Facial Expression and Peripheral Physiology Fusion to Decode Individualized Affective Experience

Yu Yin, Mohsen Nabian, Miolin Fan, ChunAn Chou, Maria Gendron, *and* Sarah Ostadabbas*

Electrical and Computer Engineering, Mechanical and Industrial Engineering, and Psychology Departments

Northeastern University

Boston, USA

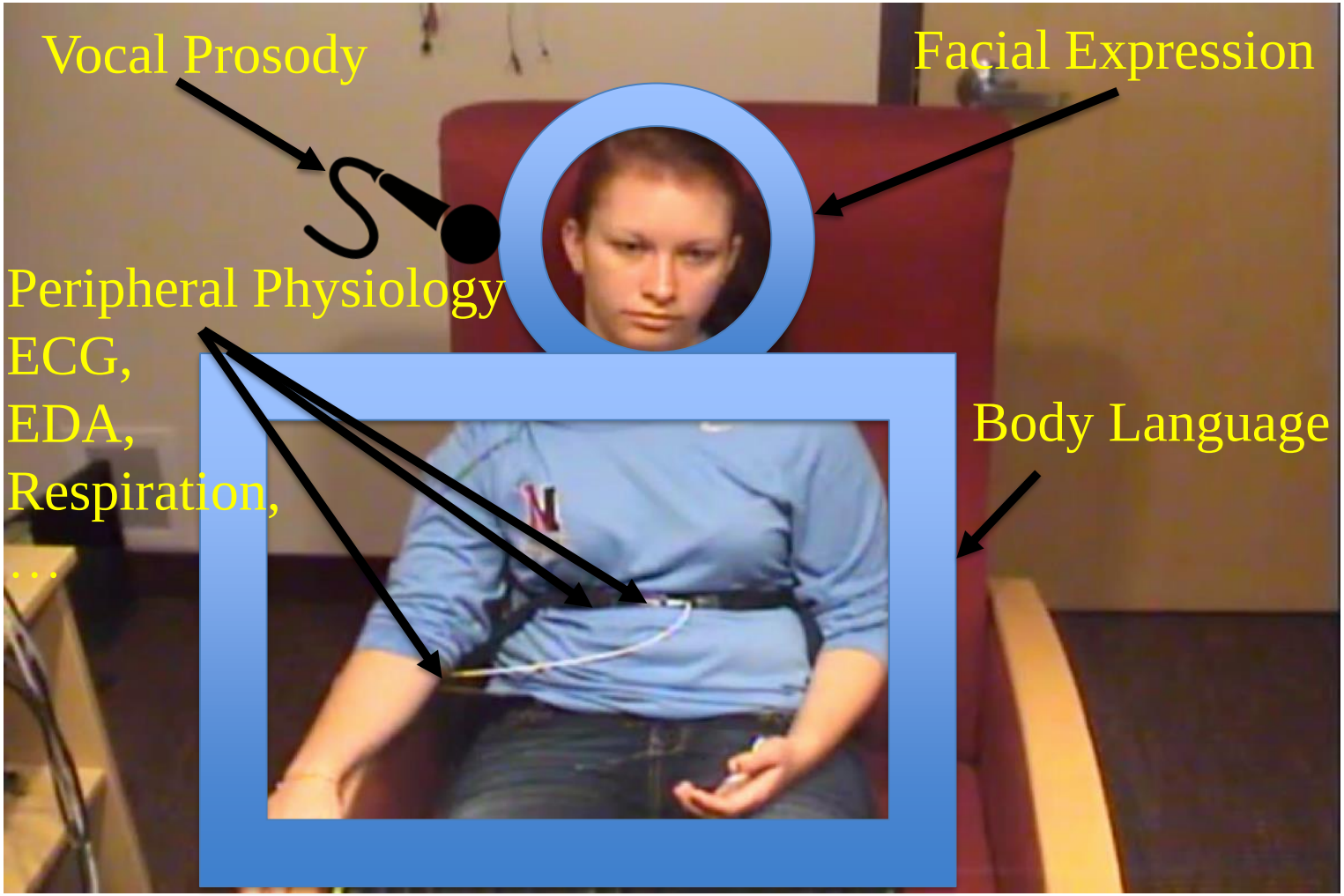
IJCAI AffyComp, July 15, 2018

Affective Experience

An Individualized Construct

Affective Experience

A Multimodal Construct



Our Goal

- Decoding individualized affective experience via analyzing multimodal external and internal sensory signals

