Proceedings of Machine Learning Research 86:10–25, 2018 2nd IJCAI Workshop on Artificial Intelligence in Affective Computing

Facial Expression and Peripheral Physiology Fusion to Decode Individualized Affective Experience

Yu Yin, Mohsen Nabian, Sarah Ostadabbas* YIN.YU1@HUSKY.NEU.EDU, MONABIYAN@ECE.NEU.EDU, OSTADABBAS@ECE.NEU.EDU
Miolin Fan, ChunAn Chou FAN.MI@HUSKY.NEU.EDU, CH.CHOU@NORTHEASTERN.EDU
Maria Gendron MARIA.GENDRON@GMAIL.COM
Northeastern University, Boston, USA

Abstract

Affective experience prediction using different data modalities measured from an individual such as their facial expression or physiological signals has received substantial research attention in recent years. However, most studies ignore the fact that people besides having different responses under affective stimuli, may also have different resting dynamics (embedded in both facial and physiological patterns) to begin with. In this paper, we present a multimodal approach to simultaneously analyze facial movements and several peripheral physiological signals to decode individualized affective experiences under positive and negative emotional contexts, while considering their personalized resting dynamics. We propose a person-specific recurrence network to quantify the dynamics present in the person's facial movements and physiological data. Facial movement is represented using a robust head vs. 3D face landmark localization and tracking approach, and physiological data are processed by extracting known attributes related to the underlying affective experience. The dynamical coupling between different input modalities is then assessed through the extraction of several complex recurrent network metrics. Inference models are then trained using these metrics as features to predict individual's affective experience in a given context, after their resting dynamics are excluded from their response. We validated our approach using a multimodal dataset consists of (i) facial videos and (ii) several peripheral physiological signals, synchronously recorded from 12 participants while watching 4 emotion-eliciting video-based stimuli. The affective experience prediction results signified that our multimodal fusion method improves the prediction accuracy up to 19% when compared to the prediction using only one or a subset of the input modalities. Furthermore, we gained prediction improvement for affective experience by considering the effect of individualized resting dynamics.

Keywords: Multimodal data fusion, individualized affective experience, facial expression, physiological signals.

1. Introduction

Affective experience is an important construct in explaining several critical aspects of human behaviors and is impaired or irregular in a number of neurodevelopmental and psychiatric disorders (Panksepp, 2004). Experimental evidence indicates that positive or negative affective experience plays an important role in motivating future actions (Barrett and Bliss-Moreau, 2009) and can promote behavioral patterns linked to compromised mental health (Wichers et al., 2015). In addition,

* This paper has code available at GitHub under 3D Facial Landmark Detection and Tracking, provided by the corresponding author's lab.

information about affective states of users has become more and more important in human-computer interaction and many other emerging areas (Schaaff and Schultz, 2009) in recent years as it greatly facilitates the ability of computers to heed the rules of human communication (Picard, 2000).

A common approach to quantify the range of human affective states and to predict human affective experiences under different circumstances is based on decoding their facial movements. Movements of the face are considered as a particularly rich source for affective display and are commonly referred to as "facial expressions" (Littlejohn and Foss, 2010; Ekman et al., 2002; Nielson et al., 2018; DURÁN et al., 2017; Russell and Fernández-Dols, 2017; Fernández-Dols and Crivelli, 2013). However, not all emotions occur with an expression in face or even distinguishable facial expression (Ekman, 1993) and the correspondence between specific facial expressions and underlying emotional experiences is not robust in psychology (Reisenzein et al., 2013). To address this problem, various physiological signals including electrocardiogram (ECG), electroencephalogram (EEG), electrodermal activity (EDA), blood pressure, and respiration patterns have been used as complementary information to decode affective states (Verma and Tiwary, 2014; Khalaf et al., 2017) based on the function of multiple physiological systems in the body (Liao et al., 2006; Perez-Rosero et al., 2017).

Among humans, there are considerable individualized differences in facial expression and peripheral physiological responses under similar affective experiences, which influence the recognition results of generalized predictive algorithms. Therefore, There is a large body of literature investigating personalized models and their impact on accurately predicting person-specific affect (Yang et al., 2014; Hernandez et al., 2011; Kandemir et al., 2014). However, it is crucial to note that human facial movements or physiological responses even without any emotional stimulus (i.e. under resting state) differ substantially from one individual to another, due to multiple intrinsic and extrinsic factors (e.g. gender, personality, cultural influences, level of education, etc.) (Hurlburt et al., 2015; Krys et al., 2016; Joo et al., 2012). In the majority of the affective computing or even psychological studies, individual baseline differences are only accounted by subtracting the mean value (and very rarely considering the standard deviation) of the collected data during resting state from the stimulus-responding state data, assuming that the resting dynamics can simply be modeled solely with the zeroth and first moments.

In this paper, we applied recurrence network analysis for multimodal data (i.e. facial movements and physiological signals) fusion to identify and decode individualized affective experiences. To extract features from facial movements, we developed a robust landmark tracking approach, in which head movement is also independently tracked and decoupled from the facial landmark movements. As for the physiological signals including ECG, EDA, and respiration, a series of signal-specific algorithms developed in (Nabian et al., 2017, 2018) were utilized to extract psychologically-related features. We used complex network metrics to assess the inter-system dynamical coupling between different response modalities of a person under a negative or positive affective experience. We used these metrics to build an inference model for individual affective experience prediction. Critically, we also modeled the resting dynamics of each individual participant before undergoing any emotional induction, to account for individualized baseline differences in affective experience. Our main contributions in this paper are as follows: (1) employing a 3D model for facial landmark localization/tracking, which decouples head motion from face expression; (2) assessing the resting/affective multimodal response of each individual through a higher order dynamics using recurrence network metrics; and (3) developing a novel multimodal feature fusion approach based on recurrence network for affective experience decoding.

