Analyzing group-level emotion with global alignment kernel based approach

Xiaohua Huang, *Member, IEEE*, Abhinav Dhall, *Member, IEEE*, Roland Goecke, *Member, IEEE*, Matti Pietikäinen, *Fellow, IEEE*, and Guoying Zhao, *Senior Member, IEEE*

Abstract—From the perspective of social science, understanding group emotion has become increasingly important for teams to considerably accomplish organizational work. Currently, automatically analyzing the perceived affect of a group of people has been received increasingly interest in affective computing community. The variability in group size makes difficulty for group-level emotion recognition to straightforwardly measure the feature distance of two group-level images. Recent works attempted to resolve the preceding problem by using feature encoding. However, the early works lack of efficiency. To alleviate this problem, this paper aims to design a new method to effectively analyze the group behavior from a group-level image. Motivated by time-series kernel approaches explored in dynamic facial expression classification, this paper mainly concentrates on global alignment kernel and design support vector machine with the combined global alignment kernels (SVM-CGAK) to better recognize group-level emotion. Specifically, we first propose to use global alignment kernel to explicitly measure the distance of two group-level images. For improving the performance of global alignment kernel, we use the global weight sort scheme based on their spatial relation information to sort the faces from group-level image, making an efficient data structure to the global alignment kernel. With this new global alignment kernel, we construct the backbone of SVM-CGAK, namely, support vector machine with global alignment kernel. Furthermore, considering the challenging environment, we construct two global alignment kernels based on Reisz-based Volume Local Binary Pattern and deep convolutional neural network features, respectively. Lastly, to make the robustness of group-level emotion recognition, we propose SVM-CGAK combining both global alignment kernels with multiple kernel learning approach. It can enhance the discriminative ability of each global alignment kernel. Intensive experiments are conducted on three challenging group-level emotion databases. The experimental results demonstrate that the proposed approach achieves promising performance for group-level emotion recognition compared with the recent state-of-the-art methods.

Index Terms—Group-level emotion recognition, Global alignment kernels, Multiple kernel learning, Facial expression analysis, Convolution neural network.

1 Introduction

With the credible progress in social media, millions of images are being made available on the Internet through social networks, such as Facebook and Twitter. The large-scale data enable us to analyze human behavior during social events (for example Figure 1) in computer vision community, such as facial expression recognition [1] and speech emotion recognition [2]. Recently, several applications in computer vision community have been developed to support other fields, such as social science and health-care. For example, computer-assisted face processing assisted students with autism spectrum disorder to improve their social skills [3]. This kind of applications primarily focus on analyzing an individual's emotion. But when we consider the mood in small groups or work teams, we are more interested in knowing group emotion. From the perspective of social science, during the past century, researchers made more contributions on understanding the structure and performance of small groups [4], [5], [6], [7]. One of notable is to define group emotion. Barsäde and Gibson in [5] made a common definition about group emotion. That is, group emotion is the moods, emotions, and dispositional affects of a group of people. Additionally, group emotion influences team processes and outcomes [8]. For example, an increase in positive mood will lead to greater cooperativeness and less group conflict [9]. On the other hand, from the perspective of computer vision, consider the mood of a family posing for a group photograph at a wedding party, it is expected that there is an automated system recognizing the mood of the family. However, it is noted that the currently designed emotion detection algorithms primarily discussed the individual's emotion. Recently, several researchers studied some tasks of group-level emotion recognition, such as group-level valence and arousal prediction [10] and group-level facial expression recognition [11]. For example, in [10], Mou *et al.* aimed to predict the valence and arousal of a group of people in an image. It may give various benefits for computer vision field in future. For example, based on the correct prediction of an image, the computer vision system can automatically select the candidate photos for people to make the photo album [12]. This kind of system may also assist social scientists/researchers in the field of education to analyze the interaction of students in collaborative learning [13], etc. Therefore, in this paper we mainly focus on analyzing the basic emotions exhibited by a group of people in an image, namely, group-level emotion. In particular, motivated

- *X. Huang is with School of Computer Engineering, Nanjing Institute of Technology, China, and is also with University of Oulu, Finland. E-mail: xiaohuahwang@gmail.com*
- *A. Dhall is with the Human-Centred Artificial Intelligence, Monash University, Australia. E-mail: abhinav.dhall@monash.edu*
- *R. Goecke is with Human-Centred Technology Research Centre, University of Canberra, Australia. E-mail: Roland.Goecke@canberra.edu.au*
- *M. Pietikäinen and G. Zhao are with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. E-mail: mkp@ee.oulu.fi, gyzhao@ee.oulu.fi*

by [10], [11], [14], [15], we are mostly concerned with three tasks in group-level emotion recognition: group-level happiness intensity estimation [15], group-level valence and arousal prediction [10], and group-level facial expression recognition [11].

Groups are referred to "emotional entities and a rich source of varied manifestations of affect" [5]. Kelly and Barsäde stated that emotion influences plentifully exist in groups/teams [6]. Earlier study discussed by Barsäde and Gibson in [5] emphasized that researchers in social science community should arise group emotions with regards to the pair of a "top-down approach" and a "bottom-up approach". A "top-down approach" indicates emotion exhibited by group is represented at the group level and is felt by individual members, while a "bottom-up approach" emphasizes the unique compositional effects of individual group member emotions. Based on the framework of [5], Kelly and Barsäde in [6] further suggested that group emotion consists of its "bottom-up" components (*i.e.*, affective compositional effects) and its "top-down" component (*i.e.*, affective context). In other words, group emotion arises from both the combinations of individual-level affective factors and group-level factors, where individual-level affective factors are posed by group members and group-level factors "shape the affective experience of the group" (p.2).

Understanding behavior of groups/teams in an image or a video has recently received much attention in computer vision community. Researchers in computer vision fields designed the methods according to group emotion theory proposed by Barsäde *et al.* [5] and Kelly *et al.* [6]. Methods in computer vision can be broadly divided into two strategies: bottom-up and top-down categories. The bottom-up category uses the subject's attributes to infer group emotion. For example, in [16], Hernandez *et al.* exploited the smile of each person as the subject's attribute for inferring the emotion of the crowd. On the other hand, the top-down method considers external attributes, such as the affect of the scene and the position of the people, to describe group members. For example, Gallagher *et al.* [17] proposed contextual features based on the group structure for computing the age and gender of individuals. However, using the bottom-up or top-down approach alone for group affective analysis may miss some useful and discriminative information in an image. For example, the bottom-up method may ignore the influence of the scene on group-level emotion, while the top-down approach does not consider the person's attributes, such as the intensity of the facial expression.

To alleviate previously mentioned problem in group affective analysis, several hybrid model methods were recently proposed by combining bottom-up and top-down components for group affective analysis. They are categorized into two branches: a group expression model [12], [15], [18] and multi-modal framework [10], [11], [19], [20], [21], [22], [23]. The group expression model encodes multiple faces in a group-level image[1] into a graph structure. It concerns with the method of modeling the global and local social attributes, such as the facial attribute and scene based on a graph [24]. The earlier group expression model appeared in [12], [15]. For example, Dhall *et al.* exploited three models, namely, average, weighted, and latent dirichlet allocation based group expression models for group-level happiness intensity estimation. In particular, they used the effect of the event and the surroundings of a group as the top-down component and used the group members together with group members' attributes, such as spontaneous expressions, clothes, age, and gender as the bottom-up component. Huang *et al.* [18] proposed another group expression model for group-level happiness intensity estimation to improve the performance. They referred to the global attributes, such as the effect of neighboring group members, as the top-down component, and the local attributes, such as an individual's feature, as the bottom-up component. Nevertheless, the group expression models are not efficient in computation due to graph construction, and they cannot perform stably due to noise in the face descriptors. For example, in [15], group expression model based on latent dirichlet allocation was seriously affected by the choice of the number of clusters in k-means. It means that a large number of clusters in k-means could make the feature very sparse, while a small number could lose the discriminative information. In [18], the graph construction suffered from the false prediction of support vector regression. Additionally, group expression model cannot directly measure the distance between images by using statistical models, such as the latent dirichlet allocation.

The multi-modal framework is an alternative method for group-level emotion recognition to combine bottom-up and top-down components of images. For example, in [11], the facial action unit and facial features are regarded as the bottom-up component, while scene features are considered the top-down component. In [25], Tan *et al.*used the Xception architecture and fused image context and facial feature to recognize group-level emotion. Similar works have also appeared in [19], [20], and [21]. Another interesting multi-modal work [10] combined face and body information to predict the valence and arousal of a group of people. Some of the works on the multi-modal framework, such as [10], prefer to set up the condition for group-level emotion recognition and experiment on specific groups based on a fixed number of faces and bodies. Further, the feature encoding methods proposed by [11] used clustering methods to construct the vocabularies and to represent each image as a frequency histogram of vocabularies. This intermediate stage may introduce some errors at the classification stage. Additionally, these methods are strongly affected by the parameter design in clustering approaches.