Affective Experience

Existing Work in Emotion Recognition

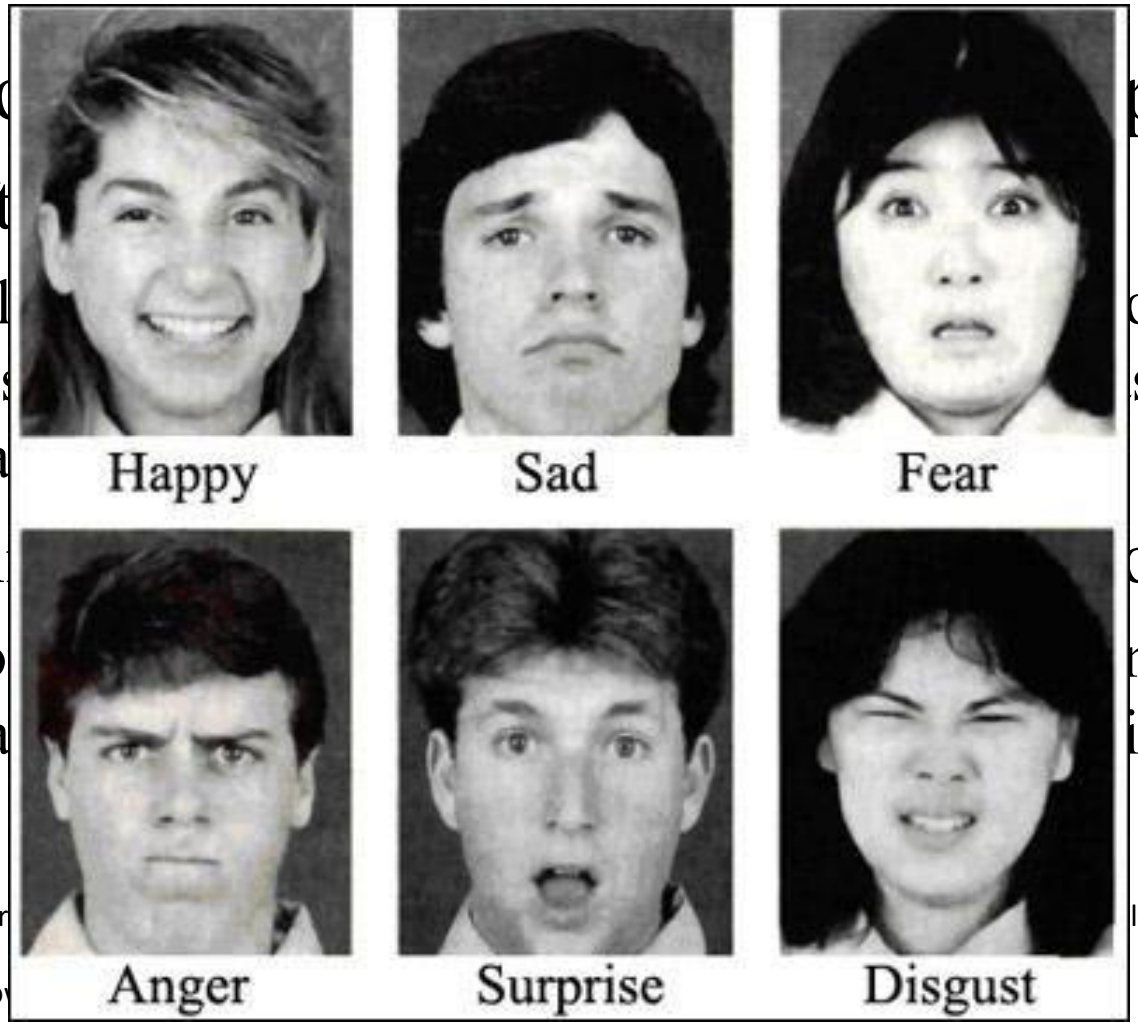
- Emotion recognition based on facial expression
  - It often takes a categorical approach
    - a label from a set of six purported basic emotions (anger, disgust, fear, happiness, sadness, surprise) is assigned to a pattern of facial movements [1]
  - Yet in real life, emotions are much more complex
    - some of the emotions do not even fit well in any of the basic categories; Theory of Constructed Emotion [2]

[1] James A Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102, 1994.
[2] Lisa Feldman Barrett. How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt, 2017.

Affective Experience

Existing Work in Emotion Recognition

- Emotic pression
  - It oft
    - a l ons (anger, di ssigned to a pa
  - Yet i omplex
    - so ny of the ba ion [2]



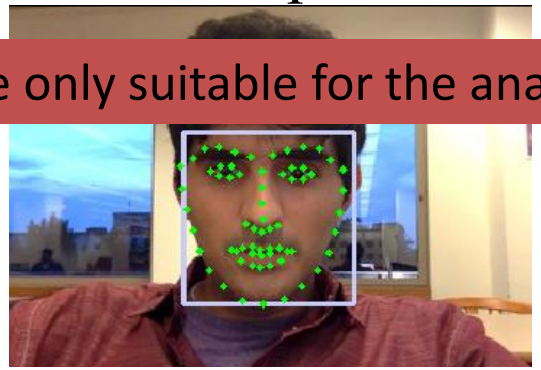
[1] James A Russell. Is ther ltural studies. *Psychological bulletin*, 115(1):102, 1994.
[2] Lisa Feldman Barrett. Ho

Affective Experience

Facial Expression Tracking

- Movements of face or facial expressions: a rich information source for affective display [1]
- Action Units (AUs):
  - Defined a the movements of specific facial muscle for finer-grained assessment of facial expressions

2D feature-based methods are only suitable for the analysis of frontal-view face [2]



[1] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. Facial Action Coding System, Investigator's Guide. 2002.
[2] Maja Pantic and Ioannis Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 36(2):433–449, 2006.