2. Related Work

Emotion recognition studies based on facial expression analysis often take a categorical approach, where a label from a set of six purported basic emotions (anger, disgust, fear, happiness, sadness, surprise) is assigned to a pattern of facial movements (Russell, 1994). Yet in real life, emotions are much more complex (Barrett et al., 2016), and specificity and consistency of facial movements to emotions is often lacking (Fernández-Dols and Crivelli, 2013; Reisenzein et al., 2013). Moreover, some of the emotions do not even fit well in any of the basic categories (Koelstra and Patras, 2013). A finer-grained assessment of facial expressions is to directly detect specific facial muscle actions (action units; AUs), including but not limited to the facial movements on which the basic emotion expressions were based (Tian et al., 2001; Cohn, 2007; Sánchez-Lozano et al., 2016). The facial features used in AUs recognition studies are often either geometric features indicating the location of facial characteristic points (mouth, eyes, chin, etc.) or appearance features representing the facial textures (Zeng et al., 2009; Valstar et al., 2015). In (Pantic and Patras, 2006), a set of facial points was used as geometric feature to recognize AUs in frontal-view face images. In (Bartlett et al., 2006; Guo and Dyer, 2005), appearance-based methods, such as Gabor wavelets or eigenfaces were applied to classify facial expressions or AUs. Some other work used both geometric and appearance features. In (Ringeval et al., 2015), appearance features were extracted using local Gabor binary patterns and geometric features were extracted based on 49 facial landmarks. Similarly, (Benitez-Quiroz et al., 2016) used second-order statistics of facial landmarks (i.e., distances and angles between landmark points) for geometric features and Gabor filters for appearance features.

However, most 2D feature-based methods are only suitable for the analysis of frontal-view face, which means only a small range of head movement is allowed. Fewer works of facial expression analysis have been done based on 3D face models. In (Zeng et al., 2006), a 3D face tracker was used to handle the arbitrary behavior of the person in the natural setting. Both geometry and appearance features were extracted based on the 3D face model. In (Cohn et al., 2004), authors focused on recognizing two of the most important facial actions in brow (brow raising and brow lowering) measured in spontaneous facial behavior with non-frontal pose, moderate out-of-plane head motion, and occlusion. A cylindrical head model was applied to estimate head movements in (Xiao et al., 2002).

To date, research on fusion of facial expression and physiological data in order to improve performance of emotion recognition algorithms is continuing to attract the attention of academia and industry alike. In (Koelstra and Patras, 2013), multimodal approaches based on both feature-level and decision-level fusion were applied to analyze facial expressions and EEG signals for generation of affective tags. In feature-level fusion, the authors simply stacked all the feature vectors together. In decision-level fusion, they first classified each modality individually and then combined the classifier outputs in a linear fashion. In (Liao et al., 2006), authors focused on recognizing only two different affective states, stress and fatigue. They applied a decision-level fusion approach to recognize these two affective states from multiple modalities including physical appearance (e.g. facial expression, head movement), physiological measures (e.g. EEG, ECG), behavioral data (e.g. mouse movement, type speed), and user performance (e.g. response time). In (Fan and Chou, 2018), authors proposed to apply recurrence network analysis to quantify the dynamics and to extract nonlinear features from EEG signals for classifying affective states. Based on existing knowledge and methods, the fusion of multimodal person-specific data in levels prior to the emotion experience inference is largely unexplored (Zeng et al., 2006). It is worth mentioning that data fusion in earlier

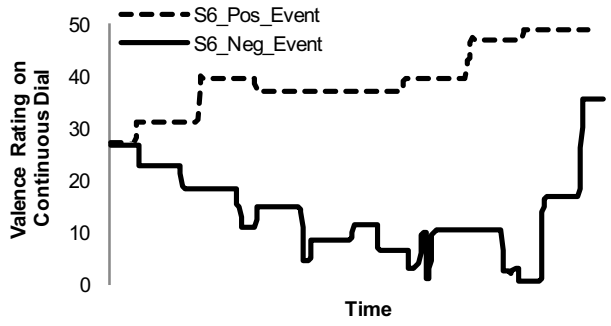

Figure 1: Continuous rating of an example video stimulus using rating dial (ranging from negative = 0 to positive = 50) reveals dynamics in videos across time.

stages, before decision, can lead to capturing the higher order information presents in the multimodal data, as well as the dynamical coupling between different input modalities.

3. Materials & Methods

3.1. Dataset

In this work, we used a multimodal dataset collected by the Psychology Department of Northeastern University, which contains both facial video recording and synchronous physiological data including ECG, EDA, and respiration signals. These data were obtained from 12 consenting participants during two data collection phases: (Phase I) each participant described their two most positive and their two most negative emotional experiences, and (Phase II) each participant watched their own 4 recorded videos as stimuli. In both phased, facial videos were recorded by a frontal camera at 25 frames per second. The three physiological signals were sampled at 1000Hz using BioLab v.3.0.13 (Mindware Technologies; Gahanna, OH) via a BioNex 8-Slot chassis (Model50-3711-08).

The recorded videos in Phase I (that formed our video-based stimuli) were played back to the same individual who recorded them (i.e., participants watched themselves), such that each participant viewed video-based stimuli in Phase II. Participants provided continuous ratings of their affective feelings while watching their video-based stimuli, using a rating dial (ranging from unpleasant to neutral to pleasant). Continuous ratings were obtained since this method is non-disruptive (are not requiring stopping of the stimulus), allows online ratings that previous research suggests are less subject to recall bias, and can provide idiographic data at a high temporal resolution (Ruef and Levenson, 2007). These self-ratings, as shown in Fig. 1, reveal dynamics across the video-based stimulus segments in the degree of positive or negative affect.

