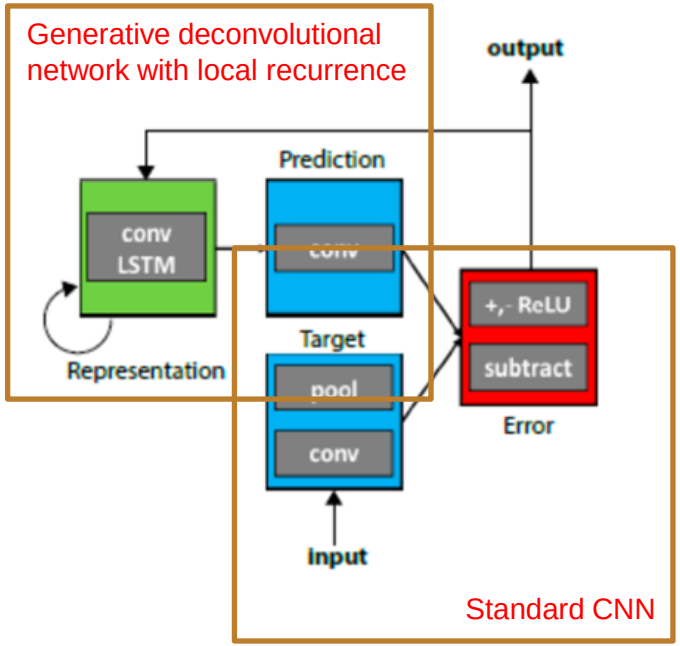
Dataset

See.4C

European Commission

Baseline method

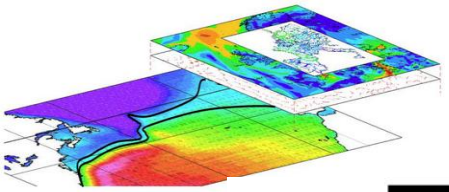




•These networks learn to predict future frames, with **each layer in the network making local predictions and forwarding deviations from those predictions to subsequent network layers.**

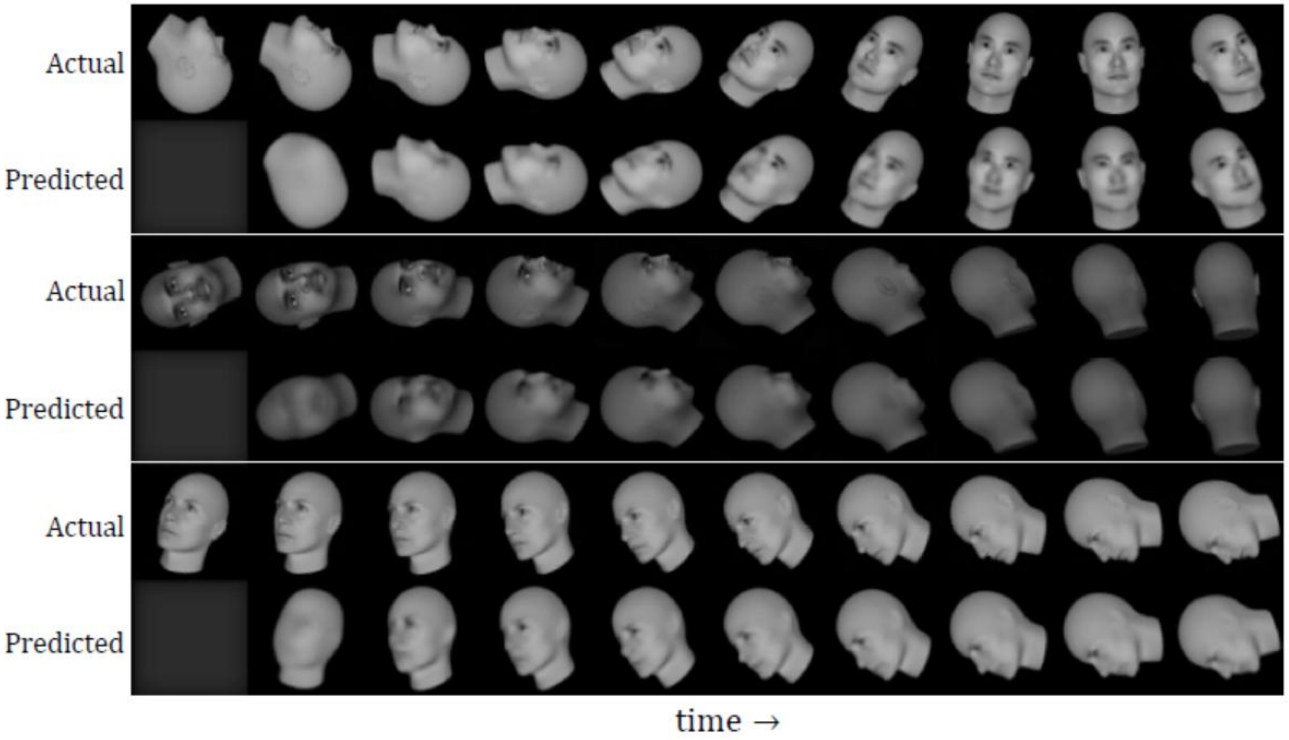
These operators define a functional block. **Many such blocks are concatenated to predict sequences frame by frame**.