Affective Experience
Physiology Analysis

- However, not all emotions occur with an expression in face!
  - *correspondence between specific facial expressions and underlying experiences is not robust in psychology [1].*
- Complementary Information: Peripheral Physiology
  - electrocardiogram (ECG),
  - electroencephalogram (EEG),
  - electrodermal activity (EDA),
  - respiration patterns,
  - …

[1] Rainer Reisenzein, Markus Studtmann, and Gernot Horstmann. Coherence between emotion and facial expression: Evidence from laboratory experiments. Emotion Review, 5(1):16–23, 2013.

Affective Experience
Physiology Analysis

- However, not all emotions occur with an expression in face!
  - *correspondence between specific facial expressions and underlying experiences is not robust in psychology [1].*
- Complementary Information: Peripheral Physiology
  - electrocardiogram (ECG),

The individualized physiological patterns during rest is only measured using a mean value!

  - respiration patterns,
  - …

[1] Rainer Reisenzein, Markus Studtmann, and Gernot Horstmann. Coherence between emotion and facial expression: Evidence from laboratory experiments. Emotion Review, 5(1):16–23, 2013.

Affective Experience

Fusing Different Sources of Information

- Fusion to improve performance of emotion recognition algorithms
- Fusion Levels:
  - Feature-level fusion
    - stacks all the feature vectors together [1]
  - Decision-level fusion
    - first classifies each modality individually and then combines the classifier outputs [2]

[1] Sander Koelstra and Ioannis Patras. Fusion of facial expressions and EEG for implicit affective tagging. Image and Vision Computing, 31(2):164–174, 2013.
[2] Wenhui Liao, Weihong Zhang, Zhiwei Zhu, Qiang Ji, and Wayne D. Gray. Toward a decision-theoretic framework for affect recognition and user assistance. International Journal of Human Computer Studies, 64(9):847–873, 2006.

Affective Experience

Fusing Different Sources of Information

- Fusion to improve performance of emotion recognition algorithms
- Fusion Levels:

The fusion of multimodal person-specific data in levels prior to the emotion experience inference is largely unexplored.

  - first classifies each modality individually and then combines the classifier outputs [2]

[1] Sander Koelstra and Ioannis Patras. Fusion of facial expressions and EEG for implicit affective tagging. Image and Vision Computing, 31(2):164–174, 2013.
[2] Wenhui Liao, Weihong Zhang, Zhiwei Zhu, Qiang Ji, and Wayne D. Gray. Toward a decision-theoretic framework for affect recognition and user assistance. International Journal of Human Computer Studies, 64(9):847–873, 2006.

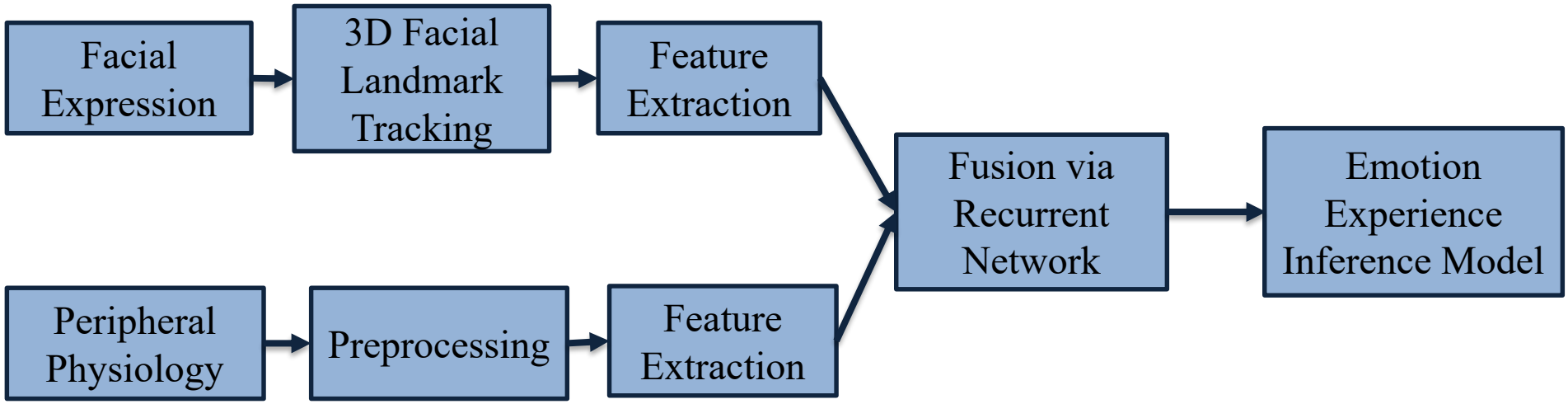
Affective Experience
Our Contributions

To address these challenges, in this work, we:

1. Employ a 3D model for facial landmark localization/tracking, which decouples head motion from face expression;
2. Assess the resting/affective multimodal response of each individual through a higher order dynamics using recurrence network metrics;
3. Develop a novel multimodal feature fusion approach based on recurrence network for affective experience decoding.