According to our empirical analysis on group expression models and multi-modal frameworks, they lacked of adaptation to varied tasks. For example, group expression model with continuous conditional random fields [18] is not suitable to classify emotion category, as it was originally designed to estimate group-level happiness intensity. Moreover, they suffered from heavy computations due to many adjustable parameters. For example, multi-modal framework [26] contained three important parameters, *i.e.*, the dimensionality of principle component analysis, the number of kernels, and the number of face blocks. Therefore, it is worth considering whether there is an efficient and effective method for us to straightforwardly compute the distance between images, so it can be flexibly and adaptively embedded into any classifier, such as the nearest neighbor classifier or support vector machine for various tasks in group-level emotion recognition. This question leads to a relatively unexplored and new topic in group-level emotion recognition: how to formulate the distance metric for calculating the distance between images (as illustrated in Figure 1). Mathematically, we assumed two images are represented by $\mathbf{\Sigma_a} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ and $\mathbf{\Sigma_b} = \{\mathbf{y_1}, \ldots, \mathbf{y_m}\}$, respectively, we aimed to find out the distance metric function $F(\mathbf{\Sigma_a}, \mathbf{\Sigma_b})$ for better describing the distance between images.

Differing from group expression model and multi-modal

1. Group-level image is defined as an image containing more than two faces as a group. For the purpose of simplicity, we use 'image' to represent 'group-level image'.

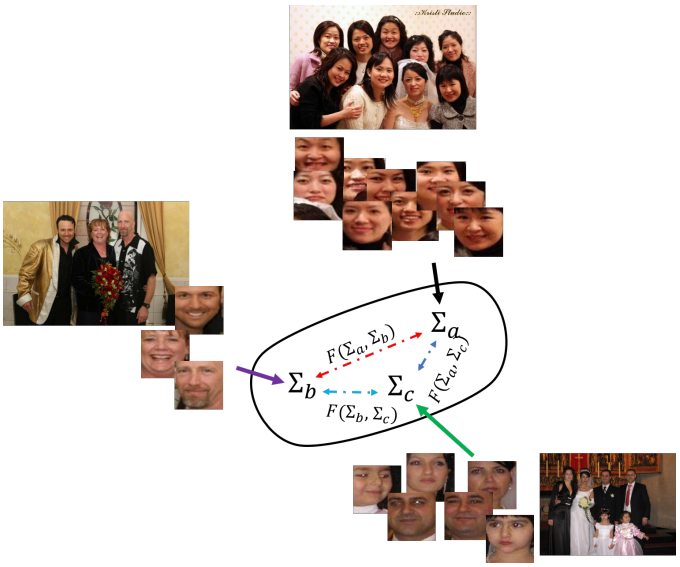


Fig. 1: An illustration of our proposed new topic in group-level emotion recognition, *i.e.*, how to formulate the distance measurement for feature sets in images. $\mathbf{\Sigma_a}$, $\mathbf{\Sigma_b}$ and $\mathbf{\Sigma_c}$ represent three sets containing the faces in three image, respectively, and F is the distance formulation which will be proposed in Section 3.

framework, we are concerned with a new method based on a distance metric function F between images, thus allowing us to straightforwardly measure the distance between images and to apply this distance metric for any classifier. As illustrated in Figure 1, the numbers of faces are not always consistent between two images. In other words, two images contain different number of faces. It is tough to directly use distance measurement, such as Euclidean distance, to measure the distance between two images $\mathbf{\Sigma_a}$ and $\mathbf{\Sigma_b}$. Recently, a family of time series kernels based on dynamic programming was exploited for constructing kernels in speech, bio-informatics, and text-processing. These time series kernels can resolve two critical issues: (1) the time series might be a variable length and (2) standard kernels for vectors cannot be captured by constructing the local dependencies between neighboring states of their time series when measuring a varied length sequence. The time series kernel approach, such as dynamic time warping [27], [28], has been investigated for action recognition [29], [30] and music retrieval [31]. However, such distances cannot be translated easily into positive definite kernels, which is an important requirement for kernel machines during the training phase. To address the positive definite problem of time series kernels, Cuturi *et al.* proposed a global alignment kernel (**GAK**) method with applications to speech recognition [32] and handwriting recognition [33]. The global alignment kernel was used for dynamic facial expression recognition to align the temporal information and demonstrated its effectiveness on facial expression recognition [34], [35]. It is observed that the global alignment kernel can better measure the time series with a variable length than other time-series kernel methods and capture the local dependencies between neighboring states of the time series. Therefore, we propose a global alignment kernel based method to directly measure the distance between two images. We first regarded the faces in an image as a set. Next, we used the global alignment kernel to measure the distance between two sets $\mathbf{\Sigma_a}$ and $\mathbf{\Sigma_b}$. For example, as illustrated in the upper image of Figure 1, we may consider the image a face sequence containing 9 faces. Then, measuring the distance between two images can be explicitly formulated as the alignment between two image sequences.

Prior to making global alignment kernel for group-level emotion recognition, it is noted that global alignment kernel suffers from the disorder of faces on the images. For example, as illustrated in Figure 1, persons in three images have different spatial positions. It is wondered how to set an appropriate and good face set on images. It aims to reduce the influence of disorder of faces and to enhance the efficiency of the global alignment. In [34], [35], they used the global alignment kernel to measure the similarity between facial expression sequences. It is observed that the facial expression videos used in their experiments occur from neutral to apex. In other words, these videos has the same phenomenon with regards to the intensity of expressions. This phenomenon makes time-series kernel, such as dynamic time warping, easily and straightforwardly find the best alignment path between two facial expression sequences. Therefore, we design a method for constructing consistent face set between two images to further enhance the good and discriminative distance metric function of global alignment kernel. The global alignment kernel will make the optimal search path from the beginning nodes of two face sets. A good face set may be useful to better calculate the distance between two images. On the other hand, we assume the group-level emotion behavior is confined in a path which people perform in orderly fashion. On the other hand, a critical problem is commonly existing in facial expression recognition: face may suffer from problems caused by challenging environments, e.g., bad illumination and head pose change. In general, we can considerably explore multiple robust feature descriptors for describing faces in images, but it is non-trivial to compute distance between multiple feature sets of multidimensional features. Here, we develop low-level and high-level features for enhancing the robustness of facial expression representation to the challenging environments and feed them into two separate global alignment kernels. Next, we propose to exploit multiple kernel learning method to combine two global alignment kernels for group-level emotion recognition, since multiple kernel learning has been commonly used and demonstrated to achieve promising performance in many fields [36], [37].

The key-contributions of this paper are described as follows: (1) Global weight sorted scheme is presented to construct efficient face sets amongst images and further evaluated its importance to global alignment kernel, such that it can enhance global alignment kernel more effectively by comparing with randomly sort; (2) global alignment kernel with global weight sorted scheme is proposed for measuring the distance between two images and is embedded into support vector machine for group-level emotion recognition; (3) multiple kernel learning approach is used to learn the optimal weights for two global alignment kernels based on two respective features, and Support vector machine with combined global alignment kernels is proposed to infer the perceived group-level emotion; and (4) comprehensive experiments on three 'in-the-wild' databases demonstrate the superiority of the proposed methods over most of state-of-the-art methods in three different tasks of group-level emotion recognition: group-level happiness intensity estimation, group-level valence and arousal prediction, and group-level facial expression recognition.

The remainder of the paper is organized as follows. Section 2 describes three challenging databases related to group-level emotion recognition. Section 3 presents the method for formulating the

distance metric function and derives our proposed approach for group-level emotion recognition. Section 4 presents and discusses the experiment results with empirically and statistically significance analysis. Section 5 concludes the paper.

2 Description of Group-level Emotion Databases

In this paper, we will conduct algorithm analysis and evaluate our proposed methods for group-level emotion recognition on the followed three databases: Happy People Images (**HAPPEI**) database [15], Multi-modal Emotion Valence and Arousal (**MultiEmoVA**) dataset [10], and Group Affective Database 2.0 (**GAFF**) [38]. The corresponding ground truth used in the experiments is provided by the authors of these three databases. Data collection, ground truth, and experimental protocols in experiments are summarized in Table 1.