3.2. Facial Landmark Localization and Tracking

In order to track relative movements of the facial landmarks solely generated by the facial muscles in the videos, we developed a robust tracking approach, in which the head movement is also tracked and decoupled from the facial landmark movements. We first employed a state-of-the-art 2D facial alignment algorithm presented in (Kazemi and Sullivan, 2014) to automatically localize 68 landmarks for each frame of the face video. Then, a 3D face model based on (Kittler et al., 2016) is used to estimate the depth information from the 2D frames and thus to achieve 3D landmark tracking.

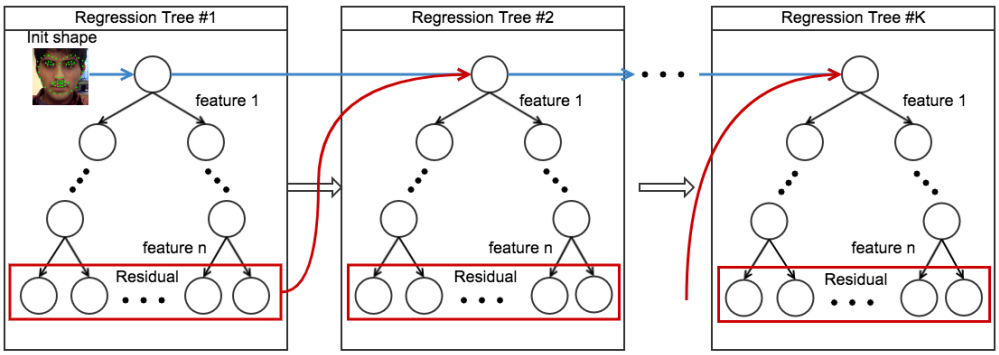

Figure 2: The framework of the cascade of regression trees designed for 2D facial landmark detection. In each level of cascade, estimated landmarks are refined by adding residuals produced by the previous regression tree.

3.2.1. 2D FACIAL LANDMARK LOCALIZATION

A cascade of trained regressors is utilized to localize the facial landmarks for each video frame as described in (Kazemi and Sullivan, 2014) (see Fig. 2). To train each regressor, the gradient tree boosting algorithm is used with a sum of square error loss (John Lu, 2010). 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 $\widehat{S}_i^{(0)}$ 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:

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

The initial shape $\widehat{S}_i^{(0)}$ for each frame is simply chosen as the mean shape of the training dataset centered and scaled according to the bounding box of the full face, detected with the histogram of oriented gradients (HOG) features. In each level of cascade, estimated landmarks are refined by adding residuals produced by the previous regression tree. Note that all frames in video are normalized to have the same Euclidean distance of pixels between the middle of the two eyes. Hence the landmark movements would be comparable across a given individual. Fig. 3 shows the result of 68 landmark detection on a video frame captured by a laptop webcam.

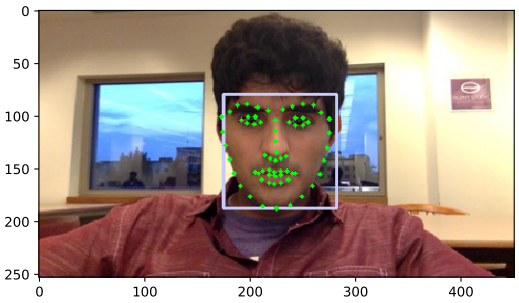

Figure 3: Results of 2D landmark detection on a webcam video frame. It outputs 68 landmarks inside a bounding box detected using HOG features.

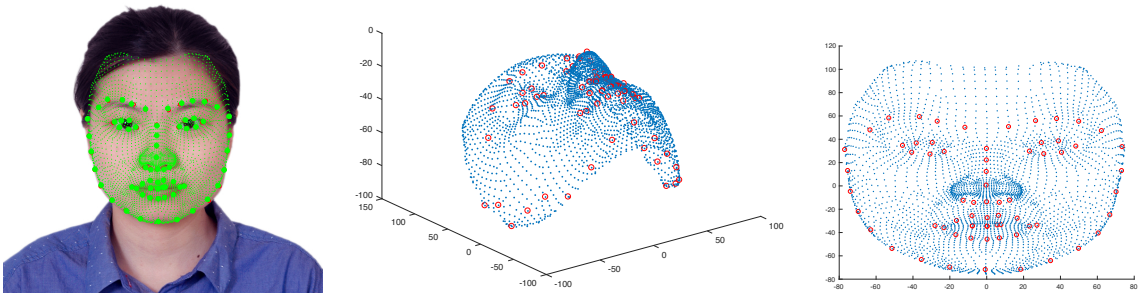

Figure 4: An example result of the landmark fitting: resulting shape and model fitting (left), 3D model fitting result (middle), and frontal view of the 3D facial model (right).