Lotter, W., Kreiman, G., & Cox, D. (2016). **Deep predictive coding networks for video prediction and unsupervised learning**. *arXiv preprint arXiv:1605.08104*.

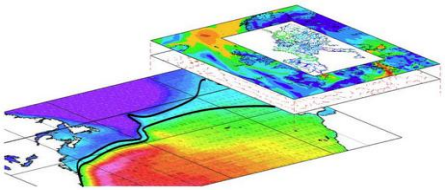


Dataset - Baseline method

See.4C

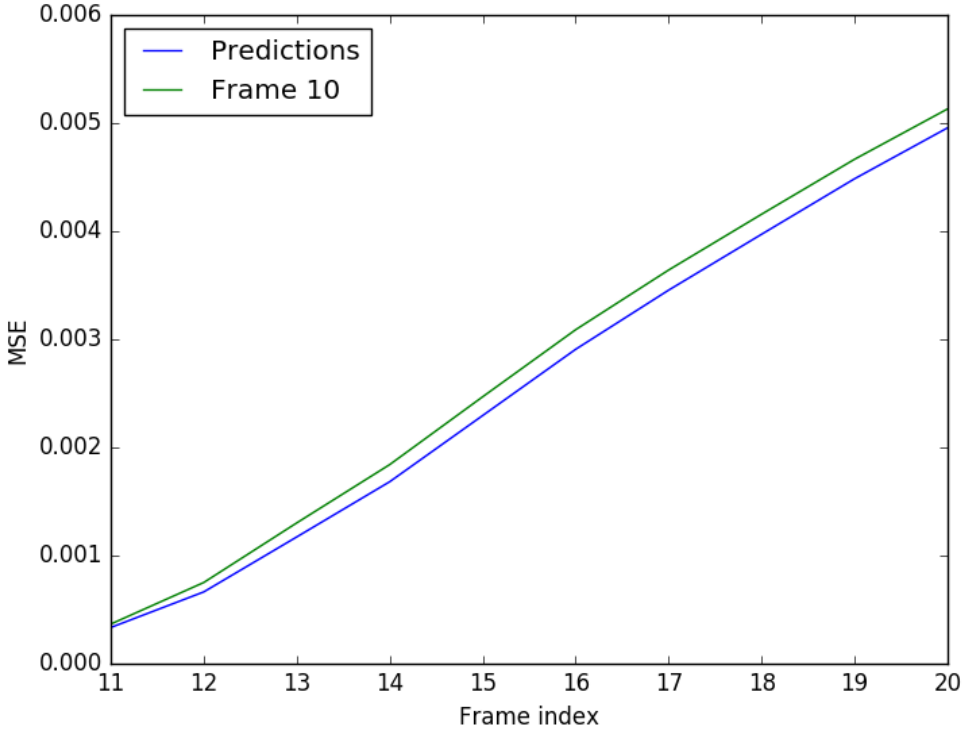


Lotter, W., Kreiman, G., & Cox, D. (2016). **Deep predictive coding networks for video prediction and unsupervised learning**. *arXiv preprint arXiv:1605.08104*.



Dataset - Baseline method
Results with our data

See.4C



14000 sequences for training

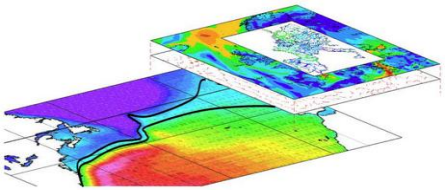
1366 sequences for validation

10 frames train – next 10 test

MSE results

Predictions – Deep baseline method

Frame 10 – Persistent (last frame replication)



Related works

See.4C

European Commission

•Nitish Srivastava, Elman Mansimov and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. ICML'2015.

•Michael Mathieu, Camille Couprie and Yann LeCun. Deep Multi-Scale Video Prediction Beyond Mean Square Error. ICLR'2016.

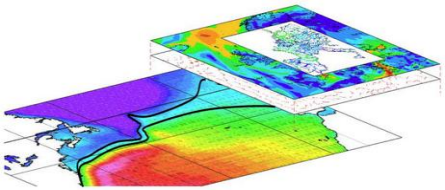
•Chelsea Finn, Ian Goodfellow and Sergey Levine. Unsupervised Learning for Physical Interaction through Video Prediction. NIPS'16.

•Carl Vondrick, Hamed Pirsiavash, Antonio Torralba. Generating Videos with Scene Dynamics. NIPS'16.

•Francesco Cricri, Xingyang Ni, Mikko Honkala, Emre Aksu, Moncef Gabbouj. Video Ladder Networks, arkiv.

•Nal Kalchbrenner, A ¨aron van den Oord, Karen Simonyan Ivo Danihelka, Oriol Vinyals, Alex Graves, Koray Kavukcuoglu (DeepMind). Video Pixel Networks. (probably) submitted to CVPR'17.

•Deep Predictive Coding Networks For Video Prediction And Unsupervised Learning. William Lotter, Gabriel Kreiman & David Cox. Submitted to ICLR'17.



See.4C

European Commission

Thank you!