HAPPEI: The HAPPEI database (in Figure 2) was collected by Dhall *et al.* [15] in 2015. This database contains 2,638 images. All images were annotated with a group-level mood intensity by four human labellers. The mood was represented by the happiness intensity corresponding to six stages of happiness (0-5). That is, Neutral, Small smile, Large smile, Small laugh, Large laugh and Thrilled. In this database, the labels are based on the perception of the labelers. The number of images with regard to classes is 92, 147, 774, 1256, 331, and 38 for neutral, small smile, large smile, small laugh, large laugh, and thrilled, respectively. The aims of this database in [15] are to infer the perceived group mood as closely as possible to human observers and to estimate the happiness intensity of images. Following the experimental protocol of [15], we chose the first 2,000 images in the experiment. Therefore, the updated number of images in each class is 73, 122, 600, 929, 241, and 35 for neutral, small smile, large smile, small laugh, large laugh, and thrilled, respectively.

Fig. 2: Six sample images containing groups in social events, annotated with six happiness intensities in the HAPPEI database [15].

MultiEmoVA: The MultiEmoVA database was collected by Mou *et al.* [10] from Google Images and Flickr by using several key words, such as graduate ceremony and party, etc. There are 250 color images annotated by 15 labelers along valence-arousal dimensions (negative/neutral/positive for valence and low/medium/high intensity for arousal). Figure 3 illustrates six images along valence/arousal dimension. As reported by Mou *et al.* [10], the inter-labeler agreement based on Cronbach's α was 0.85 and 0.96 for arousal-level and valence-arousal, respectively. In addition, they re-organized the annotation by fusing arousal-level and valence-level, therefore, it contains 46, 64, 31, 27, 10, and 72 images for high-positive, medium-positive, high-negative, medium-negative, low-negative, and neutral categories, respectively. Figure 4 illustrates six examples on the fused valence-arousal dimensions picked from the MultiEmoVA database.

Fig. 3: Six ground truth images along valence and arousal dimensions in the MultEmoVA database [10].

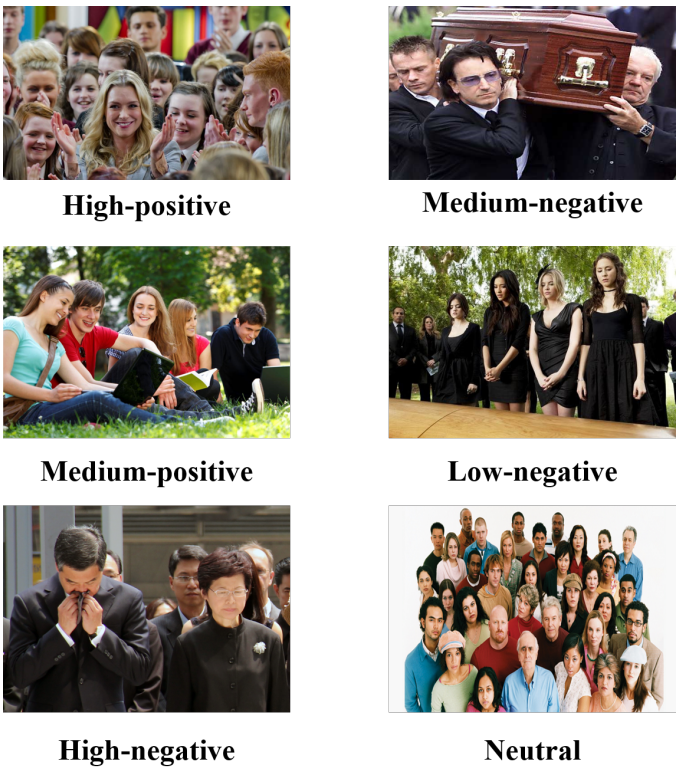


Fig. 4: Six ground truth images along fused valence-arousal dimensions in the MultEmoVA database [10].

TABLE 1: Emotion category, the number of images and our experimental protocol of three databases for group-level emotion recognition.

Databases	Emotion Category	Image Size	Experimental Protocol
HAPPEI	Six-level happiness intensity	2,000 images	4-fold cross validation
MultiEmoVA	3 categories for arousal and 3 categories for valence	250 images	5-fold cross validation
GAFF	Positive, neutral and Negative	5695 images	3630 for training and 2068 for testing

GAFF: The GAFF database was firstly proposed in [11], created from Flickr and Google images according to the keyword search, such as festival, silent protest, and violence. All the images were annotated as 'positive', 'neutral', and 'negative'. However, there were around 504 images in the first version. Recently, for an open emotion challenge competition, the number of images in the GAFF database greatly increased to 5,698 by adding more images from Flickr and Google images [38]. All the images were divided into Train (3,630 images) and Validation (2,068 images) sets for experiments. Figure 5 shows six images, two for each emotion category.

Fig. 5: Two ground truth images of each emotion category in the GAFF database [38].

3 METHODOLOGY

3.1 Problem Formulation

Given an image $\mathbf{Y}$, we aim to predict the group-level emotion using the faces based information. In this paper, the support vector machine (**SVM**) is used as the classifier. Its basic formulation is described as follows,

$$\begin{aligned} g(\mathbf{Y}) &= \sum_{i=1}^{N} \omega_i l_i \Phi(\mathbf{X}_i) \cdot \Phi(\mathbf{Y}) + b \\ &= \sum_{i=1}^{N} \omega_i l_i K(\mathbf{X}_i, \mathbf{Y}) + b, \end{aligned} \tag{1}$$

where N is the number of images in the training set, Φ is a non-linear mapping function, '$\cdot$' denotes the inner product operator, $\mathbf{X}_i$, l_i and ω_i are the i-th training sample, the corresponding class label, and its Lagrange multiplier, respectively, K is a kernel function, and b is a bias of SVM.

It is noted that the kernel function K plays an important role in Equation 1. The variability in group size [2] makes difficulty to construct the kernel function K for group-level emotion recognition. For example, in Figure 6, there are 3 and 9 faces existing in two upper-part and bottom-part images, respectively. Basically, we can use the fixed group size strategy proposed in [10]. That is, Mou *et al.* designed several specific groups based on the fixed number of faces for group-level emotion recognition. However, the fixed group size strategy seriously restricts the application of group-level emotion recognition, as in the real-world situation group size in images may be not fallen in these specific groups. Therefore, it make us reconsider how to measure the distance of two images for the kernel function K. For simplicity, we name this case "group size variability problem".

Currently, there are numerous methods to resolve that problem above. For example, in [11], Dhall *et al.* used Bag-of-Visual-Words, where image features are regarded as the words, to accumulate a histogram from multiple faces for representing the feature of an image. However, the obtained feature is very sparse. In [26], Huang *et al.* proposed an information aggregation method to encode the histograms of blocks of faces for representing the feature of an image. Although this approach can make the feature of images not sparse, it suffers from quite many parameters, such as block number and reduced dimension of Principal Component Analysis, need to be manually adjusted. Thus, we primarily consider how to search a simple but effective way to address "group size variability problem" and how to construct the kernel function for group-level emotion recognition.

In summary, the mathematical description is described as follows:

Assuming two images $\mathbf{X}_i$ and $\mathbf{X}_j$ contain M_i and M_j faces, respectively, we extract their corresponding faces features denoted as $\{\mathbf{f}_m\}_{m=1}^{M_i}$ and $\{\mathbf{g}_n\}_{n=1}^{M_j}$. Thus, the distance measurement between $\mathbf{X}_i$ and $\mathbf{X}_j$ is formulated as follows:

$$s(\mathbf{X}_i, \mathbf{X}_j) = f(\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{M_i}\}, \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{M_j}\}), \tag{2}$$

