Cross Corpus Physiological-based Emotion Recognition Using a Learnable Visual Semantic Graph Convolutional Network

Woan-Shiuan Chien, Hao-Chun Yang, Chi-Chun Lee
wschien@gapp.nthu.edu.tw, hgy@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw
Department of Electrical Engineering, National Tsing Hua University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

ABSTRACT

Affective media videos have been used as stimulus to investigate an individual's affective-physio responses. In this study, we aim to develop a network learning strategy for robust cross-corpus emotion recognition using physiological features jointly with affective video content. Specifically, we present a novel framework of Visual Semantic Graph Learning Convolutional Network (VGLCN) for individual emotional state recognition using physiology on transfer learning tasks. The stimulus of videos content is integrated into learnable graph structure to weight the importance of physiology on the two emotion dimensions, valence and arousal. Furthermore, we evaluate our proposed framework on two public emotion databases with a rigorous cross validation method, and our model achieves the best unweighted average recall (UAR), which is 67.9%, 56.9% for arousal and 79.8%, 70.4% for valence on the cross datasets recognition experiments respectively. Further analyses reveal that 1) VGLCN is especially effective on transfer valence binary-task, 2) the physiological features (ECG, EDA) are very informative features for emotion recognition and 3) the affective media videos are important constraint to be included in the framework to stabilize the performance power.

CCS CONCEPTS

• **Information systems** → **Personalization**; • **Human-centered computing** → **Ubiquitous computing**.

KEYWORDS

affective multimedia, emotion recognition, transfer learning, physiology, graph convolution network

ACM Reference Format:
Woan-Shiuan Chien, Hao-Chun Yang, Chi-Chun Lee. 2020. Cross Corpus Physiological-based Emotion Recognition Using a Learnable Visual Semantic Graph Convolutional Network. In *28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394171.3413552

MM '20, October 12–16, 2020, Seattle, WA, USA.

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413552

1 INTRODUCTION

A content-centric understanding of affect has become an increasingly "hot topic" in psychology, researchers have shown great interest in computationally understanding human's emotional state as an internal generative process when exposing to affective media [19]. At the same time, in the field of multimedia processing, the ability to automatically characterize a large amount of unstructured media content with relevant, reliable and discriminative tags is critical for intelligent indexing, retrieval and recommendation. Affective characteristics are important features for describing multimedia content, especially with its impact on viewer-induced internal affective responses [14]. These affective responses often underlie many decision making and design choices of technological solutions. For example, Chanel et al. [3] modified the difficulty of a video game according to the user's emotional state to maintain high engagement. Baveye et al. [1] provided insights on understanding the intended emotion stimulation that film maker attempts to perform in order to boost the box office through video content analysis. Developing algorithms to automatically infer subject's emotions while viewing affect-rich video data not only provides additional analytics for commercial purposes, many of the derived insights bring additional understanding in the content-centric aspect of these influential affective multimedia.

Most of the recent studies in using physiology to perform automatic emotion recognition rely on using films, which contains both visual and audio scenes resembling real-life scenarios, as stimuli to elicit physiological changes [9, 16]. In fact, there exist several public affective databases, such as Amigos [4] and Ascertain [24], that use short film clips as emotion stimuli in triggering affective internal responses that are further captured through physiological wearable devices. Some notable works of using physiology for emotion recognition includes: Yang et al. [30] proposed a network in learning personality attribute-invariant physiological representation to enhance emotion discriminability; Santamaria et al. [22] used CNN for automatic feature learning from physiological signals to predict emotion states; Feng et al. [6] and Xu et al. [29] also presented an approach of emotion recognition with complexly-designed affective features using physiology from multiple subjects. While many of these research demonstrate promising accuracies when training and testing on a single database, only few studies work on cross-corpus emotion recognition using physiological signals to investigate the generalizability of such a technology. For example, Lan et al. [15] proposed an ECG-based transfer learning framework. Sun et al. [25] investigated the relationship between two different databases via SEMG. However, to the best of our knowledge, there has not been any works in developing algorithms that generalize

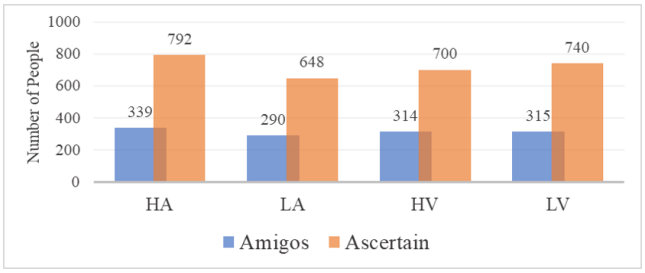


Figure 1: The number of labels of high/low arousal and high/low valence utilized in our dataset. ("H": High, "L": :Low, "A": Arousal, "V": Valence)

in performing cross corpora emotion recognition using both EDA and ECG.

