Video Representation: More than Feature Mean

Xiang Xiang, Dept. of Computer Science, Johns Hopkins University

Monitoring pain through cameras.

Patients are typically not monitored in the waiting room.
- Deteriorations happen unnoticed.

e.g., In US 1.3 million deteriorations in waiting rooms per year.

Patients may move from place to place at any time, making real-time computing a necessity.

Patients unaware of being monitored don't adjust behavior.

Problem

Recognizing one source of variation among others in a video.
Temporal feature mean is widely-used yet not a robust statistic.

- How to represent pain visually?
- How to represent a patient visually?
- How to represent pose variation?

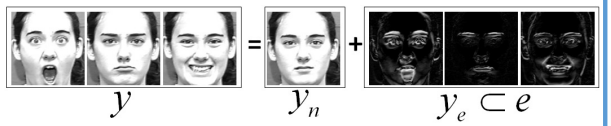
How to represent pain visually?

Pain can be expressed as a weighted combination of pains.
- Learn representation from examples. Same for other emotions.

Pain can be expressed as a sparse representation of emotions.

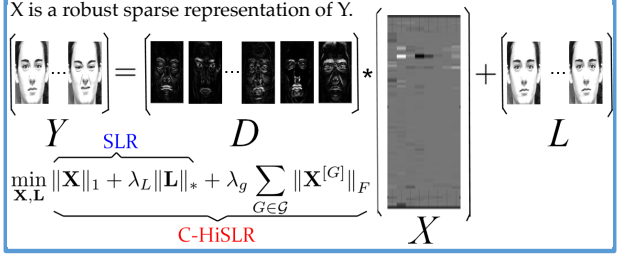
Will only the examples of pains be activated?
- No, the expressive face is not sparsely representable.

But we observe that there is always an underlying neural face.



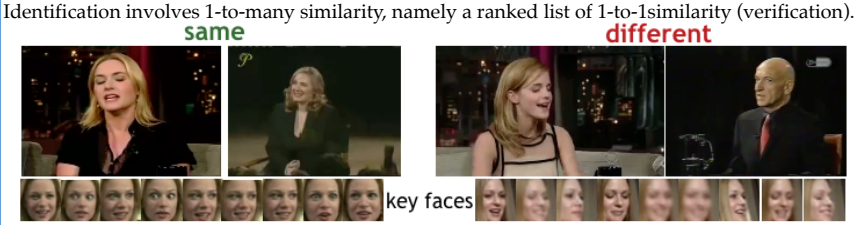
The residue after getting rid of neural face is sparsely representable.

Represent a video Y as a sparse matrix X robust to identity variation.
X is a robust sparse representation of Y.



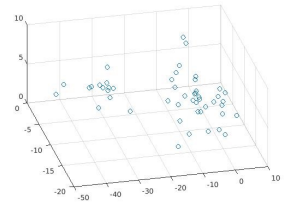
How to represent an identity visually?

Identification involves 1-to-many similarity, namely a ranked list of 1-to-1similarity (verification).



How to represent pose variation?

K-means clustering poses estimated as rotation angles. Selecting frames by distances to centroids.
Pros: reducing num of frames from tens or hundreds to K while preserving the overall diversity.

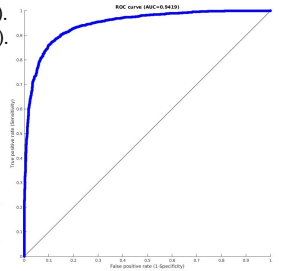


Experimental Results

Left: facial expression on Cohn Kanade plus dataset (321 videos).
Right: identification on YouTube Faces dataset (5000 video pairs).
AUC is 0.9419 close to VGG-Face using feature mean.

Model	An	Co	Di	Fe	Ha	Sa	Su
SRC	0.71	*0.60*	**0.93**	*0.25*	**0.96**	*0.24*	**0.98**
SLR	*0.51*	*0.63*	*0.74*	**0.51**	*0.85*	**0.70**	***0.94***
C-HiSLR	**0.77**	**0.84**	**0.93**	**0.53**	**0.93**	**0.65**	***0.95***

Table 4. Comparison of sensitivity. The **bold** and *italics* denote the highest and lowest respectively. Difference within 0.05 is treated as comparable. C-HiSLR performs the best.



Conclusion: possible to recognize pain and identity visually even in crowd waiting room.
Future work: handling facial expression, identity, pose in a holistic framework.
Take-home message: realistic data are noisy; robust analysis for sequential data is possible.

Reference: X. Xiang. Temporal Selective Max Pooling for Face Verification. FFER @ ICPR 2016; arxiv 1609.07042.
X. Xiang et al. Hierarchical Sparse and Collaborative Low-Rank Representation for Emotion Recognition. ICASSP 2015.
K Sikka et al. LOMo: Latent Ordinal Model for Facial Analysis in Videos. CVPR 2016. arXiv 1604.01500.

Contact: xxiang@cs.jhu.edu 410-446-4338
or via Greg Hager, Trac Tran, Alan Yuille.
Codes: https://github.com/eglxiang/