2. Group size means the number of peoples in an image.

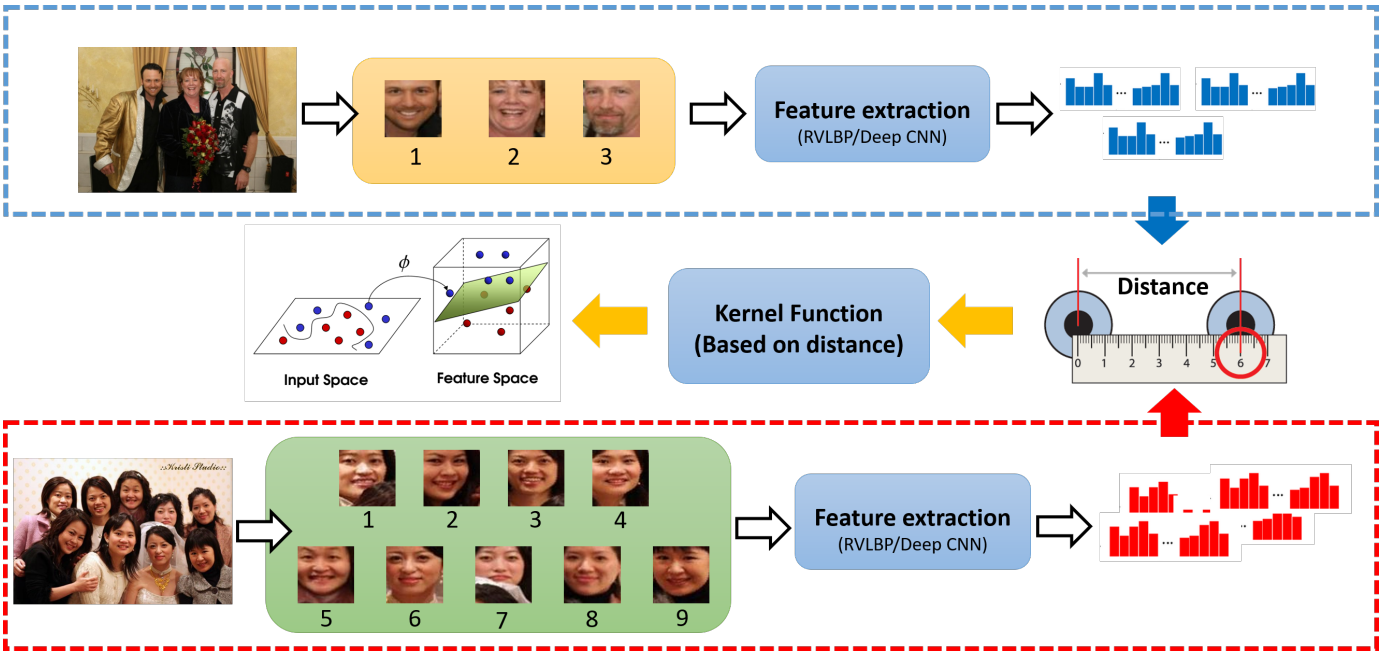


Fig. 6: Illustration of the 'group size variability problem' for measuring the distance between two images. The number under the face image represents the order index of an image obtained by face detection. The upper and bottom dotted-line blocks show the pipeline of feature extraction. The objective of addressing the 'Group Size Variability Problem' is to search proper distance function and classifier, as shown in the middle pipeline.

where s is the distance between images, and f represents the distance measurement function calculating the distance of $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{M_i}\}$ and $\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{M_j}\}$. In next section, we will discuss how to derive the distance measurement function f in Equation 2 and how to construct the kernel function in Equation 1 for group-level emotion recognition.

3.2 SVM with the Combined Global Alignment Kernels

In this section, we first utilized the 'global weight sort' method for obtaining an efficient data structure. We second detailed the method to derive the distance measurement function f in Equation 2 and the backbone of support vector machine based on combined global alignment kernels method, namely, SVM based on global alignment kernel (**SVM-GAK**) for group-level emotion recognition. Additionally, we took two examples of measuring the distance of two images and analyzed the influence of 'Global weight sort' to SVM-GAK. Lastly, we proposed the SVM with the combined global alignment kernels (**SVM-CGAK**) by combining two global alignment kernels for group-level emotion recognition.

3.2.1 Global weight sort

Given an image $\mathbf{X}_i$ containing M_i faces, denoted as $\mathbf{x}_1, \ldots, \mathbf{x}_{M_i}$, a fully connected graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is constructed to map the global structure of faces in a group, where $\mathbf{V}$ is a non-empty set of faces $\mathbf{x}_1, \ldots, \mathbf{x}_{M_i}$, and an edge $\mathbf{E}_{k,l}$ represents the link between $\mathbf{x}_k$ and $\mathbf{x}_l$. For obtaining $\mathcal{G}$, the minimal spanning tree algorithm [39] is implemented, providing the location and minimally connected neighbors of a face. Based on $\mathcal{G}$, we presented relative face size S_k estimating relationship of faces and relative distance δ_k representing the influence of neighboring faces for each face. They are obtained as follows:

- Relative face size: For $\mathbf{x}_k$, the face size is taken by $d_k = \| p_{L,k} - p_{R,k} \|$, where $p_{L,k}$ and $p_{R,k}$ are the coordinate of the left and right eyes, respectively. Next, the relative face size S_k of $\mathbf{x}_k$ is given by $\frac{d_k}{\sum_{j=1}^{n} \frac{d_j}{n}}$, where n is the number of neighboring faces of $\mathbf{x}_k$.
- Relative distance: Based on the nose tip locations of all faces in an image, their centroid c_g is computed by using $\frac{\sum_{k=1}^{M_i} p_k}{M_i}$, where p_k is the coordinate of the nose tip of the k-th face. Furthermore, the relative distance δ_k of the k-th face is described as $\delta_k = \| p_k - c_g \|$, and δ_k is further normalized based on the mean relative distance.

Therefore, the global weight w_k of $\mathbf{x}_k$ is obtained by:

$$w_k = \| 1 - \lambda \delta_k \| * S_k, \tag{3}$$

where λ controls the effect of these weight factors on the global weight. In our method, we empirically set λ as 0.1.

Then, face feature set of $\mathbf{X}_i$ is sorted according to decreasing global weights, denoted as $\widehat{\mathbf{X}}_i = \{\widehat{\mathbf{f}}_1, \ldots, \widehat{\mathbf{f}}_{M_i}\}$. For the sake of simplicity, we remove $\widehat{\ }$ out of $\widehat{\mathbf{X}}_i = \{\widehat{\mathbf{f}}_1, \ldots, \widehat{\mathbf{f}}_{M_i}\}$ in the following discussion.

3.2.2 Construction of distance measurement

We aim to calculate the distance between $\mathbf{X}_i$ and $\mathbf{X}_j$ in various ways by distorting them. An optimal search path π has a length P and $P < M_i + M_j - 1$, since the two face sets have $M_i + M_j$ points and they are matched at one point of the search path. A path π is a pair of non-decreasing integral vectors (π_1, π_2) of a length p such that $1 = \pi_1(1) \leq \ldots \leq \pi_1(P) = M_i$ and $1 = \pi_2(1) \leq \ldots \leq \pi_2(P) = M_j$, with unitary increments and no simultaneous repetitions. Let $\mid \pi \mid$ denote the length of path π, the distance measurement function for $\mathbf{X}_i$ and $\mathbf{X}_j$ for Equation 2 can be defined as follows:

$$s(\mathbf{X}_i, \mathbf{X}_j) = \sum_{p}^{|\pi|} \phi(\mathbf{f}_{\pi_1(p)}, \mathbf{g}_{\pi_2(p)}), \tag{4}$$

where ϕ is a local divergence that measures the discrepancy between any two points $\mathbf{f}_{\pi_1(p)}$ and $\mathbf{g}_{\pi_2(p)}$.

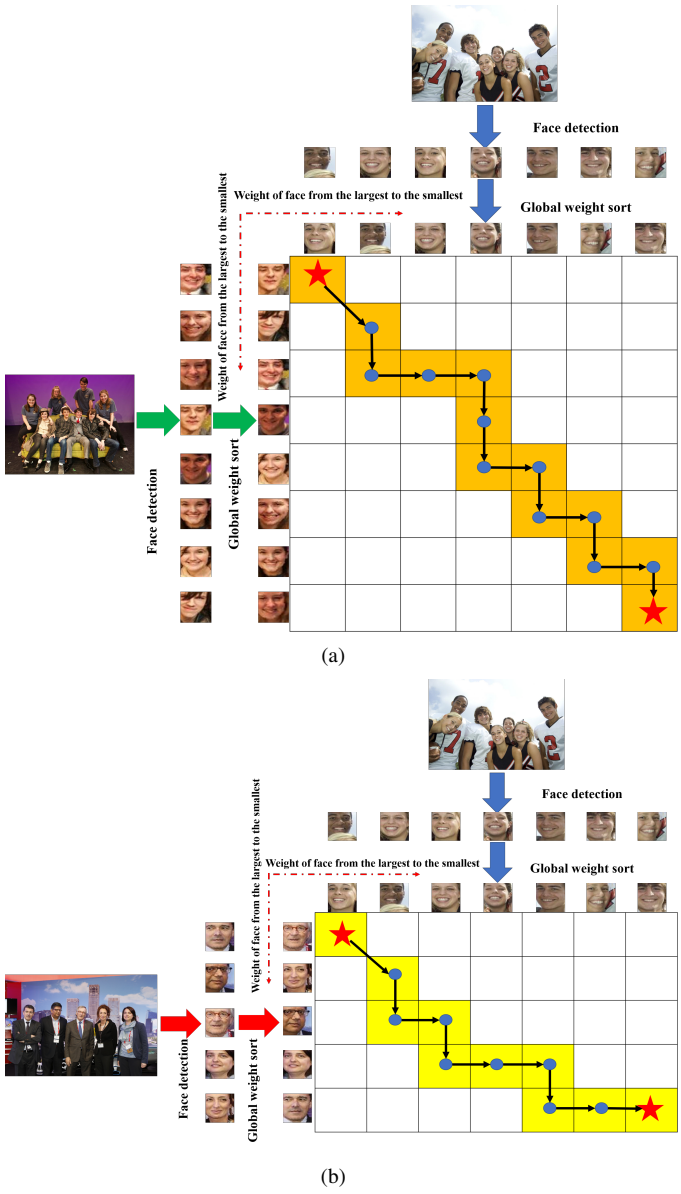


Fig. 7: Two examples of the distance measurement based on global alignment kernel, where the upper image in (a) and (b) is with 'positive' class label, and the left images in (a) and (b) are with 'positive' and 'neutral' categories, respectively. Face set is sorted by global weight sort method. The red star shapes mean the beginning and end nodes of the search path, respectively. The black arrow is the optimal search path direction. According to the optimal search path by Equation 6, the distances between images are 0.42143 and 0.48024 for (a) and (b), respectively. Source: the GAFF database [38].