Multimodal Data Fusion

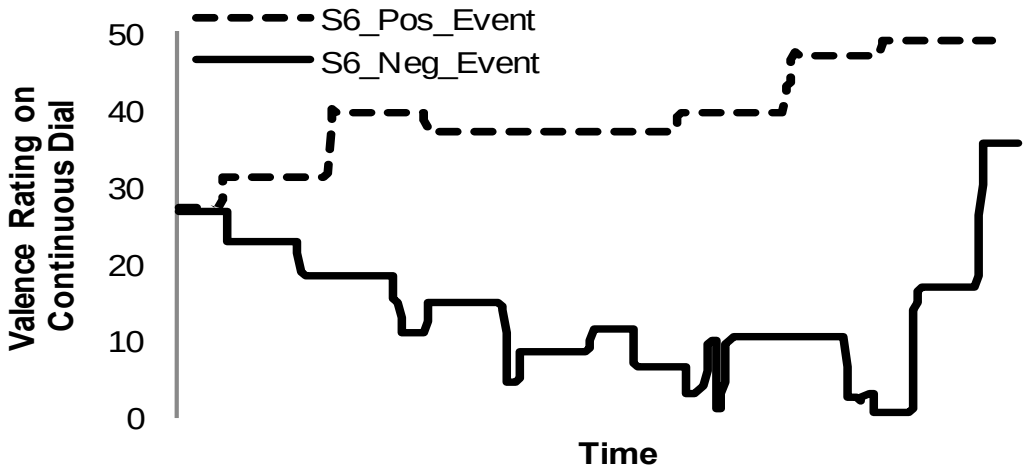
Our Proposed Framework



Dataset Forming

- Phase I: 12 participants described their two most positive and their two most negative emotional experiences.

- Phase II: each participant watched their own 4 recorded videos as stimuli.



Continuous rating of an example video stimulus using rating dial (ranging from negative = 0 to positive = 50) reveals dynamics in videos across time.

Methods

(1) Decoupling of the head from 3D facial landmark movements
(2) Multimodal data fusion
(3) Affective experience decoding

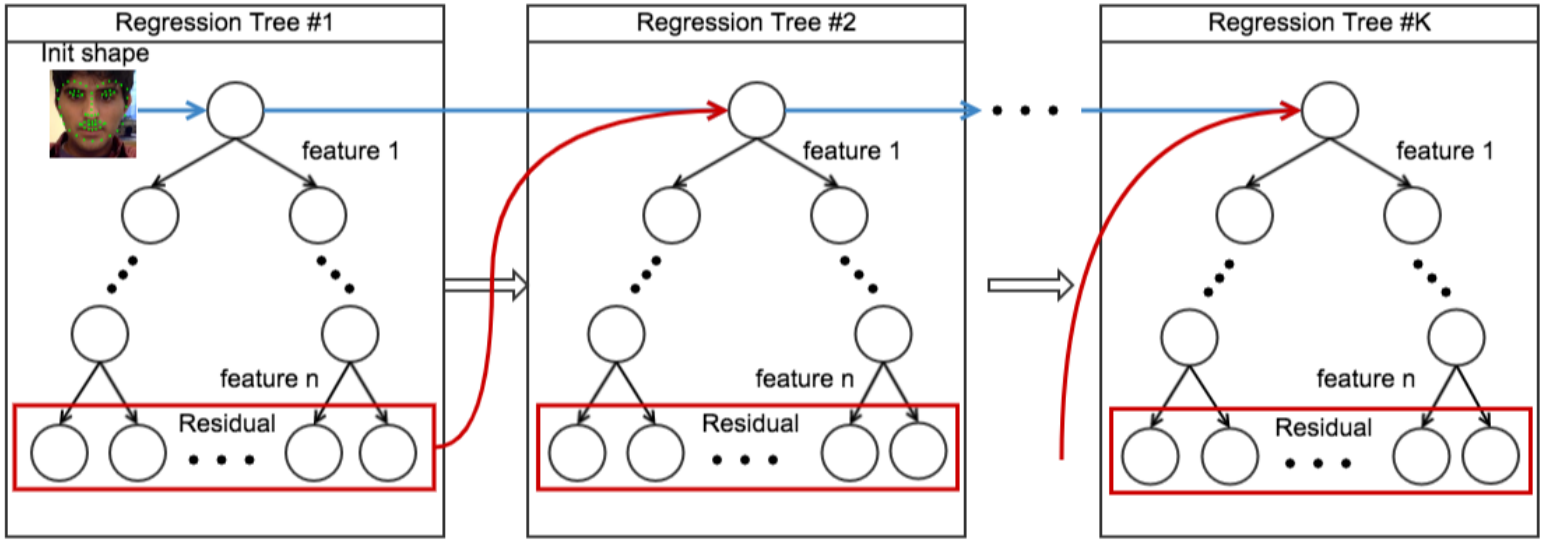
Facial Landmarks
Localization and Tracking

- A two-step process:
  1. Using a state-of-the-art 2D facial alignment algorithm ([1]) to automatically localize 68 landmarks for each frame of the face video
  2. Then, a 3D face model is used ([2]) to estimate the depth information from the 2D frames and thus to achieve 3D landmark tracking

[1] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1867–1874, 2014.
[2] Josef Kittler, Patrik Huber, Zhen-Hua Feng, Guosheng Hu, and William Christmas. 3d morphable face models and their applications. International Conference on Articulated Motion and Deformable Objects, pages 185–206, 2016.