3.2.2. 3D FACIAL LANDMARK TRACKING

Our facial landmarks tracking algorithm needs to remain invariant across head movement including its translation, scaling (getting closer or further from camera), and rotations (i.e. roll, yaw, pitch). To eliminate the interference of head movement, we first extract the depth information of each face pixels from 2D video frames using the 3D morphable face model described in (Kittler et al., 2016). This model consists of a principal component analysis (PCA) model of face shapes, which could be used for reconstructing a 3D face from a single 2D image. The PCA model consists of a set of principal components $V = [\upsilon_1, ..., \upsilon_K]$, the mean value of all the facial meshes $\overline{\upsilon}$, and their standard deviation σ_k. The shape of a novel face is then generated with:

$$S_i = \overline{\upsilon} + \sum_{k=1}^{K} \alpha_k \sigma_k \upsilon_k, \tag{2}$$

where K is the number of principal components and α_k's are the representation of S_i in the coordinates of the PCA shape space. The 3D face shapes were then reconstructed by fitting 68 detected landmarks to a PCA shape model. For the purpose of model fitting, the gold standard algorithm of Hartley and Zisserman (Hartley and Zisserman, 2003) were implemented, which finds a least squares approximation of an affine camera matrix given 2D-3D point pairs. An example result of the landmark fitting is shown in Fig. 4.

Using the 3D geometric transformation matrix, every frame could be transformed into frontal face, enabling us to track and compare the movement of the facial landmarks throughout the video. The landmarks detection algorithm is not accurate for some frames with large head rotation angles or poor lighting. We address this issue by applying landmark geometric constraints.

3.2.3. FACIAL EXPRESSION FEATURES

Rather than few prototypic facial expressions, such as happiness, anger, surprise, and fear, it is shown that the dynamics and temporal combination of facial action units (AUs) may provide more reliable and specific quantification of the expressive movements of the human face during emotion (Tian et al., 2001). Guided by the work done by Y. Tian (Tian et al., 2001), we reduced the facial landmarks feature dimensions from 2×68 to 12 features described as follows: (1-2) left and right eyebrow y-values (corresponding to AUs 1, 2 and 5); (3) inner corners differences of eyebrows (corresponding to AU 4); (4) horizontal distance of the the two corners of lips (corresponding to AUs 12 and 20); (5) vertical distance of the two lips (corresponding to AUs 25, 26 and 27); (6)

average vertical positions of the two corners of the lips (corresponding to AU 15); (7-9) head rigid displacement in X, Y, and Z direction, respectively; (10-12) head rigid rotation in roll, pitch, and yaw direction, respectively. It is noted that combining the facial landmarks into AUs leads to the reduction of the stochastically distributed noise in landmark positioning as well as resolving the issue of subject-specific and camera-specific variations.

Moreover, comparing to other open source tools that can directly detect AUs, our method can provide more information related to the facial movements for two reasons. Firstly, most AU detection algorithms could only deal with frontal-view face, while our 3D landmark tracking method could also extract head movement directly from the video. Secondly, our method provides continues measures of facial landmarks instead of only 6 discrete numbers representing AUs and its intensities.

3.3. Peripheral Physiological Signal Processing

3.3.1. SIGNAL PREPROCESSING

To eliminate or reduce the noise and artifacts carried by the physiological signal measurements, signal-specific filtering is required prior to applying any feature extraction algorithm. An elliptic bandpass filter with the cut-off frequencies of 5Hz and 45Hz was applied on the ECG signals. This cut-off frequency range was selected based on the power spectral density analysis of the ECG signals and the elliptic filter type was selected to ensure the amplitude of the peak points on the signal were not significantly suppressed by the filter (Chavan et al., 2005). With the similar investigation, a low-pass filter with cutoff frequency of 1Hz was selected as the optimal choice and was applied on the EDA signals (De Luca et al., 2010). As for the respiration signal, we applied a Butterworth lowpass filter with a cutoff frequency of 20Hz.

3.3.2. PSYCHOLOGICALLY-INSPIRED PHYSIOLOGICAL FEATURES

The dimensionality reduction of the peripheral physiological time series signals is done by employing a series of signal-specific algorithms to extract informative physiological features from each signal (Nabian et al., 2017). Physiological signals including ECG, EDA, and respiration were processed using the biosignal processing MATLAB toolbox accessible at (ACl).

ECG signals contain rich information relevant to human health, sleep quality, and emotional states (Rameshwari S Mane, 2013). For the ECG signals, the features are based upon the detected QRS points on the signal (Pan and Tompkins, 1985) and after successfully detecting these points, relevant physiological features can be computed (see Table 1).

The EDA signal is composed of two types of activity, tonic and phasic. The slowly varying base signal is the tonic aspect, and is also called skin conductance level (SCL). The faster-changing part is called phasic activity or skin conductance response (SCR). SCRs are related to more acute exterior stimuli or non-specific activation (Gamboa and Fred, 2005). Many important features for this purpose are extracted from SCRs. The occurrence of the SCR is detected by finding two consecutive zero-crossings, from negative to positive and positive to negative of the bartletted differentiated EDA signal. Most of EDA features listed in Table 1 are based on the detection of SCRs. It is noted that for given windows of EDA in which no SCR was found, feature values were set to 0.

Features were also extracted from respiration signal as its pattern may vary in distinct affective states. Regular respiration is linked to relaxation, while fast and shallow breathing might correspond

Table 1: Physiology-specific feature extraction from physiological signal modalities.

Modality	Extracted Features
ECG	Features based on QRS detection: mean R-R intervals (the time between consecutive heartbeats), standard deviation of R-R intervals, standard deviation of the differences between adjacent R-R intervals, the square root of the mean of the sum of the squares of differences between adjacent R-R intervals, the number of pairs of adjacent R-R intervals where the first R-R interval exceeds the second R-R interval by more than 50ms, the number of pairs of adjacent R-R intervals where the second R-R interval exceeds the first R-R interval by more than 50ms, mean area of each QRS complex and its standard deviation.
EDA	Signal mean, numbers of detected SCRs, mean SCR duration, mean SCR amplitude, mean SCR rise-time (where rise-time of an SCR is defined as the time between the initial rise and the peak of an SCR).
Resp	Respiration rate (peak to peak in ms), amplitude (height of peak), percent inhalation (the proportion of rising part of the signal in each cycle) and percent exhalation (the proportion of falling part of the signal in each cycle).

to more aroused emotions, such as acute anxiety and emotional tension (Koelstra and Patras, 2013). Important features extracted from respiration signal are provided in Table 1.