For resolving Equation 4, the global alignment is proposed to calculate the distance between the images, as it considers that the minimum value of alignments may be sensitive to peculiarities of the time series and uses all alignments weighted exponentially. It can be further defined as the sum of exponentiated and sign-changed costs of the individual alignments such as $k(\mathbf{X},\mathbf{Y}) = \sum_{\pi\in A(m,n)} e^{(-s_{\mathbf{X},\mathbf{Y}}(\pi))}$. According to divergence ϕ and Equation 4, the global alignment kernel is therefore formulated as follows,

$$k(\mathbf{X}_i,\mathbf{X}_j) = \sum_{\pi\in A(m,n)} \prod_{p}^{|\pi|} e^{-\phi(\mathbf{f}_{\pi_1(p)},\mathbf{g}_{\pi_2(p)})}, \tag{5}$$

where $A(m,n)$ is the set of all paths between two feature sets $\mathbf{X}_i$ and $\mathbf{X}_j$.

It has been argued by [32] that $e^{-\phi}$ in Equation 5 goes through the whole spectrum of the costs along with all paths. Additionally, it gives rise to a smoother measure than the minimum of the costs of some classical time-series align kernel such as DTW. Following the suggestion by [32], we use a local kernel described as follows:

$$k_{\mathrm{GA}}(\mathbf{X}_i,\mathbf{X}_j) = \sum_{\pi\in A(m,n)} \prod_{p}^{|\pi|} e^{-\phi_\sigma}, \tag{6}$$

where $\phi_\sigma = \frac{1}{2\sigma^2} d(\mathbf{f}_{\pi_1(p)},\mathbf{g}_{\pi_2(p)}) + \log(2 - e^{-\frac{d(\mathbf{f}_{\pi_1(p)},\mathbf{g}_{\pi_2(p)})}{2\sigma^2}})$, d is the distance function, and σ is the standard deviation.

Figure 7 describes the work flow of the global alignment kernel on measuring the distance between images. The global alignment kernel will find the optimal search path π between two face sets, and then calculate the distance with respect to the optimal path. With the global alignment kernel, the distances between the images in the Figure 7(a) and the Figure 7(b) are 0.42143 and 0.48024, respectively. On the other hand, with the averaging distance approach, the distances between the images in the Figure 7(a) and the Figure 7(b) are 1.2856 and 1.2287, respectively. It is seen that the global alignment kernel can make the images of the same class be close to each other while the images from different classes be far from each other. It implies that the global alignment kernel can reserve the discriminative information. The global alignment kernel can provide the discriminative information to SVM. It is also seen that the global alignment kernel is flexible when it calculates the distance between two images with various group size.

Based on Equation 6, the basic form of SVM in Equation 1 is rewritten as follows:

$$g(\mathbf{Y}) = \sum_{i=1}^{N} \omega_i l_i K_{\mathrm{GA}}(\mathbf{X}_i,\mathbf{Y}) + b, \tag{7}$$

where N is the number of images in the training set.

3.2.3 *Analysis of the 'global weight sort'*

A general method, namely, 'holistic method', uses all neighboring faces to each face for obtaining the graph. However, this method may not provide the relative position of each face in a group. Also, it may introduce noise to the graph, caused by the isolated faces. Here, we discuss the influence of 'global weight sort' to the global alignment kernel by comparing with two other sorting methods. We conducted an experiment to compare the 'global weight sort' method based on the minimal spanning tree algorithm with the 'global weight sort' method based on the holistic method. Additionally, we compared with SVM-GAK without global weight sort on the HAPPEI database [15]. Specifically, 'without global weight sort' means that we used face detector [42] to automatically localize multiple faces and output its subsequent face results according to its search order. Following the experiment protocol of Huang *et al.* [18], we used a 4-fold-cross-validation protocol to analyze the influence of 'global weight sort' to SVM-GAK,

TABLE 2: Performance comparison of SVM-GAK without using global weight sort, SVM-GAK using 'holistic method' and SVM-GAK based on 'minimal spanning tree algorithm', where mean absolute error is used as a performance metric. The last column is the average mean absolute error of all features corresponding with global weight sort method.

Global weight sort method	**Features**			**Average**
	Local binary pattern [40]	Local phase quantization [41]	CNN	
Without global weight sort	0.5992	0.5810	0.5006	0.5603
Holistic method	0.5717	0.5681	0.5032	0.5477
Minimal spanning tree algorithm	0.5690	0.5674	0.4999	0.5454

TABLE 3: Two-sample left tailed t-test results of three pairs (HM vs. NO, MST vs. NO, and MST vs. HM) of Table 2. For the sake of analysis, we abbreviate 'without global weight sort', 'holistic method' and 'minimal spanning tree algorithm' as NO, HM, and MST, respectively.

	HM vs. NO	MST vs. NO	MST vs. HM
p-value	0.0664 ($> .05$)	0.0454 ($< .05$)	0.5263 ($> .05$)

where 1,500 images were chosen for training and 500 for testing, repeating four times. Mean absolute error is used as the metric for estimating happiness intensity of images. We also used Local Binary Pattern [40], Local Phase Quantization [41] and deep convolutional neural network (CNN) (*i.e.*, L2-normalization on output of FC6-layer of VGG-face) as features. To make a fair comparison, σ for SVM-GAK is set as 10 and Euclidean distance is used for $d(\mathbf{f}_{\pi_1(p)}, \mathbf{g}_{\pi_2(p)})$.

We performed statistical analysis for a pair of algorithms between the SVM-GAK without global weight sort, the SVM-GAK with holistic method and the SVM-GAK with minimal spanning tree algorithm. Here, for the sake of analysis, we abbreviated 'without global weight sort', 'holistic method', and 'minimal spanning tree algorithm' as NO, HM, and MST, respectively. Firstly, we assumed the null hypothesis that the data from these three sorting methods come from normal distribution with the same variance. Specifically, we used two-sample F-test for equal variances. The results of F-test indicated that all data coming from these three methods have equal variances. Further, we conducted two-sample left tailed t-test for the pair between these three methods. The null hypotheses are described as following:

- For HM vs. NO, the null hypothesis is the mean absolute error of HM is more than NO.
- For MST vs. NO, the null hypothesis is the mean absolute error of MST is more than NO.
- For MST vs. HM, the null hypothesis is the mean absolute error of MST is more than HM.

We aimed to see to see (1) whether MST or HM is significantly better than NO and (2) whether MST is significantly better than HM. The comparative results in terms of mean absolute error are presented in Table 2. Additionally, the p-values results are reported in Table 3.

As seen from Table 3 in HM vs. NO, the null hypothesis the mean absolute error of HM is more than NO' is accepted, but for MST vs. NO, the null hypothesis the mean absolute error of MST is more than NO' is rejected. It indicated that the improvements boosted by 'minimal spanning tree algorithm' for SVM-GAK are more considerable than without using 'minimal spanning tree algorithm'. Furthermore, we took the CNN feature for SVM-GAK for example as analysis. SVM-GAK with 'minimal spanning tree algorithm' obtained the lowest mean absolute error of 0.4999 than SVM-GAK 'without global weight sort'. This may be explained by: (1) we used the 'global weight sort' scheme to extract the consistent structure of faces in an image and (2) SVM-GAK obtained the better optimal path between faces in two images based on that structure. On the other hand, compared with SVM-GAK 'without global weight sort', SVM-GAK 'with the global weight sort' obtained considerable improvement over all features. Global weight sort scheme sorts the faces according to their importance in the image. It can efficiently provide the consistent graph structure when SVM-GAK computes the optimal path between faces in two images. The comparative results showed that the significant sorted related position of faces can affect the performance of SVM-GAK.