In this work, our goal is to develop a robust transfer learning strategy to perform emotion recognition using physiology across databases. Our core hypothesis is that individuals would have similar subjective feelings when exposing to similar audio-visual stimuli that triggers internal physiology. Hence, by explicitly using media content as constraint in our cross corpora emotion algorithm, it would lead to an effective cross corpus recognition accuracy across databases. Specifically, we propose a Visual Semantic Graph Learning Convolutional Network (VGLCN) for cross corpora emotion recognition using physiology. Our framework is evaluated on publicly available Amigos (Am) dataset [4] and Ascertain (As) dataset [24], in which each subject is exposed to a set of audio-visual movie clips with varying degree of intended affect-triggering content. We jointly model how a person's internal physiology response to these multiple stimuli using a graph structure, then further perform the transfer tasks on the subject-wise graph-embedding learned from the collected physiological signals. In this work, we pay extra attention in carrying out our experiments, i.e., the cross validation experiments are not only subject-independent but also video stimuli-independent, when learning to transfer between these two databases. Our framework achieves the state of the art emotion recognition unweighted average recall (UAR) using physiology, which is 67.9%, 56.9% for arousal and 79.8%, 70.4% for valence in **As→Am** and **Am→As** respectively (we will explain the "cross dataset" scenario setting in 3.1).

2 METHODOLOGY

2.1 Datasets

In this study, we use two large emotional datasets collected under a similar scenario for algorithm development. In each dataset, a series of emotional videos with intended emotional elicitation (annotated with high/low arousal or valence, *-Int*) were delivered to the participants. The participants were asked to label their subjective feelings (*-Sb*) at the end of each video clips, while their physiological responses (Electrocardiogram (ECG), Electrodermal activity (EDA)) were recorded with sensors simultaneously. Our goal is to use these physiological data for recognizing each individual's self-rated emotion attributes. Specifically, we carry out the emotion recognition experiments as a binary classification problem, i.e., *-Sb* cut-off at the mean of each subject is used as label for final prediction. Figure 1 shows the number of samples in each emotion category. Several details of the dataset are listed below:

- **Amigos (Am)** [4]: A total of 16 short emotional videos (duration<250s) were carefully chosen from previous research as physiology elicitation. 40 participants aged between 21 and 40 (mean age 28.3) were recruited in a laboratory environment.
- **Ascertain (As)** [24]: Ascertain is one of the largest datasets aiming for studying physiological responses under emotional content stimuli. There are 36 short movie clips (duration 51~127s) for emotion elicitation with 58 university students (mean age 30) recruited in this dataset. The whole data collection was conducted in the laboratory environment using the commercial physiological sensor. Note that there are 11 video stimuli list in Table 1 that are the same across both the Amigos and the Ascertain databases.

Table 1: The list of the repetitive video stimuli used in both the Amigos and the Ascertain database.

Video ID		Source Movie
Amigos	Ascertain	
1	10	August Rush
2	13	Love Actually
4	18	House of Flying Daggers
6	20	My Girl
7	23	My Bodyguard
9	31	Prestige
10	34	Pink Flamingos
11	36	Black Swan
12	4	Airplane
13	5	When Harry Met Sally
16	9	Hot Shots

Table 2: An overview of physiological low-level descriptors extracted from [5]. "F*" indicates 15 statistical functions[1].

Modality	Low-Level Descriptors
ECG(51)	number_of_artifacts, RMSSD, meanNN, sdNN, cvNN, CVSD, medianNN, madNN, mcvNN, pNN50, pNN20, Triang, Shannon_h, ULF, VLF, LF, HF, VHF, Total_Power, LFn,HFn, LF/HF, LF/P, HF/P, DFA_1, DFA_2,Shannon, FD_Higushi, Average_Signal_Quality, F* Cardiac_Cycles_Signal_Quality
EDA(68)	F*SCR_Onsets, F*SCR_Peaks_Amplitudes, F*EDA_Phasic, F*EDA_Tonic_Component

2.2 Computational Framework

2.2.1 Physiological and Visual Content Descriptors. We first preprocess physiology data, i.e., a low-pass filter cut-off at 60Hz is first applied on ECG and EDA signals. Several standard low-level physiological descriptors (LLDs) are listed in Table 2 and extracted using the NeuroKit [5]. A standard z-normalization is performed subject-wise on each feature dimension to mitigate the issue of individual differences. In addition, the visual content vector of each emotion stimuli video (an individual items is a matrix v) is extracted by using the pre-trained deep 3D CNNs model with Kinetics proposed in [10] and results in frame-level descriptors of dimension 1024[13].

[1]max, min, mean, median, std, skewness, kurtosis, min position, max position, 25_percentile,75_percentile, 75_percentile-25_percentile, 1_percentile, 99_percentile, 99_percentile-1_percentile