2D Landmark Localization

- A cascade of trained regressors is utilized to localize the facial landmarks for each video frame
  - In each level of cascade, estimated landmarks are refined by adding residuals produced by the previous regression

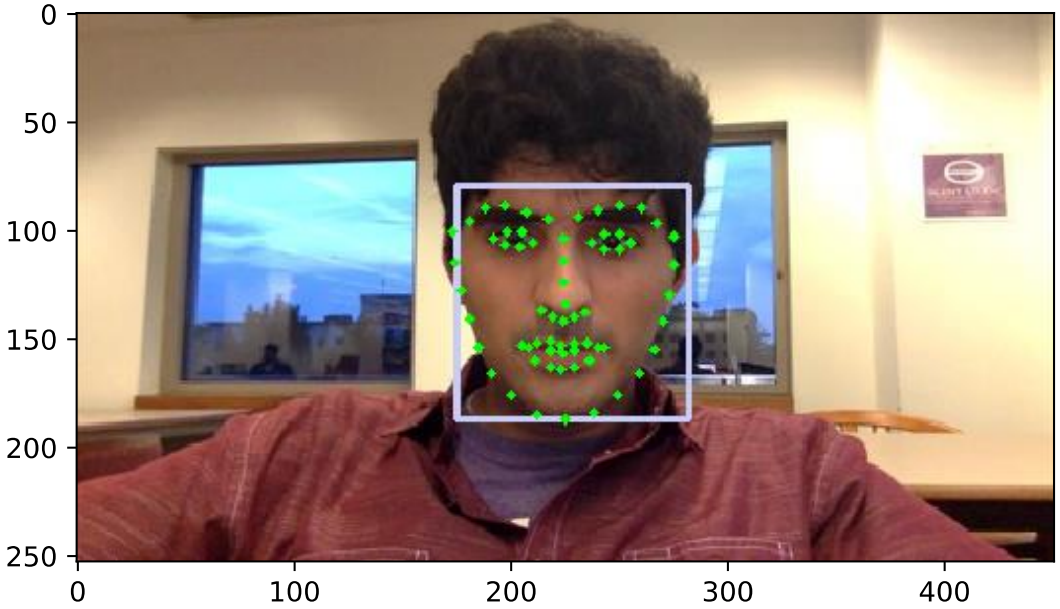


2D Landmark Localization Algorithm

- Assume we have training dataset $\{(I_1, S_1), \ldots, (I_n, S_n)\}$, where each I_i is a face image and S_i is its shape vector.
  - We set an initial shape estimate $\hat{S}_i^{(0)}$ for every face image
  - In each regression tree, the regression function r_t is learned using the gradient tree boosting algorithm, and then the estimation of every shape is updated as:

$$\hat{S}_i^{(t+1)} = \hat{S}_i^{(t)} + r_t(I_i, \hat{S}_i^{(t)})$$

2D Landmark Localization Algorithm



3D Landmark Tracking

- Our facial landmarks tracking algorithm needs to remain invariant across head movement.
- To eliminate the interference of head movement, we extract the depth information of each face pixels from 2D video frames using a 3D morphable face model [1]

[1] Josef Kittler, Patrik Huber, Zhen-Hua Feng, Guosheng Hu, and William Christmas. 3d morphable face models and their applications. International Conference on Articulated Motion and Deformable Objects, pages 185–206, 2016.

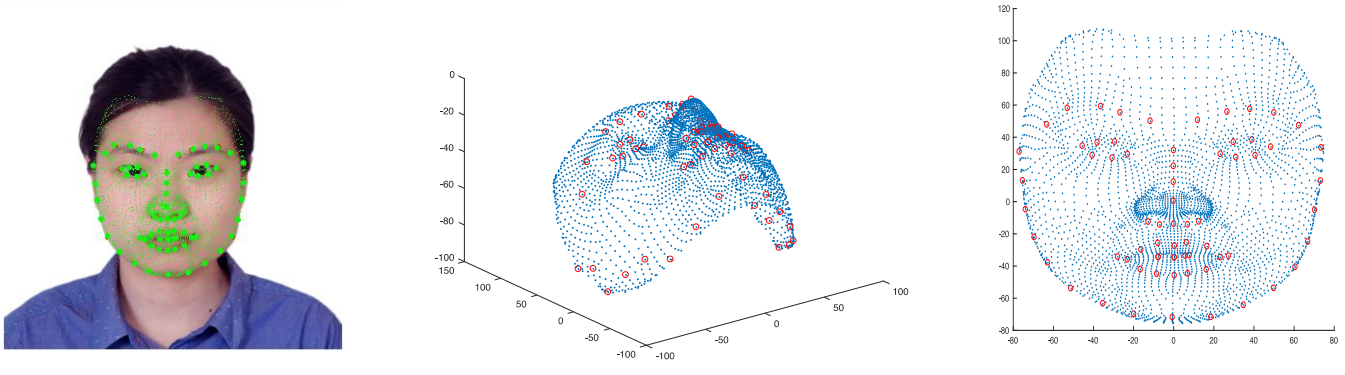
3D Landmark Tracking Algorithm

- This model consists of a PCA model of face shapes,
  - for reconstructing a 3D face from a single 2D image.
  - The PCA model consists of a set of principal components $V = [v_1, \ldots, v_k]$, the mean value of all the facial meshes $\bar{v}$, and their standard deviation σ_k.
- The shape of a novel face is then generated with:

$$S_k = \bar{v} + \sum_{k=1}^{K} \alpha_k \, \sigma_k \, v_k$$

where K is the number of principal components and α_k's are the representation of S_k in the coordinates of the PCA shape space.