3.3.3. WINDOWING AND OVERLAPPING SIZES FOR FEATURE EXTRACTION

Multimodal data fusion using a recurrent network requires all feature vectors to be the same length. Therefore, values for window sizing and overlapping were carefully selected for different facial expression signals as well as physiological signals to achieve same size feature vectors in all of the modalities as well as to reasonably capture the temporal dynamics of the signals. Please note that the larger window size is chosen for the signal with slower changing rate to make sure appropriate features are correctly extracted. To align different modalities in time, we applied different overlaps to force corresponding windows of different modalities to have the same starting point in time. Windowing and overlapping sizes for each signal modality are provided in Table 2.

Table 2: Signal-specific feature extraction parameters.

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

3.4. Multimodal Data Fusion using Recurrent Network

Here, we demonstrate a method introduced earlier by (Zou et al., 2015) to build an inter-system network model that represents the joint contribution between different response modalities of a per-

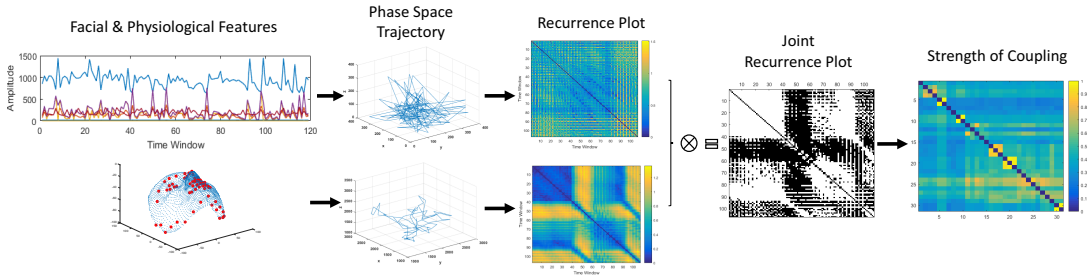

Figure 5: The framework for multimodal data fusion using recurrent network. First, signal-specific facial and physiological features are windowed and extracted. Second, the phase space trajectory for each modality is reconstructed and the corresponding recurrence plot is obtained. Then, a joint recurrence plot (JRP) is calculated by multiplying recurrence plots together. Finally, complex network metrics are extracted to assess the inter-system dynamical coupling.

son under an affective experience context. In our study, the nodes in the network represent features extracted from multimodal data (facial expressions and physiological responses) and edges are defined based on the directional coupling between modalities. The network construction consists of the following steps: (1) time-delayed embedding for reconstructing phase space trajectory (Takens, 1981); (2) recurrence plot (RP) construction (Eckmann et al., 1987); (3) extension of RP to multiple systems (i.e. modalities) to obtain a joint RP (JRP) (Romano et al., 2004); and (4) extraction of complex network metrics to assess the inter-system dynamical coupling. The general framework of the proposed fusion approach applied on our multimodal dataset is shown in Fig. 5.

3.4.1. RECURRENCE PLOT FOR SINGLE MODALITY

Introduced by Eckmann in 1987 (Eckmann et al., 1987), RP is a visualization to represent the temporal dependency relationships between all states in a time series data using a binary, squared matrix (Eckmann et al., 1987). Suppose the state of system (or modality) X at time i and j is represented by $\mathbf{x}_i, \mathbf{x}_j$, recurrence can be recorded by the binary function as:

$$\mathbf{R}^X_{i,j} = \Theta(\epsilon_X - ||\mathbf{x}_i - \mathbf{x}_j||_1), \mathbf{x}_i \in \mathbb{R}^m, i,j = 1, ...N, \tag{3}$$

where Θ is a Heaviside function and the RP puts a point at coordinates (i, j) if $R^X_{i,j} = 1$, any time the state trajectory gets sufficiently close (within the system threshold ϵ_X) to a point it has been previously.

3.4.2. EXTENSION TO MULTIMODALITY AND INVESTIGATION OF COUPLING

Although the original method was developed for a single time series, later variations of RP included consideration of multivariate time series from different aspects. To best capture the dynamic coupling between multiple modalities, we adopted JRP since it represents when a recurrence occurs simultaneously in two or more time series (Romano et al., 2004). Suppose we have modalities X and Y, for which the individual RPs can be obtained. In general, JRP is obtained by the product of multiple systems:

$$\mathbf{JR}^{X,Y}_{i,j} = \Theta(\epsilon_X - ||\mathbf{x}_i - \mathbf{x}_j||_1)\Theta(\epsilon_Y - ||\mathbf{y}_i - \mathbf{y}_j||_1). \tag{4}$$

3.4.3. COMPLEX NETWORK-BASED FEATURE EXTRACTION

We enabled the usage of complex network analysis by converting the JRP matrix to the adjacency matrix for a network, which serves as a graphical representation of the temporal neighborhood relations between system states across entire time series. This network, referred to as recurrence networks (RN), is expressed in the following formula:

$$\mathbf{A}_{i,j}(\varepsilon) = \mathbf{JR}_{i,j}(\varepsilon) - \mathbf{I}, \tag{5}$$

where $\mathbf{A}_{i,j}(\varepsilon)$ is the adjacency matrix, $\mathbf{JR}_{i,j}(\varepsilon)$ is the JRP (a binary matrix), and $\mathbf{I}(T)$ is an identity matrix for removing the elements on the main diagonal line that creates self-loops in the network.