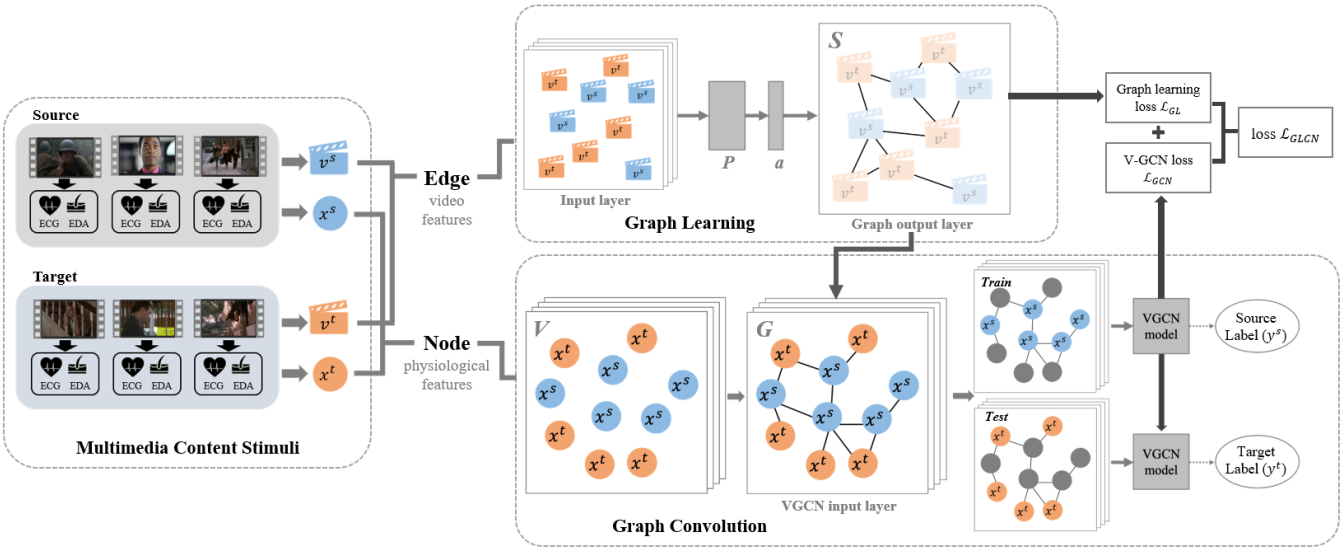


Figure 2: Our proposed Visual Semantic Graph Learning Convolution Network (VGLCN) on transfer learning. First, We retrieve all video features in both v^s and v^t for the edges $\mathcal{S}$ to learn a visual content based graph, and the node-set $\mathcal{V}$ is consist of both x^t and x^s. Then, we build a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{S}\}$. We mask x^t when training the VGCN model and mask x^s to perform prediction. Moreover, the VGLCN model will be optimized by considering both $\mathcal{L}_{\text{GCN}}$ and $\mathcal{L}_{\text{GL}}$.

To prevent the curse of dimensionality, the UMAP algorithm[18] is applied over the dataset to reduce the dimension to 32, and video-level content vectors are further aggregated using mean pooling over frames.

2.2.2 Graph Convolutional Network. Recently, Graph Convolutional Network (GCN) has received growing attention for its use in capturing structural inter-relationship among instances (nodes) and demonstrates its superior modeling power on various recognition tasks [28]. In this research, our goal is to perform emotion recognition in a transfer learning setting, i.e., a strict independently subject and stimuli while emotion recognition scenario, and learn an adaptive (or optimal) graph representation for GCN architecture. Thus, we regard it as an unsupervised domain adaptation problem. We first build a large graph $\boldsymbol{\mathcal{G}} = \{\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}\}$, where $\boldsymbol{\mathcal{V}}$ is the node-set which comprised of the physiological data and $\boldsymbol{\mathcal{E}}$ is built based on the video data. Specifically, given a video stimulation data $\boldsymbol{v}$, LLDs $\boldsymbol{x}$ and an emotional label $\boldsymbol{y}$, the objective of GCN is trained on source domain $D_s = (v_k^s, x_{ik}^s, y_{ik}^s)^{n_s}$, while preserving the transferability toward unlabeled target domain $D_t = (v_l^t, x_{jl}^t)^{n_t}$, where i and j refer to the non-overlapped subjects and k and l are the non-overlapped emotional elicitations.

Insipred by [28], we proposed a variation of GCN which performs a spectral convolution for modeling unstructured graphical data to handle the problem of transfer learning. The core GCN layer can be interpreted as a special case of a first-order differentiable message-passing framework:

$$H^{(l+1)} = \sigma(\mathrm{D}^{-\frac{1}{2}}\mathrm{A}\mathrm{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}), \tag{1}$$

where H^l denotes the l^{th} layer in the network, and D, A refers to the degree and adjacency matrix decomposed from the graph $\boldsymbol{\mathcal{G}}$. In addition, The model input H^0 is equivalent to the node matrix $\boldsymbol{\mathcal{V}}$ of the graph with shape $N \times d$, where $N = n^s + n^t$ is the number of all nodes with feature dimension d. Besides, during the forward pass, each node would perform message sharing among the linked nodes, then multiplied by a learnable weight matrix W of shape $d^l \mathrm{x} d^{l+1}$, and finally activated by a non-linearity function σ. Thus, the whole network would output a $N \times 1$ emotion state prediction for both source and target data, and all of parameters are updated through standard cross-entropy loss by giving it the source data label only:

$$\mathcal{L}_{\text{GCN}} = -\frac{1}{n^s}\sum[y^s \log \mathcal{F}(\mathcal{G}) + (1 - y^s)(1 - \log \mathcal{F}(\mathcal{G}))]. \tag{2}$$

Moreover, the overall network could be implemented and backpropagated using the sparse matrix multiplication kernel [26].