According to p-value in MST vs. HM, we accepted the null hypothesis for MST vs. HM. It indicated that the improvements obtained by the 'minimal spanning tree algorithm' are not significant. However, compared with the performance over all features, the improvement obtained by the minimal spanning tree algorithm was still competitive. For 'holistic method', the poor performance may be caused by the isolated faces.

3.2.4 SVM based on Combined Global Alignment Kernel

In Section 3.2.2, we presented the SVM-GAK method for group-level emotion recognition. It is important to extract the features from the faces for the global alignment kernel, as the appropriate feature can better measure the distances of faces. The existing approaches on multiple feature fusion [43], [44] show using multiple features can use the benefit of different features and obtain the better performance than using sole feature. Additionally, the existing group-level emotion recognition databases were collected from the Internet and suffer from the noise caused by poor illumination, head pose and bad image quality. Therefore, we proposed an effective way to combine multiple features in the global alignment kernel for group-level emotion recognition. To deal with the problems caused by the challenging environments, such as blurred faces and poor illumination, we used deep convolutional neural network (**CNN**) feature [45], [46] as the high-level feature, and Riesz-based Volume Local Binary Pattern (**RVLBP**) [18] as the low-level feature for global alignment kernel. The description of features and their corresponding parameter setups can be referred to Supplementary Materials - A and B.

Considering deep CNN and RVLBP features for SVM-GAK, a simple method is to concatenate them into one feature vector $\mathbf{X}$ and then to input them into Equation 7. However, the feature concatenation method fails to consider the complementary information between both features. Instead, we combined them in an alternative way. Among kernel combination methods, multiple

kernel learning has been demonstrated as a simple yet effective method to combine different features [47], [48]. Due to the efficiency of multiple kernel learning, we used multiple kernel learning to learn the optimal weights for two global alignment kernels. And then, we combined two global alignment kernels with the optimal weights. Finally, we proposed an support vector machine with combined global alignment kernels, namely, **SVM-CGAK**, for group-level emotion recognition. It is illustrated in Figure 8.

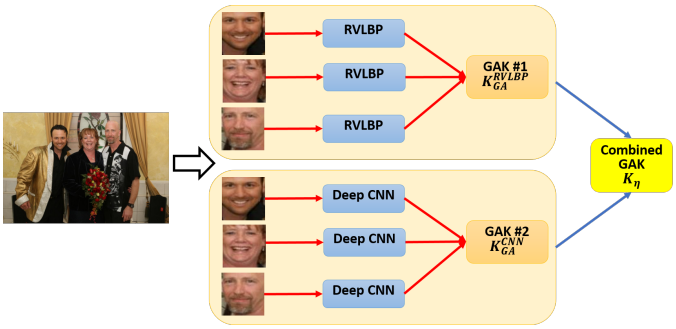


Fig. 8: Illustration of constructing SVM-CGAK. It consists of three stages: (1) RVLBP and deep CNN features are separately extracted from faces; (2) two global alignment kernel is generated, which are denoted as $K_{\text{GA}}^{\text{RVLBP}}$ and $K_{\text{GA}}^{\text{CNN}}$ for RVLBP and deep CNN features, respectively; (3) subsequently, the combination strategy is used to fuse both kernels.

In our method, we chose the Chi-Square distance [49] and the square Euclidean distance to define the local divergence in Equation 4 for RVLBP and deep CNN features, respectively. Based on these pre-designed kernels, we further combined both kernels by using multiple kernel learning, which was described as follows,

$$K_\eta = \beta_{\text{RVLBP}} K_{\text{GA}}^{\text{RVLBP}}(\mathbf{X}_i, \mathbf{Y}) + \beta_{\text{CNN}} K_{\text{GA}}^{\text{CNN}}(\mathbf{X}_i, \mathbf{Y}), \quad (8)$$

where β_{RVLBP} and β_{CNN} are the weights for RVLBP and deep CNN features, respectively.

With the combination strategy, the SVM can be given by

$$g(\mathbf{Y}) = \sum_{i=1}^{N} \omega_i l_i K_\eta(\mathbf{X}_i, \mathbf{Y}) + b. \quad (9)$$

Lemma 1. Let A_i be a positive definite matrix. If $\lambda_i > 0$ is a real number, then $\lambda_i A_i$ is positive definite. The sum $\sum_i \lambda_i A_i$ and multiplication $\Pi_i \lambda_i A_i$ are positive definite.

According to Lemma 1 observed by [32] and [50], Equation 9 is positive definite. Therefore, the SVM-CGAK is a convex optimization problem which can be efficiently solved by a quadratic programming algorithm. For obtaining β_{RVLBP} and β_{CNN}, we proposed to use the localized multiple kernel regression [51]. The detailed solution can be referred to [51].

4 Experiments

We firstly intensively evaluated the effect of a parameter to SVM-GAK on the HAPPEI database [15] in Section 4.2. Next, we compared SVM-GAK with sequential methods on the HAPPEI database [15] in Section 4.3. Furthermore, we evaluated the benefit of multiple kernel learning for SVM-CGAK. Finally, we compared SVM-GAK and SVM-CGAK with the state-of-the-art methods on the HAPPEI [15], MultiEmoVA [10], and GAFF [11] databases.

4.1 Experiment protocols

The experiment protocols used in this paper are described as follows:

- HAPPEI database: Following the protocol in [18], we implemented a 4-fold-cross-validation in our experiments, where 1,500 images were used for training and 500 for testing, repeating four times. The mean absolute error was used as the metric for estimating the happiness intensity of images.
- MultiEmoVA database: Following the experiment setup in [10], we divided experiments into two parts (Experiments #1 and #2) with respect to the three categories of arousal (low, medium and high) and valence (negative, neutral and positive). Experiment #1 was to conduct the experiments on each dimension separately and report the performances, respectively, while Experiment #2 was to formulate arousal-valence categories as 5-class classification task (*i.e.*, medium+negative, high+negative, medium+positive, high+positive and neutral). We used 5-fold-cross-validation protocol and reported the average recognition accuracy for Experiments #1 and #2.
- GAFF database: According to [38], we chose Train set (3630 samples) for training and Validation (2068 samples) set for testing. Recognition accuracy was used.

4.2 Parameter evaluation of SVM-GAK on the HAPPEI database

We evaluated the effect of the standard deviation σ in $\{0.1, 1, 2, 10, 100, 1000\}$ to SVM-GAK based on RVLBP/CNN. The parameter evaluation of the standard deviation σ on the HAPPEI database is illustrated in Figure 9.

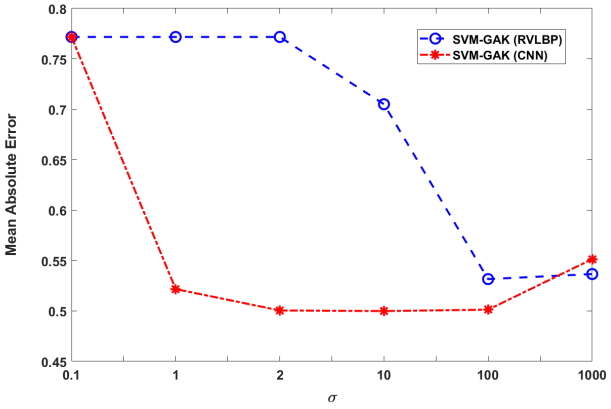


Fig. 9: Performance influence of the standard deviation σ to SVM-GAK, where SVM-GAK was separately based on RVLBP/CNN. For the purpose of simplicity, we named them as SVM-GAK (RVLBP) and SVM-GAK (CNN), respectively.

As we can see, the performance of SVM-GAK was sensitive to the change of standard deviation. In particular, the performance was obviously improved by an increasing standard deviation. A small standard deviation will make the global alignment kernel function in Equation 6 have a large variance. The performance implies that the support vector obtained in small standard deviation will influence on regression. The increasing standard deviation will lead to high bias and low variance models. It implies that the support vector does not have wide-spread influence. The

TABLE 4: Performance comparison in terms of mean absolute error amongst SVM-GAK, hidden markov model (HMM) [52], continuous conditional random field (CCRF) [53], and long short-term memory (LSTM) [54]. The fourth column means the average mean absolute error along local binary pattern, local phase quantization, and CNN for each compared algorithm. The best result is in bold. The asterisk represents the p-value is less than 0.05.

Methods	**Features**			**Average**	**p-value**		
	Local binary pattern [40]	Local phase quantization [41]	CNN		HMM	CCRF	LSTM
HMM	0.6308	0.6198	0.5562	0.6023	NaN	0.9936	0.9993
CCRF	0.5886	0.5953	0.4916	0.5585	0.0064*	NaN	0.8934
LSTM	0.5864	0.5732	0.4951	0.5516	0.00007*	0.1066	NaN
SVM-GAK	0.569	0.5674	**0.4999**	**0.5454**	0.00004*	0.0949	0.1969

promising standard deviation dropped into the range from 2 to 100.