In our study, RN is used to describe the dynamical behaviors (or patterns) of a person under an affective experience context. Further, to characterize the network features, a set of metrics are computed for quantitative assessment of the network topology, and further provide information about coupling dynamics in a different view. The network measures we computed include two general classes: (1) global measures: transitivity, global efficiency, and out-strength/in-strength correlation, and (2) local measures: in-strength/out-strength, local efficiency, edge/node betweenness centrality, diversity, and clustering coefficients. The global measures are related to the topological structures of the entire network, while the local measures are related to the attribute of individual nodes. These metrics are widely used for describing the connectivity patterns in complex systems; the computational details are included in (Rubinov and Sporns, 2010).

3.5. Inference Models for Affective Experience Prediction

3.5.1. PREDICTION BASED ON EMOTIONAL VIDEO STIMULUS CONTENT

We performed a binary across-individual classification on video stimuli with positive or negative contents. Since each subject experienced two positive and two negative video stimuli, the class distribution was balanced. First, we extracted the signal-specific features from the multimodal input signals collected from each subject while he/she was watching a video stimulus. To fuse these features, we then constructed their JRPs, as well as the network metrics containing global and local measures. We used a support vector machine (SVM) with a linear kernel function as the inference model and the stimuli context prediction accuracy was defined as the percentage of stimuli correctly classified based on their positive or negative contents. We performed a per-subject leave-one-out cross validation method for the model evaluation.

3.5.2. PREDICTION BASED ON AFFECTIVE SELF-RATING SCALES

Here, we trained an inference model to predict individualized affective experiences of participants while watching a video stimulus, rather than classifying the content of the stimulus. This inference model accounts for individual differences in responding to emotional stimuli, by assuming that the multimodal facial and physiological data collected during an affective experience is more directly linked to the person's internal affective state rather than the content of the stimulus as positive or negative.

The same signal-specific features are fused by JRP for predicting the valence self-rating of the person. The original range of valence rating was [0, 50]. We applied a Min-Max normalization to scale the valence ratings to [0, 1] for each subject. Since the actual rating range of different subjects could differ a lot, normalized values allow the comparison across subjects in a way that eliminates

the effects of certain gross influences. For each video stimulus, participants provided continuous ratings of their positive/negative affect. We trained our inference model to predict the median value of continuous ratings, since self-ratings might become dramatically high or low at the end of the video. Instead of predicting the mean value of continuous ratings, we assumed the median value is a more robust representation of the subject's valenced experience during the whole video stimulus.

We employed support vector regressor (SVR) with a ridge penalty as the self-rating regression model. The same leave-one-out cross-validation approach is applied as in the previous classification task. Root mean square error (RMSE) and mean absolute error (MAE) of the predicted valence self-rating scores are employed here to indicate the performance of our trained regression model.

4. Experimental Results

4.1. Emotional Video Stimulus Content Prediction Results

We performed a per-subject leave-one-out cross validation, where the classifier is trained on a total of 44 trials from 11 participants, and tested on the 4 videos from the remaining one participant. For each participant, classification accuracy and F1 score were used to evaluate the performance of our stimulus content classification model. Table 3 gives the classification results over each modality and video stimulus content. As a baseline, we also gave the expected values of the random classifier. According to the significance test with the resulted p-values in Table 3, the obtained classification accuracies of the trained model are not significantly more accurate than the random classifier. Overall, the result shows that video content ratings might not be predictable from the recorded data. One explanation for the low prediction accuracy of our trained model could be the fact that the data was not recorded from the participants while they were verbally explaining their positive or negative experiences, but rather we used the facial and physiological signal recordings of the participants while they were watching back their own recorded videos. In other words, all the recorded signals are not directly related to the video-based stimulus contents. In the following section, we instead predict the self-ratings, which may be more closely related to the recorded multimodal signals.

4.2. Affective Self-Rating Scale Prediction Results

Table 4 provides the results for our regression model to predict self-ratings of the subjective experience of emotion trained on the multimodal fusion data. We reported the RMSE and MAE, as the two widely used metrics for the accuracy evaluation of continuous variable estimation. Both of these values are negatively-oriented scores, in which the lower values represent higher accuracy

Table 3: Emotional video-based 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

Table 4: 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**
Random	0.40	0.37

performance. As a baseline, we also gave the expected values of a random regressor, which were found by randomly generating the value that are drawn from a normal distribution of the training data.

As evident in the Table 4, the prediction accuracy measures obtained from the models trained on either single modality and multimodal fusion data are higher than the prediction accuracy of the random classifier, except for the models trained solely on head movement features. Moreover, the highest accuracy was achieved with our proposed network-based multimodal fusion method. Note that the physiological signals are also well ranked in terms of performance for rating prediction: respiration performs second best on valence and EDA performs third best. This implies that alongside the facial video data, the physiological signals provide informative complementary descriptions of the affective experience.

In addition to our multimodal fusion method based on the recurrence network, another basic fusion method was also employed and evaluated. In the basic method, each modality of facial (i.e. landmarks and head movement) or physiological signals (i.e. ECG, EDA and Respiration) is weighed equally for the fusion. The results of facial and physiological features fused by the basic fusion method were also given in Table 4. It is observed that the prediction accuracy from fusion of facial signals outperforms the prediction accuracy based on single modalities. However, the model based on fusion of the physiological signals only outperformed the model trained on the ECG features and had lower performance than models trained on EDA and respiration data.