3D Landmark Fitting

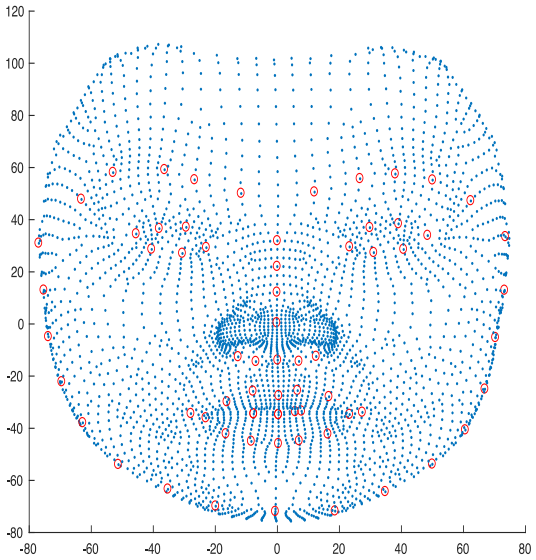


An example result of the landmark fitting: resulting shape and model fitting (left), 3D model fitting result (middle), and frontal view of the 3D facial model (right).

Facial Landmark Features

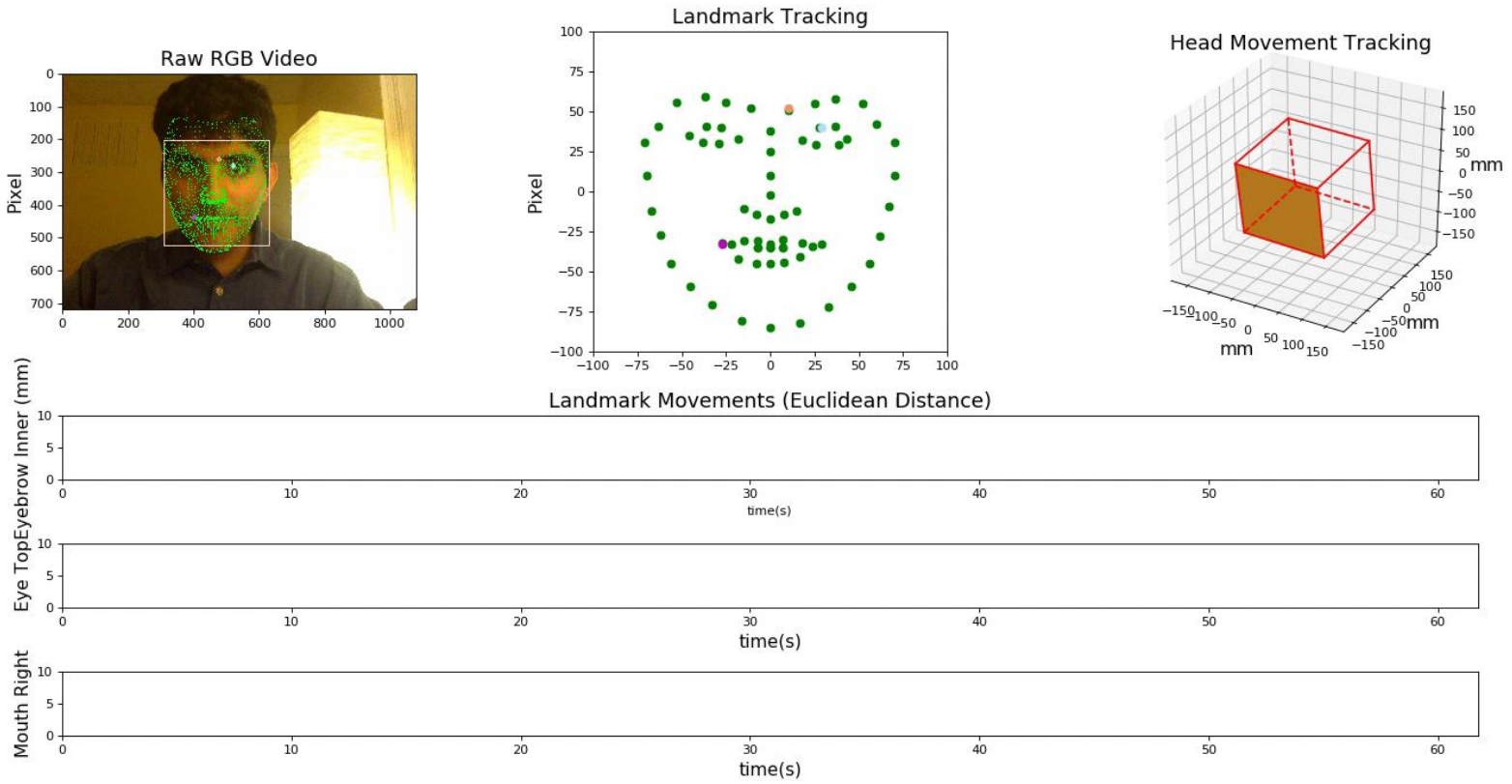
- Guided by the work in [1], we reduced the facial landmarks feature dimensions from 2×68 to 12 features.