As well, we can see the effect of two different feature descriptors to SVM-GAK. When the standard deviation reached 10 and 100, respectively, CNN and RVLBP performed considerably better than other standard deviation values. It may be explained that SVM-GAK reached a good balance between bias and variance. In the followed experiments, σ was set as 100 and 10 for RVLBP and CNN, respectively.

4.3 Performance comparison of SVM-GAK with with sequential methods on the HAPPEI database

Here, we compared SVM-GAK with hidden markov model [52], continuous conditional random field [53], and long short term memory [54]. The comparative results in terms of mean absolute error on the HAPPEI database are presented in Table 4.

As seen from Table 4, hidden markov model obtained the result of 0.6023 in terms of mean absolute error. The result indicated hidden markov model cannot better describe the relationship amongst multiple people. Next, the continuous conditional random field outperformed hidden markov model and long short term memory, as the continuous conditional random field was suitable to model the relationship between faces and intensity [53]. Lastly, according to average mean absolute error over all the three features, the SVM-GAK had a considerable performance. It implies the intensity of happiness estimated by SVM-GAK can be closed to the human annotation.

We performed a statistical analysis for a pair of two different algorithms including HMM vs. CCRF, HMM vs. LSTM, CCRF vs. HMM, CCRF vs. LSTM, LSTM vs. HMM, LSTM vs. CCRF, SVM-GAK vs. HMM, SVM-GAK vs. CCRF, and SVM-GAK vs. LSTM. We assumed the null hypothesis that the data from hidden markov, continuous conditional random field, long short term memory, and SVM-GAK come from normal distribution with the same variance. The two-sample F-test was used to judge equal variances. The corresponding results indicated that the all data have equal variances. Subsequently, two-sample left tailed t-test was implemented to analyze the algorithm pair. The null hypothesis for the pair of methods can be described as: 'for Algorithm #1 vs. Algorithm #2, the mean absolute error of Algorithm #1 is more than Algorithm #2'. The analysis concerned whether SVM-GAK is significantly better than hidden markov model, continuous conditional random field, and long short term memory. These statistical results of Table 4 indicate that SVM-GAK achieved significant improvement compared with hidden markov model, but not significant improvement compared with continuous conditional random field and long short term memory. It firstly indicates that global alignment kernel can better model the distance than hidden markov model. It also implies that random field or the deep net could be considered as a potentially explored method for group-level happiness intensity estimation in future. According to the performance comparison in terms of mean absolute error, our method had competitiveness to continuous conditional random field and long short term memory. Overall, compared with hidden markov model, continuous conditional random field, and long short term memory, SVM-GAK had considerably competitive performance.

4.4 Comparison of SVM-CGAK with decision-level and feature concatenation fusion methods on HAPPEI database

We evaluated the performance of SVM-CGAK based the previously well-designed σ. In order to show the ability of the optimal weights learning method, we fixed the weights for the two kernels as 1 in Equation 9. SVM-CGAK obtained the results of 0.5082 and 0.4920 in terms of mean absolute error by using same weights and the optimal weights learning method, respectively. This increase was due to the learned weights for the two kernels better extracted the importance of the two kernels than by directly assigning equal weights. It is concluded that the MKL strategy estimated the optimal weights of the basis kernels through optimizing a parametric function of the kernel weights.

TABLE 5: Performance comparison of 'SVM-GAK-FC', 'SVM-GAK-DC' and SVM-CGAK on the HAPPEI database, where the mean absolute error was used as a performance metric.

Methods	Decision-level	MAE
SVM-GAK-FC	-	0.5082
SVM-GAK-DC	Summation	0.5066
SVM-GAK-DC	Weighted summation	0.5069
SVM-CGAK	-	0.4920

To justify that the proposed SVM-CGAK works consistently well, we compared SVM-CGAK with SVM-GAK based on the feature concatenation method, namely, 'SVM-GAK-FC' and SVM-GAK based on decision-level, namely, 'SVM-GAK-DC'. Both methods can be referred to Supplementary Material - C. For 'SVM-GAK-FC', we set σ as 10 and used Squared Euclidean for GAK, while for 'SVM-GAK-DC', σ was set as 100 and 10 for RVLBP and deep CNN, respectively. The comparative results are presented in Table 5. It is seen that SVM-GAK-FC obtained the result of 0.5082 in terms of mean absolute error, while SVM-GAK-DC based on summation and weighted summation decision-level rules obtained the results of 0.5066 and 0.5069 in terms of mean absolute error, respectively.

It is observed that SVM-CGAK outperformed SVM-GAK-FC. It may be explained as follows: (1) feature concatenation may result in a feature vector with very large dimensionality leading to the 'curse of dimensionality' problem, and (2) the concatenated features may be incompatible to a distance metric. Combined global alignment kernels can provide more efficient dimensionality reduction for SVM-GAK than the feature concatenation method. Additionally, they can consider more complementary information of two features than the feature concatenation method. On the other hand, SVM-CGAK achieved better performance than SVM-GAK-DC. For SVM-GAK-DC, the poor performance may have been caused by the assumption of feature distribution independent for classifier fusion. Through these comparisons, SVM-CGAK overcame the SVM-GAK-FC and SVM-GAK-DC.

4.5 Evaluation of SVM-CGAK on the HAPPEI, MultiEmoVA and GAFF databases

Based on the well-designed parameters in the previous sections, we evaluated the performance of SVM-CGAK on the HAPPEI [15], MultiEmoVA [10], and GAFF [11] databases for group-level happiness intensity estimation, group-level arousal and valence prediction, and group-level facial expression recognition, respectively.

4.5.1 Group-level happiness intensity estimation

We compared SVM-CGAK with weighted group expression model ($GEM_{weighted}$) [18], latent dirichlet allocation based group expression model (GEM_{LDA}) [18], continuous conditional random field based group expression model (GEM_{CCRF}) [18], and Information aggregation on the face [26] for group-level happiness intensity estimation as described as follows:

- $GEM_{weighted}$: The group expression model was defined as the weighted average of estimated happiness intensities of all faces.
- GEM_{LDA}: Topic modeling and manually defined attributes were proposed to combine the global and local attributes for estimating happiness intensity.
- GEM_{CCRF}: Continuous conditional random field was exploited to model the content information of the faces and the relation information between faces.
- Information aggregation: In [26], Huang *et al.* proposed an information aggregation to encode the facial regions from an image into a compact feature for an image. Specifically, they divided facial images into several blocks and extracted their corresponding features. Furthermore, they employed Gaussian mixture models to obtain K visual background probability model, where K is the number of Gaussians. For each image, the feature was obtained by stacking the first- and second-order differences between the regional features and each visual background probability model.

The comparative results in terms of mean absolute error are reported in Table 6. Compared with GEM_{CCRF}, SVM-CGAK promisingly increased the performance by 0.0372 in terms of mean absolute error. Different from group expression models, SVM-CGAK obtained the benefit of combining multiple features and the advantage of using adaptive weights. On the other hand, SVM-CGAK borrowed the advantage of global alignment kernel. That is, the global alignment kernel can better describe the distance of two images than the existing group expression models methods.

TABLE 6: Comparative results of the state-of-the-art algorithms and our proposed methods on the HAPPEI database, where results of compared algorithms are directly from [18], [26]. The bold number is the best performance.

Methods	Mean absolute error
$GEM_{weighted}$ [18]	0.5469
GEM_{LDA} [18]	0.5407
GEM_{CCRF} [18]	0.5292
Information aggregation [26]	0.5187
SVM-GAK (RVLBP)	0.5316
SVM-GAK (CNN)	0.4999
SVM-CGAK	**0.4920**

It is seen that Information aggregation method obtained the best result amongst the state-of-the-art methods, which was 0.5187, while our method SVM-CGAK achieved the result of 0.4920 in terms of mean absolute error. For group-level happiness intensity estimation, SVM-CGAK considerably obtained the increasing performance by 0.0267 in terms of mean absolute error. It is glad to see that SVM-CGAK obtained better performance than Information aggregation. The promising performance improved by SVM-CGAK is explained as follows: SVM-CGAK can better reserve the label information for classification. However, information aggregation method ignored the label information in the encoding method and lacked of the discriminative information. Therefore, the comparative results presented in Table 6 indicate that SVM-CGAK performed promisingly better than the state-of-the-art methods. Moreover, we went ahead by comparing SVM-GAK with SVM based on sole global alignment kernel, which was represented by SVM-GAK (RVLBP) and SVM-GAK (CNN). It is seen that SVM-GAK (RVLBP) and SVM-GAK (CNN) obtained the results of 0.5316 and 0.4999 in terms of mean absolute error. It is seen that SVM-CGAK outperformed SVM based on sole global alignment kernel. The comparative results indicate that combining two global alignment kernels can considerably improve the performance compared with the sole kernel. According to intensive comparisons on the HAPPEI database, SVM-CGAK achieved considerable performance for group-level happiness intensity estimation.