2.2.3 Learning Visual Semantic Graph. Better modeling of these videos to include a prior description of latent control of the physiology resulting from these video contents. Since the physiological responses are induced through the contents of videos, we include prior descrition as a latent control of the physiology resulting from such video contents. That is, we train a graph to learn the relationships between the contents of videos, and then integrate this with physiological responses to the GCN model. To be more specific, in order to integrate video content into the GCN, we construct a self-learning graph $\boldsymbol{\mathcal{G}} = \{\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{S}}\}$ to learn structural relationships between video contents. We follow work in [21] that looks for a nonnegative function $\boldsymbol{\mathcal{S}}$ as $S_{ab} = g(v_a, v_b)$ to represent the pairwise relationship between video data $\boldsymbol{v_a}$ and $\boldsymbol{v_b}$ from both source and target domain. More specifically, the implementation of S_{ab} is calculated in a single-layer low-dimensional embedding neural network parameterized by a projection matrix $P \in \mathbb{R}^{p \times d}$, $d < p$ and

Table 3: A summary of the experimental recognition results. 'Am' represents the dataset of Amigos. 'As' expresses the dataset of Ascertain. In VGCN and VGLCN model, "(f^v, f^p)" means the edges are created by video features and the nodes are using physiological features.

	Am → Am		**As → As**		**As → Am**		**Am → As**	
	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
SVM (f^p)	0.476	0.537	0.512	0.552	0.517	0.514	0.501	0.490
DNN (f^p)	0.562	0.599	0.536	0.573	0.560	0.527	0.518	0.513
SVM (f^p+f^v)	0.477	0.596	0.488	0.569	0.471	0.450	0.500	0.448
DNN (f^p+f^v)	0.533	0.570	0.535	0.549	0.556	0.517	0.533	0.521
DANN (f^p+f^v)	0.537	0.590	0.531	0.548	0.577	0.576	0.533	0.543
CDAN-E (f^p+f^v)	0.536	0.588	0.520	0.549	0.574	0.561	0.544	0.528
VGCN (f^p, f^p)	0.528	0.544	0.539	0.578	0.553	0.586	0.558	0.594
VGCN (f^v, f^p)	0.515	0.591	0.499	0.527	0.631	0.676	0.536	0.598
VGLCN (f^p, f^p)	-	-	-	-	0.556	0.600	0.550	0.603
VGLCN (f^v, f^p)	-	-	-	-	**0.679**	**0.798**	**0.569**	**0.704**

a weight vector $z = (z_1, z_2, \cdots, z_p)^T \in \mathbb{R}^{p\times 1}$. We then transform our input video features using $\tilde{v_a} = v_b P$, for $a = 1, 2, \cdots n$. Formally, we learn the graph edges $\boldsymbol{\mathcal{S}}$ as:

$$\tilde{S}_{ab} = g(v_a, v_b) = \frac{\exp(\mathrm{ReLU}(z^T\,|\tilde{v_a} - \tilde{v_b}|))}{\sum_{b=1}^{n}\exp(\mathrm{ReLU}(z^T\,|\tilde{v_a} - \tilde{v_b}|))}, \tag{3}$$

where $\mathrm{ReLU}(\cdot)$ is an activation function that equals to $\max(0, \cdot)$. That is to say, $\mathrm{ReLU}(\cdot)$ guarantees the nonnegativity of S_{ab}. In addition, the above operation applied on each row of $\boldsymbol{\mathcal{S}}$ is to ensure that the learned edge $\boldsymbol{\mathcal{S}}$ will be summed to one by b.

Finally, we optimize the optimal weight vector z by minimizing the following loss function,

$$\mathcal{L}_{\mathrm{GL}} = \gamma_v \sum_{a,b=1}^{n} \|v_a - v_b\|_2^2 S_{ab} + \gamma_S \|S_{ab}\|_F^2 . \tag{4}$$

With this loss function, the larger distance of $\|v_a - v_b\|_2$ between v_a and v_b is encouraged with a smaller value of S_{ab}. In other words, S_{ab} tends toward a larger weight when the distance between v_a and v_b is shorter. Besides, the second term is used to control the sparsity of learned graph $\boldsymbol{\mathcal{G}}$ because of the simplex property of $\boldsymbol{\mathcal{S}}$.

2.2.4 Visual Semantic Graph Learning Convolutional Network. Our proposed Visual Semantic Graph Learning Convolutional Network (VGLCN) integrates the self-learning mechanism of edge weights (derived from video data) into the GCN. We follow the work in [11] to learn an optimal graph representation and to simultaneously integrate graph learning and convolution to improve the unsupersived transfer recognition performance.

$$H^{(l+1)} = \sigma(\mathrm{D}_s^{-\frac{1}{2}}\mathrm{SD}_s^{-\frac{1}{2}}H^{(l)}W^{(l)}), \tag{5}$$

where the degree and adjacency matrix that is decomposed from visual semantic edge $\boldsymbol{\mathcal{S}}$ has the identical definition D_s as described in Section 2.2.2. In addition, the parameters of this network are trained by minimizing the following loss function:

$$\mathcal{L}_{\mathrm{GLCN}} = \mathcal{L}_{\mathrm{GCN}} + \lambda\mathcal{L}_{\mathrm{GL}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{GCN}}$ and $\mathcal{L}_{\mathrm{GL}}$ are defined in Eq.(2) and Eq.(4), respectively, and parameter $\lambda \geq 0$ is a tradeoff parameter.