Feature No.	Explanation
1, 2	left and right eyebrow y-values
3	inner corners differences of eyebrows
4	horizontal distance of the the two corners of lips
5	vertical distance of the two lips
6	average vertical positions of the two corners of the lips
7, 8, 9	head rigid displacement in X, Y, and Z direction
10, 11, 12	head rigid rotation in roll, pitch, and yaw direction



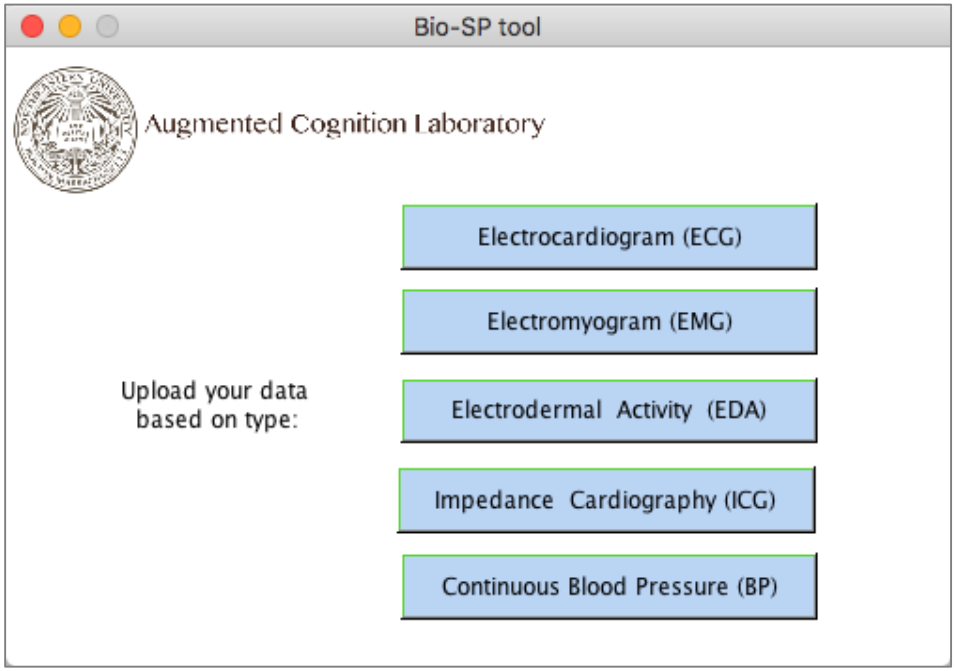
[1] Y Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. IEEE Transactions on pattern analysis and machine intelligence, 23(2):97–115, 2001.

3D Landmark Tracking



Physiological Signal Processing

- The preprocessing and feature extraction of physiological signals including ECG, EDA, and respiration were done with our Bio-SP tool.



Biosignal-Specific Processing (Bio-SP) Tool. http://www.northeastern.edu/ostadabbas/ software/.