5. Conclusion

In this paper, we presented a multimodal approach that analyzes both facial expressions and peripheral physiological signals concurrently to identify and decode the individualized affective experiences of participants watching a series of emotional video-based stimuli. We developed a robust 3D face tracking approach, in which head movement is also independently tracked and decoupled from the facial landmark movements. Signal-specific features were then extracted from both facial and physiological signals (ECG, EDA, and respiration). We applied recurrence network for multimodal data fusion and complex network-based features were extracted from the fusion of different modalities. Finally, we validated our approach using a multimodal dataset consists of (i) facial videos and (ii) several peripheral physiological signals, synchronously recorded from 12 participants while watching 4 emotion-eliciting video stimuli. The experimental results for binary classification of video-based stimuli with positive vs. negative content showed that video content ratings might not be predictable from the recorded data from the participants when they are watching them.One poten-

tial explanation is that people might recall their feelings while watching themselves talking about past experiences, but they may not feel the same at that moment. However, the recorded signal modalities were shown to be reliable predictors of the self-rating of their affective experience at the moment of watching the videos. Our feature-level fusion approach based on the recurrence network demonstrated to improve upon single modality results, suggesting the these modalities contain complementary information in accounting for person's affective experiences.

References

Biosignal-Specific Processing (Bio-SP) Tool. http://www.northeastern.edu/ostadabbas/software/. Accessed: 2018.

Lisa Feldman Barrett and Eliza Bliss-Moreau. Affect as a Psychological Primitive. *Advances in Experimental Social Psychology*, 41:167–218, 2009. ISSN 00652601. doi: 10.1016/S0065-2601(08)00404-8.

Lisa Feldman Barrett, Karen S. Quigley, and Paul Hamilton. An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371 (1708):20160011, 2016.

Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Fully Automatic Facial Action Recognition in Spontaneous Behavior. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 223–230, 2006. doi: 10.1109/FGR.2006.55.

C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.600.

Mahesh S. Chavan, R. A. Agarwala, and M. D. Uplane. Digital elliptic filter application for noise reduction in ecg signal. *4th WSEAS International Conference on ELECTRONICS, CONTROL and SIGNAL PROCESSING*, pages 58–63, 2005.

Jeffrey F Cohn. Foundations of human-centered computing: Facial expression and emotion. *International Joint Conference on Artificial Intelligence*, pages 233–238, 2007. doi: 10.1145/1180995.1181043.

Jeffrey F Cohn, Lawrence Ian Reed, Zara Ambadar, Jing Xiao, and Tsuyoshi Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 1:0–6, 2004. ISSN 1062-922X. doi: 10.1109/ICSMC.2004.1398367.

Carlo J De Luca, L Donald Gilmore, Mikhail Kuznetsov, and Serge H Roy. Filtering the surface emg signal: Movement artifact and baseline noise contamination. *Journal of Biomechanics*, 43(8):1573–1579, 2010.

JUANI DURÁN, Rainer Reisenzein, and JOSÉMIGUEL FERNÁNDEZ-DOLS. Coherence between emotions and facial expressions. *The Science of Facial Expression*, page 1849, 2017.

J-P Eckmann, S Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. *EPL (Europhysics Letters)*, 4(9):973, 1987.

Paul Ekman. Facial expression and emotion. *The American psychologist*, 48(4):384–392, 1993. ISSN 0003-066X. doi: 10.1037/0003-066X.48.4.384.

Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. *Facial Action Coding System - Investigator's Guide*. 2002. ISBN 0-931835-01-1. doi: 10.1016/j.msea.2004.04.064.

Miaolin Fan and Chun-An Chou. Recognizing Affective State Patterns Using Regularized Learning with Nonlinear Dynamical Features of EEG. *IEEE Conference on Biomedical and Health Informatics (BHI) 2018*, 2018.

José-Miguel Fernández-Dols and Carlos Crivelli. Emotion and expression: Naturalistic studies. *Emotion Review*, 5(1):24–29, 2013.

HFS Gamboa and ALN Fred. An electrodermal activity psychophysiologic model. 2005.

Guodong Guo and Charles R. Dyer. Learning from examples in the small sample case: Face expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 35(3):477–488, 2005. ISSN 10834419. doi: 10.1109/TSMCB.2005.846658.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. Call center stress recognition with person-specific models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6974 LNCS(PART 1):125–134, 2011. ISSN 03029743. doi: 10.1007/978-3-642-24600-5_16.

Russell T. Hurlburt, Ben Alderson-Day, Charles Fernyhough, and Simone Kühn. What goes on in the resting-state? A qualitative glimpse into resting-state experience in the scanner. *Frontiers in Psychology*, 6(OCT), 2015. ISSN 16641078. doi: 10.3389/fpsyg.2015.01535.

ZQ John Lu. The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3):693–694, 2010.

Hong Jin Joo, Bora Yeon, and Kyoung Uk Lee. The impact of personality traits on emotional responses to interpersonal stress. *Clinical Psychopharmacology and Neuroscience*, 10(1):54–58, 2012. ISSN 20934327. doi: 10.9758/cpn.2012.10.1.54.

Melih Kandemir, Akos Vetek, Mehmet Gönen, Arto Klami, and Samuel Kaski. Multi-task and multi-view learning of user state. *Neurocomputing*, 139:97–106, 2014. ISSN 18728286. doi: 10.1016/j.neucom.2014.02.057.

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.

Aya Khalaf, Mohsen Nabian, Miaolin Fan, Yu Yin, Jolie Wormwood, Erika Siegel, Karen Quigley, Lisa Feldman Barrett, Murat Akcakaya, Chun-An Chou, and Sarah Ostadabbas. Analysis of multimodal physiological signals within and across individuals to predict psychological threat vs. challenge. *Available via PsyArXiv*, 2017.