3 EXPERIMENTAL SETUP AND RESULTS

3.1 Experimental Setup

There are two scenarios to evaluate our method:

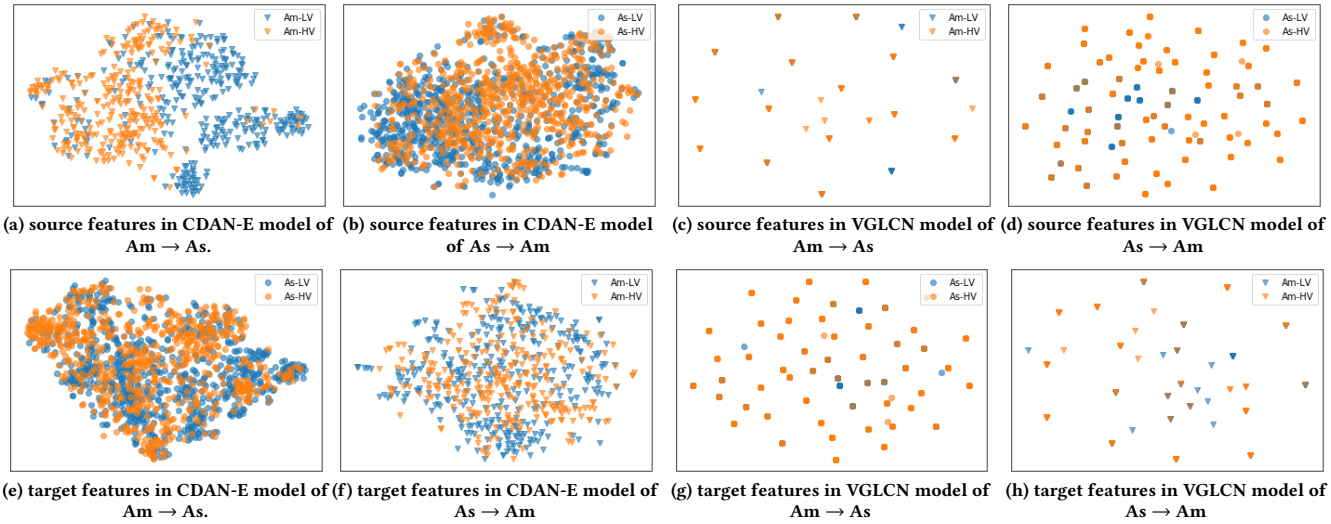
(1) Within dataset: source and target are all from either Amigos (**Am→Am**) or all from Ascertain (**As→As**) separately. A subject independent 10-fold cross-validation and a video independent 4-fold cross-validation were jointly conducted results in a total of 40-fold cross-validation to evaluate the transferability for unseen subjects stimulated under unknown video in a single dataset.
(2) Cross dataset: train the model on either Amigos or Ascertain and test on the other (**As→Am** or **Am→As**). The repeated stimuli's Table 1 are excluded to guarantee the robustness of transferability on unseen videos.

Several hyperparameters are grid-searched: dropout rate between [0.2,0.5], learning rate among [0.005,0.001,0.0005]. Batch size is fixed as [16,32], the max epoch is 500, and optimizer is Adam. In VGLCN, graph-related parameters are chosen with γ_v between [0.1,0.01], γ_S among [0.01,0.001,0.0001], and λ in loss function is setting between [0.1,0.01]. The final evaluation metric used is the unweighted average recall (UAR).

3.2 Comparison Models

We first conduct our experiments utilizing linear SVM and vanilla DNN only with the physiological features. The architecture of our DNN models includes three dense layers with dimension [d, $d/3$, $d/5$]. Then we carry out the experiments with both physiological features and video semantic features, and we compare them with the following models to examine the effectiveness of our proposed VGLCN.

- **CDAN-E:** Conditional Domain Adversarial Adaptation With Entropy Constraint. CDAN-E and DANN (Domain Adversarial Neural Networks) are used here as another approach of unsupervised domain adaptation method. These frameworks are all proposed



(a) source features in CDAN-E model of Am → As. (b) source features in CDAN-E model of As → Am (c) source features in VGLCN model of Am → As (d) source features in VGLCN model of As → Am

(e) target features in CDAN-E model of Am → As. (f) target features in CDAN-E model of As → Am (g) target features in VGLCN model of Am → As (h) target features in VGLCN model of As → Am

Figure 3: Scatter plot of the results by t-SNE for source and target features derived from CDAN-E model (a)(b)(e)(f) and VGLCN model (c)(d)(g)(h) with respect to the 2 classes of high and low valence.

by Ganin et al. [8, 17] for image classification. CDAN-E is considered as one of the state of the art methods on unsupervised domain adaptation that performs domain alignment on image features conditioned on the output of classifiers. In addition, it utilizes entropy minimization on target samples. We implement the adversarial network includes two dense layers with dimension of $[d/5, 10]$.