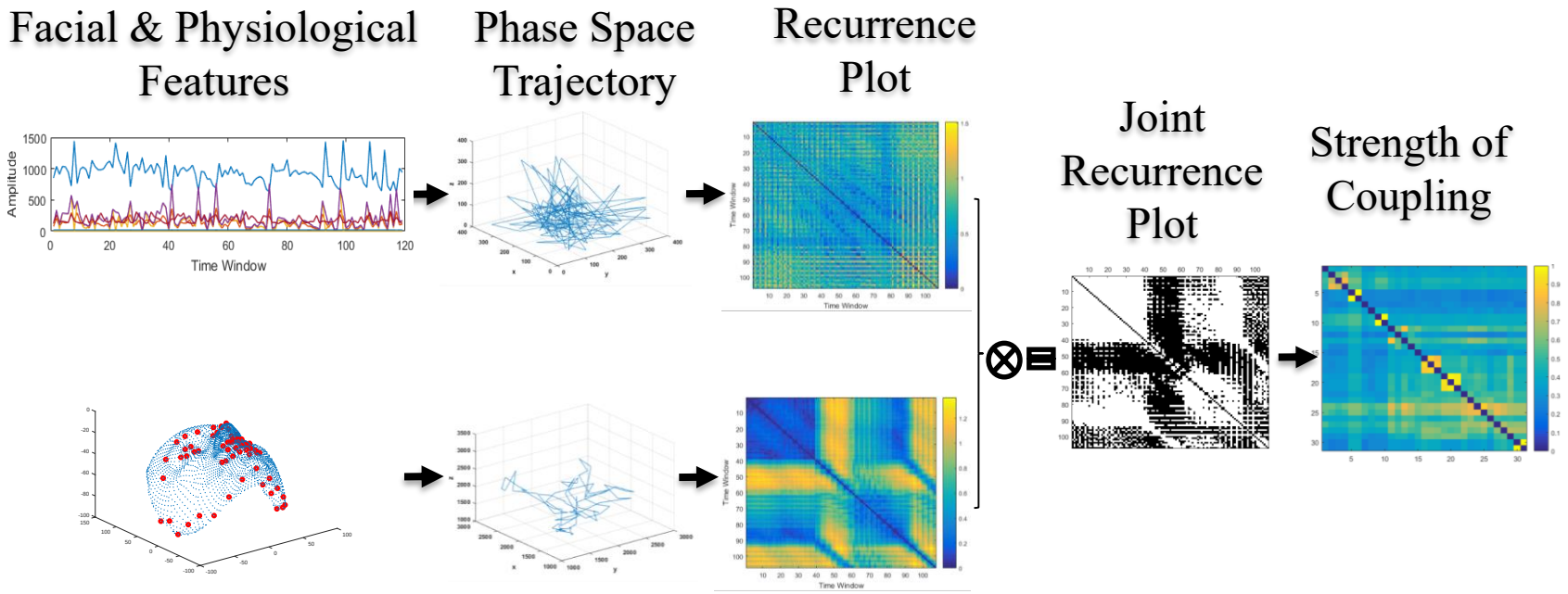
Methods

(1) Decoupling of the head from 3D facial landmark movements
(2) Multimodal data fusion using recurrent networks
(3) Affective experience decoding

Modality Alignment
in Heterogeneous Data Streams

Modality	Window Size (sec)	Overlap (%)	Features (#)
Face	5	50	6
Head	5	50	6
ECG	5	50	10
EDA	20	88.2	5
Resp	30	92.4	4

Multimodal Data Fusion Framework



Recurrence Plot (RP)
Single Time-Series

- RP is a visualization to represent the temporal dependency relationships between all states in a time series data using a binary, squared matrix

$$\mathbf{R}_{i,j} = \Theta\left(\varepsilon_i - \|\mathbf{x}_i - \mathbf{x}_j\|\right), \quad \mathbf{x}_i \in \Re^m, \quad i,j=1\ldots N,$$

Threshold

Heaviside function

System's states at time *i* and *j*, based on some pre-defined time lag

Binary value

Joint RP
Extension to Multimodal

- To best capture the dynamic coupling between multiple modalities, we adopted the joint RP (JRP) since it represents when a recurrence occurs simultaneously in two or more time series:

$$\mathbf{JR}_{i,j}^{X,Y} = \Theta(\epsilon_X - ||\mathbf{x}_i - \mathbf{x}_j||_1)\Theta(\epsilon_Y - ||\mathbf{y}_i - \mathbf{y}_j||_1)$$

Network-Based Feature Extraction

- The JRP matrix is then converted to the adjacency matrix for a network called recurrence networks (RN)

$$\mathbf{A}_{i,j}(\varepsilon) = \mathbf{JR}_{i,j}(\varepsilon) - \mathbf{I}$$

- The network measures we computed include two general classes:
  - (1) global measures: transitivity, global efficiency, and out-strength/in-strength correlation.
  - (2) local measures: in-strength/out-strength, local efficiency, edge/node betweenness centrality, diversity, and clustering coefficients.

Methods

(1) Decoupling of the head from 3D facial landmark movements
(2) Multimodal data fusion using recurrent networks
(3) Affective experience decoding and quantitative results

Inference Models

- We performed two decoding tasks:
  1. A binary across-individual classification on video stimuli
     - The stimuli context prediction accuracy was defined as the percentage of stimuli correctly classified based on their positive or negative contents.
  2. An individualized affective experience prediction of participants while watching a video stimulus
     - We employed support vector regressor (SVR) with a ridge penalty as the self-rating regression model.

Results I

Emotional video stimulus content prediction results

Modality	Accuracy	p-value	F1	Precision	Recall
ECG	51.4	0.29	51.3	51.3	52.1
EDA	48.6	0.18	47.7	48.6	47.2
Resp	49.3	0.20	43.4	49.1	40.3
Face	45.1	0.10	44.6	44.9	44.4
Head	50.7	0.27	52.4	50.6	**55.6**
Fusion	**55.3**	0.49	**54.1**	**56.7**	51.8
Random	50.4	0.24	50	50.6	48.5

Results I

Emotional video stimuli content prediction results

Modality	Accuracy	Kappa value	F1	Precision	Recall
ECG	51.4	0.29			52.1
EDA	48.6				47.2
R					0.3
					44.4
				50.6	**55.6**
		0.49	**54.1**	**56.7**	51.8
	50.4	0.2	50	50.4	48.5

Results II

Affective self-rating scale prediction results

Modality	RMSE	MAE
ECG	0.33	0.28
EDA	0.27	0.24
Resp	0.27	0.24
Face	0.32	0.27
Head	0.45	0.33
Facial	0.27	**0.24**
Physio	0.29	0.25
Fusion	**0.26**	**0.24**

Sarah Ostadabbas, PhD
Assistant Professor
Electrical & Computer Engineering Department
Northeastern University
Boston, MA 02115
ostadabbas@ece.neu.edu
http://www.northeastern.edu/ostadabbas/

National Science Foundation
WHERE DISCOVERIES BEGIN

Northeastern University

MathWorks® amazon web services™ NVIDIA