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.

Sander Koelstra and Ioannis Patras. Fusion of facial expressions and EEG for implicit affective tagging. *Image and Vision Computing*, 31(2):164–174, 2013. ISSN 02628856. doi: 10.1016/j.imavis.2012.10.002.

Kuba Krys, C-Melanie Vauclair, Colin A Capaldi, Vivian Miu-Chi Lun, Michael Harris Bond, Alejandra Domínguez-Espinosa, Claudio Torres, Ottmar V Lipp, L Sam S Manickam, Cai Xing, et al. Be careful where you smile: Culture shapes judgments of intelligence and honesty of smiling individuals. *Journal of nonverbal behavior*, 40(2):101–116, 2016.

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. ISSN 10715819. doi: 10.1016/j.ijhcs.2006.04.001.

Stephen W Littlejohn and Karen A Foss. *Theories of human communication*. Waveland press, 2010.

Mohsen Nabian, Athena Nouhi, Yu Yin, and Sarah Ostadabbas. A biosignal-specific processing tool for machine learning and pattern recognition. *Healthcare Innovations and Point of Care Technologies (HI-POCT), 2017 IEEE*, pages 76–80, 2017.

Mohsen Nabian, Yu Yin, Jolie Wormwood, Karen Quigley, Lisa Barrett, and Sarah Ostadabbas. An open-source feature extraction tool for the analysis of peripheral physiological data. *IEEE Journal of Translational Engineering in Health and Medicine*, 2018.

Catie Nielson, Mohsen Nabian, Yu Yin, Jolie Wormwood, D DeSteno, Lisa Barrett, Karen Quigley, and Sarah Ostadabbas. Extracting facial synchrony from videos of naturalistic dyadic interaction. *2018 annual conference of the Society for Affective Science (SAS)*, 2018.

Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *Biomedical Engineering, IEEE Transactions on*, (3):230–236, 1985.

Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 2004.

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. ISSN 10834419. doi: 10.1109/TSMCB.2005.859075.

Maria S Perez-Rosero, Behnaz Rezaei, Murat Akcakaya, and Sarah Ostadabbas. Decoding emotional experiences through physiological signal processing. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 881–885, 2017.

Rosalind W Picard. Toward computers that recognize and respond to user emotion. *IBM Systems Journal*, 39 (3.4):705–719, 2000. ISSN 0018-8670. doi: 10.1147/sj.393.0705.

Vaibhav D Awandekar Priya Rani Rameshwari S Mane, A N Cheeran. Cardiac arrhythmia detection by ecg feature extraction. *International Journal of Engineering Research and Applications (IJERA)*, 3(2): 327–332, 2013.

Rainer Reisenzein, Markus Studtmann, and Gernot Horstmann. Coherence between emotion and facial expression: Evidence from laboratory experiments. *Emotion Review*, 5(1):16–23, 2013. ISSN 17540739. doi: 10.1177/1754073912457228.

Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. *Proc. AVEC 2015, satellite workshop of ACM-Multimedia 2015*, pages 3–8, 2015. doi: 10.1145/2808196.2811642.

M Carmen Romano, Marco Thiel, Jürgen Kurths, and Werner von Bloh. Multivariate recurrence plots. *Physics letters A*, 330(3-4):214–223, 2004.

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

Anna Marie Ruef and Robert W Levenson. Continuous measurement of emotion. *Handbook of emotion elicitation and assessment*, pages 286–297, 2007.

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.

James A Russell and José-Miguel Fernández-Dols. The science of facial expression. *Oxford University Press*, 2017.

Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, and Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. *European Conference on Computer Vision*, pages 645–661, 2016.

Kristina Schaaff and Tanja Schultz. Towards emotion recognition from electroencephalographic signals. *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, 2009. ISSN 2156-8103. doi: 10.1109/ACII.2009.5349316.

Floris Takens. Detecting strange attractors in turbulence. *Dynamical systems and turbulence, Warwick 1980*, pages 366–381, 1981.

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.

Michel F. Valstar, Timur Almaev, Jeffrey M. Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F. Cohn. FERA 2015 - second Facial Expression Recognition and Analysis challenge. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2015. doi: 10.1109/FG.2015.7284874.

Gyanendra K. Verma and Uma Shanker Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 102(P1):162–172, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2013.11.007.

Marieke Wichers, Zuzana Kasanova, Jindra Bakker, Evert Thiery, Catherine Derom, Nele Jacobs, and Jim Van Os. From affective experience to motivated action: Tracking reward-seeking and punishment-avoidant behaviour in real-life. *PLoS ONE*, 10(6), 2015. ISSN 19326203. doi: 10.1371/journal.pone.0129722.

Jing Xiao, Takeo Kanade, and Jeffrey F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002*, pages 163–169, 2002. ISSN 08999457. doi: 10.1109/AFGR.2002.1004149.

Shuang Yang, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Personalized Modeling of Facial Action Unit Intensity. *Advances in Visual Computing*, pages 269–281, 2014. ISSN 16113349.

Zhihong Zeng, Yun Fu, Glenn I. Roisman, Zhen Wen, Yuxiao Hu, and Thomas S. Huang. Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5):1–8, 2006. ISSN 17962048. doi: 10.4304/jmm.1.5.1-8.

Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009. ISSN 01628828. doi: 10.1109/TPAMI.2008.52.

Yong Zou, M Carmen Romano, Marco Thiel, and Jürgen Kurths. Identifying coupling directions by recurrences. *Recurrence Quantification Analysis*, pages 65–99, 2015.