- **VGCN:** Visual Content Regularized Graphical Neural Network. In "within dataset" scenario, there would be 40 graphs built by utilizing all of the physiological data, i.e., the total combination of the subject and the video independent cross-validation setting. In "cross dataset" scenario, there is only one graph built by utilizing all of the physiological data.
  We first build a large graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. We then extract the visual semantic features then retrieve those videos with positive Spearman correlation coefficient from both v^s and v^t for each subject's physiology data (as a node). Therefore, any of the two nodes in $\mathcal{V}$ would be linked if their original video stimuli v have positive correlation coefficient. In short, our graph consists of the linked edges by considering the video content across both databases, where the nodes represent the physiology data. With this, we have bind both source and target into a large visual semantic graph G for further processing.
- **VGLCN:** Proposed Visual Content Regularized Graphical Learning Neural Network. Our proposed VGLCN is additional modification from VGCN model. Here, our objective is to enhance the transferability of physiology-based emotion recognition with the constraint of the video stumuli. We only consider "cross dataset" scenario, there is only one graph built $\mathcal{G} = \{\mathcal{V}, \mathcal{S}\}$ utilizing all of the video data v^s and v^t, then we will mask the physiological data from target domain x^t when training the model and mask x^s to perform prediction. The specifics of learning with visual content embedding is described in 2.2.3.

3.3 Emotion Recognition Results

Table 3 summarizes our experimental results. Our proposed VGLCN model outperforms all comparison methods on cross-database settings. The improvement is much more obvious which results in a relative gain of 10.5%, 2.5% for arousal and 23.7%, 14.4% for valence in **As→Am** and **Am→As**, respectively. Several observations can be made. First, there exists a large data discrepancy across datasets (even within the same dataset). This discrepancy results from either subject differences or heterogeneous video contents that deteriorate emotion recognition accuracy when using SVM or DNN without any strategy in constrain learning jointly with video information. Besides, directly concatenate the visual features for joint content-physio modeling deterioriate the recognition performances further. This suggests that the video features do not directly embed discriminative emotional information themselves, and the intricate dependency between video stimuli and physiological responses require a sophisticated algorithm to handle.

Second, we observe that although DANN and CDAN-E has been considered as a relatively strong baseline, the improvement compared with vanilla DNN is not obvious in this task. We hypothesize that since the domain adversarial invariant mechanism of DANN and CDAN-E mainly are focusing on mapping cross datasets' feature distribution conditioned on predicted labels, it could only minimize the discrepancy of physiological representation in a global (holistic) manner and would fail to consider the local variations (like subject or elicitation differences), which is especially critical for physiology-based emotion recognition for an individual. Therefore, we train VGCN to model the relationships between the contents of videos based on the correlations, and this mechanism works better and improves the model performance.

Furthermore, our proposed visual content-based VGLCN utilizing learnable visual semantic graph modeling that links physiological representations under similar visual stimuli, which helps in obtaining improved robustness results in the transfer setting. The

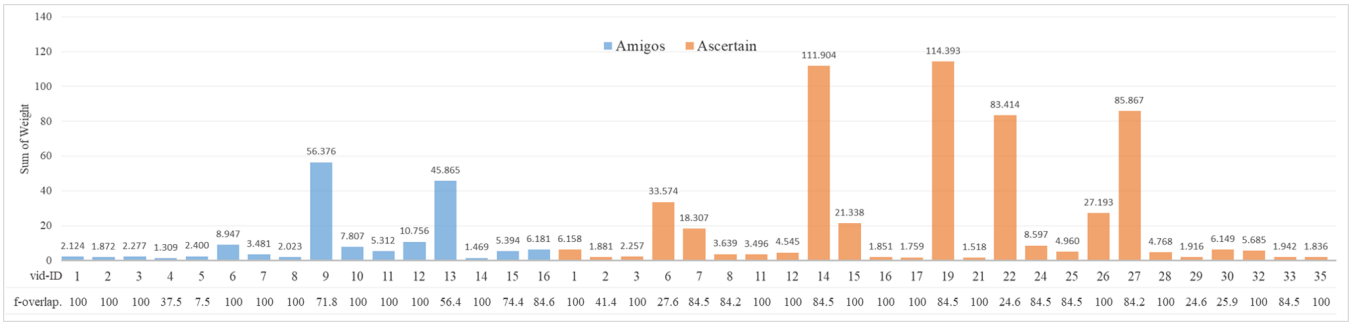


Figure 4: A histogram of the weight summation over each connected video in the learned graph. "f-overlap.": the feature identical percentage by each video ID through VGLCN model. "Sum of Weight": the total weight of each connected video in the learned graph.

major difference between VGCN and VGLCN is in the construction the graphs. VGLCN minimizes a loss function to learn the structural relationships between the video contents in a non-linear manner. On the other hand, the graph of VGCN is made according to the plain linear correlations computed between the high-dimensional features of videos. That is to say, VGLCN learns a better graph representation weights with self-learning strategy. Besides, we also observe that our VGLCN shows a larger boost on the valence prediction than arousal task. This could result from that visual contents usually delivered more valence-related messages [27]. Additional analysis result is shown in the following section.

4 ANALYSIS AND DISCUSSION

In this section, to understand the potential modulation of visual stimuli toward affective physio-responses, we specifically analyze the cross-corpus valence recognition (which has the highest UAR). Firstly, we visualize the representations along with several video-level statistics. Then, we focus on the most informative components of feature sets in this transfer task.

4.1 Visualization

To demonstrate the effectiveness of the proposed VGLCN model in the transfer tasks, we plot the features using t-Distributed Stochastic Neighbor Embedding (t-SNE) for visualization with the relatively strong baseline – CDAN-E and our proposed core method – VGLCN. In Figure 3, where (a)(b) show two-dimension visualization by reducing high-dimensional features of 2 different transfer tasks. We observe that the different labels of source features are separated, while the target features in (e)(f) are not. This indicates that the CDAN-E model can be trained better in the source dataset, but there is no transferability of emotion discriminability toward unseen data. In contrast, due to the mechanism by using video stimulus we modeled, though some of the features used to train the VGLCN model are mapped into an identical position in the plot, features in (c)(d)(g)(h) are indeed more distinguishable in either cases of using Amigos or Ascertain as source data with our proposed VGLCN.

Moreover, in order to realize what kind of features easily project the same coordinates, we calculate 2 vectors as Figure 4 shows: 1) the percentage of these identical features for each video. 2) sum up the weight of connected videos for each video from the learned graph. We manually aggregate this proportion into two groups,

Table 4: The informative features in cross dataset scenario. "*" represents the features in $p \leq 0.01$.

Modality	Low-Level Descriptors
ECG(5)	Triang*, Shannon_h*, DFA_2*, correlation_dimension*, Entropy_Multiscale_AUC*
EDA(17)	SCR_Onsets_99_percentile*, SCR_Onsets_99_percentile-1_percentile*, SCR_Onsets_max_position, SCR_Onsets_max*, SCR_Onsets_mean*, SCR_Onsets_median*, SCR_Onsets_std*, SCR_Onsets_quartile_range*, SCR_Onsets_up_quartile*, SCR_Onsets_low_quartile*, EDA_Phasic_VLF, EDA_Phasic_min position, EDA_Tonic_1_percentile, EDA_Tonic_99_percentile-1_percentile*, EDA_Tonic_std, EDA_Tonic_min, EDA_Tonic_VLF*,

all features or only partial features are mapped into an identical position, with identical video ID. The results of two-tailed Student's t-test indicate that there is a statistically difference between these two groups in both **Am** ($s = 1.931, p = 0.07394$) and **As** ($s = 2.263, p = 0.03337$), respectively. This implies that the more weights the video connects, the more diversities the feature embedding will be learned (fewer weights would result in learning an identical physiological representation in our VGLCN).

4.2 Analyses of Informative Features

We further analysis the effectiveness of our transfer between two databases by inspecting the impacts of physiological in terms of self-disclosed emotion states (*-Sb*). Firstly, according to *-Sb*, we split features into two groups, high-class or low-class valence. Moreover, we perform the two-tailed Student's t-test between two groups. The features with p-value less than 0.05 in both **Am** and **As** databases are listed in Table 4. Surprisingly, these features occupy 18.5% of all extracted physiological features. Furthermore, we train our VGLCN model only using these statistically significant physiological features, and we achieves 76% and 70% UAR in **As→Am** and **Am→As**, respectively. These promising results further indicate that these informative physiological features are very effective for cross-corpus emotion recognition and useful on the high-level emotional dimension task, valence. Besides, we also carry out the same analysis on arousal recognition tasks. There is no physiological features that pass the statistical testing result. This observation might explain the reason why our proposed model did not perform as well on arousal dimension as compared to valence. Our analysis is consistent with previous studies that physiological signals could act as internal measure for subjective feelings toward affective multimedia stimuli: Fukumotoet et al. [7] shows that the change in intensity and cycle

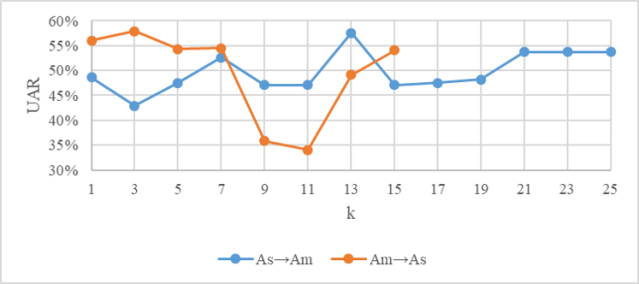


Figure 5: A line chart for the UARs under different k in As→Am and Am→As.

of respiration is related to the salient horror time points in movies. In addition, Soleymani et al. [23] shows that motion components in videos are correlated with the viewers' valence feelings when considering both multimedia content and physiological features.

Second, we have shown that the video contents are important in the modeling. Hence, in order to investigate how the videos affect the model performance, we calculate the k-nearest video stimulus for each video by directly assigning the label according to the video stimuli based on voting. That is to say, the annotation of the videos are given through the original emotional stimulation state used to trigger the viewer. However, from Figure 5, it shows the accuracy obtained with this particular set of emotion recognition for different levels of k. Here, k is the number of videos connected using k-nearest. In order to avoid the situation where the vote is tied, we only consider k to be an odd number. Specifictly, there are only 16 videos from Amigos, that is, each video from Ascertain can only find up to 15 videos from Amigos based on k-nearest algorithm. On the other hand, in "**As→Am**" scenario, the maximum value of k can be set to 25. The results also clearly show that by simply looking at similar video content does not result in a good recognition of viewer's self-reported emotion state without jointly considering physiological responses. In other words, it is important to jointly consider both the video content and the subject's physiological responses in order to achieve robust emotion recognition.

5 CONCLUSION

In this work, we present a novel framework of graph learning convolutional network for individual emotion recognition using physiological data, specifically evaluated in the context of transfer tasks (subject independent, video stimuli independent, and unsupervised cross database). The experiments show that our method reaches the state of the art emotion recognition results. To our best knowledge, this is one of the first work on emotion transfer learning that jointly considers the physiology and the video content across datasets. There are multiple future directions. An immediate one would be verifying our results on similar cross datasets such as Deap [14], Dreamer [12] and so on. Second, we will include additional modalities in the emotion stimuli, such as the sound in the video, which may help on the dimension that is more acoustically-driven (e.g., arousal). Lastly, a better understanding which components within a media clip that would trigger physiological responses linked to subject feelings would help in advancing a variety of human-centered multimedia applications [2, 20].

REFERENCES

[1] Yoann Baveye, Christel Chamaret, Emmanuel Dellandréa, and Liming Chen. 2017. Affective video content analysis: A multidisciplinary insight. *IEEE Transactions on Affective Computing* 9, 4 (2017), 396–409.

[2] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. 2017. Signal processing and machine learning for mental health research and clinical applications [perspectives]. *IEEE Signal Processing Magazine* 34, 5 (2017), 196–195.

[3] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. 2011. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 41, 6 (2011), 1052–1063.

[4] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Niculae Sebe, and Ioannis Patras. 2018. Amigos: a dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* (2018).

[5] Makowski. D. 2016. NeuroKit: A Python Toolbox for Statistics and Neurophysiological Signal Processing (EEG, EDA, ECG, EMG...). Paris, France.

[6] Huanghao Feng, Hosein M Golshan, and Mohammad H Mahoor. 2018. A wavelet-based approach to emotion classification using EDA signals. *Expert Systems with Applications* 112 (2018), 77–86.

[7] Makoto Fukumoto and Yuuki Tsukino. 2015. Relationship of Terror Feelings and Physiological Response During Watching Horror Movie. In *IFIP International Conference on Computer Information Systems and Industrial Management*. Springer, 500–507.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[9] James J Gross and Robert W Levenson. 1995. Emotion elicitation using films. *Cognition & emotion* 9, 1 (1995), 87–108.

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

[11] Bo Jiang, Ziyan Zhang, Doudou Lin, Jin Tang, and Bin Luo. 2019. Semi-supervised learning with graph learning-convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11313–11320.

[12] Stamos Katsigiannis and Naeem Ramzan. 2017. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics* 22, 1 (2017), 98–107.

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[14] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[15] Zirui Lan, Olga Sourina, Lipo Wang, Reinhold Scherer, and Gernot R Müller-Putz. 2018. Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Transactions on Cognitive and Developmental Systems* 11, 1 (2018), 85–94.

[16] Jinpeng Li, Shuang Qiu, Yuan-Yuan Shen, Cheng-Lin Liu, and Huiguang He. 2019. Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition. *IEEE transactions on cybernetics* (2019).

[17] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*. 1640–1650.

[18] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[19] Robin L Nabi and Werner Wirth. 2008. Exploring the role of emotion in media effects: An introduction to the special issue. *Media Psychology* 11, 1 (2008), 1–6.

[20] Shrikanth Narayanan and Panayiotis G Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* 101, 5 (2013), 1203–1233.

[21] Feiping Nie, Xiaoqian Wang, and Heng Huang. 2014. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 977–986.

[22] Luz Santamaria-Granados, Mario Munoz-Organero, Gustavo Ramirez-Gonzalez, Enas Abdulhay, and NJIA Arunkumar. 2018. Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS). *IEEE Access* 7 (2018), 57–67.

[23] Mohammad Soleymani, Guillaume Chanel, Joep JM Kierkels, and Thierry Pun. 2008. Affective ranking of movie scenes using physiological signals and content analysis. In *Proceedings of the 2nd ACM workshop on Multimedia semantics*. 32–39.

[24] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. 2018. ASCERTAIN: Emotion and Personality

Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing* 2 (2018), 147–160.

[25] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. 2011. A two-stage weighting framework for multi-source domain adaptation. In *Advances in neural information processing systems*. 505–513.

[26] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al. 2019. Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs. *arXiv preprint arXiv:1909.01315* (2019).

[27] Jonathan Weinel. 2018. *Inner Sound: altered states of consciousness in electronic music and audio-visual media.* Oxford University Press.

[28] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

[29] Ya Xu, Guangyuan Liu, Min Hao, Wanhui Wen, and Xiting Huang. 2010. Analysis of affective ECG signals toward emotion recognition. *Journal of Electronics (China)* 27, 1 (2010), 8–14.

[30] Hao-Chun Yang and Chi-Chun Lee. 2019. An Attribute-invariant Variational Learning for Emotion Recognition Using Physiology. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1184–1188.