4.5.2 Group-level arousal and valence prediction

In the MultiEmoVA database, we compared our methods with the baseline result presented by Mou *et al.* [10]. In [10], Mou *et al.* divided images into 3 groups based on the number of faces in each image as follows: 2 faces, 3 faces, and 4+ faces. Then, they extracted features from face, body, and context. For face, geometric feature, local quantized zernike moments and global quantized zernike moments are used. For body information, one-level pyramid histogram of oriented gradients is computed on the four equally divided sub-regions of the whole upper-body region. For context, its feature was formulated by using the relative relationship between multiple people. Finally, they concatenated multiple features from face, body, and context into one feature vector.

The comparative results on the MultiEmoVA database are presented in Table 7, where we straightforwardly reported the baseline results of Mou *et al.* [10]'s paper. In Experiment #1, according to classification accuracy, it is seen that SVM-CGAK achieved the significantly results on valence-level and arousal-

TABLE 7: Experimental results in terms of classification accuracy (%) for Experiments #1 and #2 on the MultiEmoVA database, where the results of Face, Face + Context, Face + Body and Face + Body + Context are directly extracted from [10]. The bold number means the best result in each experiment setup.

Methods	Experiment #1		Experiment #2
	Valence	Arousal	5-class
Face	48	52	33.15
Face + Context	50	53	38.70
Face + Body	53	50	39.96
Face + Body + Context	54	51	35.96
SVM-GAK (RVLBP)	58.06	57.61	49.20
SVM-GAK (CNN)	62.02	56.59	48
SVM-CGAK	**63.78**	**63.38**	**54.40**

TABLE 8: Algorithm comparison of our proposed methods with the baseline algorithm and several state-of-the-art methods on the GAFF database. The comparison results are derived from [11], [22], [23], [25], [55]. Methods in Lines 1-10 represent sole feature is used, while those in Lines 11-14 mean the multi-modal for group-level facial expression recognition. The bold number means the best result in the state-of-the-art methods and our proposed methods.

Method	**Recognition rate**
baseline [11]	52.97
VGG-face [55]	65.41
VGG-16 [55]	64.11
Resnet-50 [55]	62.65
Xception [55]	60.18
Facial emotion CNN [25]	69.97
VGG-19 scene [25]	67.2
Face-pretrained CNN [23]	60
InceptionV3-FC [23]	63.19
VGG16-FC [23]	66.30
Fusion of Scene and VGG-face [22]	65.0
Ensemble of classifiers [55]	66.51
Face-pretrained CNN + InceptionV3-FC [23]	70.09
Face-pretrained CNN + VGG16-FC [23]	**72.38**
SVM-GAK (RVLBP)	67.32
SVM-GAK (CNN)	70.67
SVM-CGAK	**72.17**

level predictions compared with the baseline algorithms (Face, Face+Context, Face+Body, Face+Body+Context). On the other hand, in Experiment #2, our proposed methods had considerable performance on 5-class classification task. Furthermore, in valence, arousal dimensions of Experiments #1 and 5-class classification of #2, SVM-CGAK increased the performance in terms of classification accuracy from the best result amongst baseline algorithms (54%, 53%, and 39.96%) to 63.78%, 63.38%, and 54.40%, respectively. We can see that SVM-CGAK obtained considerably results in both experiment protocols.

In Experiment #1 and Experiment #2, it is seen that our proposed method based on the face information had significant improvement on the performance compared with the work (Face) of [10]. In their work, Mou *et al.* [10] got the results of 48%, 52%, and 33.15% in terms of classification for valence prediction, arousal prediction, and 5-class prediction, respectively, but our proposed method SVM-CGAK obtained the classification accuracy of 63.78%, 63.38%, and 54.40% for valence prediction, arousal prediction, and 5-class prediction, respectively. The increased classification accuracy reached 15.78%, 11.38%, and 21.25% for valence prediction, arousal prediction, and 5-class prediction, respectively.

Additionally, our proposed methods based on face information obtained the best results compared with multi-modal (*i.e.*, Face + Body + Context) of [10]. It indicates that SVM-CGAK can only use face information to achieve promising performance without adding more information from body and context. It also implies that SVM-CGAK may be further improved by considering more multi-modal information in future. Furthermore, the method of Mou *et al.* [10] needed to build three separate systems to predict group-level valence and arousal, because they re-organized all images into three cases (*i.e.*, 2 faces, 3 faces, and 4+faces). This kind of system lacked of flexibility and robustness in the real-world application. Conversely, our proposed method can flexibly predict group-level valence and arousal without setting up three separate cases.

4.5.3 Group-level facial expression recognition

In the GAFF database, we compared our method with several comparative algorithms [11], [22], [23], [25], [55] in Table 8. The comparative algorithms were briefly described as follows:

- Dhall *et al.* [11] used Census transform histogram descriptor to extract features from 4×4 non-overlapped blocks of the images, and then used SVM with a non-linear Chi-square kernel to train the classification model.

- Balaji and Oruganti [22] combined features from face and scene information for group-level emotion recognition. Here, we named it as "Fusion of Scene and VGG-face".
- Abbas and Chalup [23] combined image context and facial information for group-level emotion recognition. They induced three solo models ("Face-pretrained CNN", "InceptionV3-FC", "VGG16-FC") and two multi-modals ("Face-pretrained CNN + InceptionV3-FC" and "Face-pretrained CNN + VGG16-FC").
- Tan *et al.* [25] proposed two types of CNNs, namely, individual facial emotion CNN ("Facial emotion CNN"), and global image based CNN ("VGG-19 scene").
- Rassadin *et al.* [55] extracted feature vectors of detected faces using the convolutional neural network (CNN) trained for face identification task. In the final pipeline an ensemble of random forest classifiers on four features from VGG-face, VGG-16, ResNet-50, and Xception on the detected faces, was learned to predict emotion score using available training set. Here, we named their proposed methods as "VGG-face", "VGG-16", "Resnet-50", "Xception", and "Ensemble of classifiers", respectively.

It is noted that the results are directly comparable due to the same experiment setups. As seen from Table 8, SVM-GAK achieved the recognition rate of 67.32% and 70.67%, respectively, when we used RVLBP and CNN as feature descriptor, respectively. Compared the sole-model approaches listed in the lines 1-10, when RVLBP was fed into SVM-GAK, SVM-GAK outperformed most of the sole-model methods except Facial emotion CNN [25]. However, the gap of performance between SVM-GAK and Facial emotion CNN was not too much. This gap may be caused by the RVLBP, since RVLBP requested the strict face alignment. When we used CNN as feature descriptor for SVM-GAK, the performance is boosted to 70.67% in terms of recognition rate. It is also seen that SVM-GAK works better than Facial emotion CNN [25]. The performance difference between RVLBP and CNN confirms that the SVM-GAK was affected by different feature descriptor. It also implies that the feature obtained by using deep learning networks may be more suitable to SVM-GAK than hand-crafted features.

Comparing with multi-modal approaches listed in the lines 11-14, SVM-GAK can as well obtain the promising results. SVM-CGAK achieved higher recognition rate (72.17%) than Fusion of Scene and VGG-face (65.0%) [22], Face-pretrained CNN + InceptionV3-FC (70.09%) [55] and Ensemble of classifier (66.51%) [55], while SVM-CGAK obtained a little worse performance comparing with hybrid network (72.38%) [55]. Different from the hybrid network [55], they combined face and scene information. However, SVM-CGAK exploited the information from face. From the perspective of sole-model, without scene information, Face-pretrained CNN of [55] only obtained the recognition rate of 60% in face information, while SVM-GAK based on CNN and SVM-CGAK obtained 70.67% and 72.17% in terms of recognition rate. The comparative results demonstrate that SVM-GAK and SVM-CGAK still lead comparative performance to group-level facial expression recognition. On the other hand, our previous work in [26] demonstrated that adding scene information can significantly improve the performance in group-level facial expression recognition. We believed that adding more information from multi-modal would make more benefit and improvement to SVM-CGAK.

5 Conclusion

To advance the research in affective computing, it is important to understand the affect exhibited by a group of people in images. In this paper, we proposed a new simple but effective method for analyzing group-level emotion. First, we proposed to use a global alignment kernel based on the efficient data structure as a novel metric, which can explicitly measure the distance between two images. Furthermore, based on the global alignment kernel, we proposed support vector machine learning with the combined global alignment kernels, namely, SVM-CGAK, for group-level emotion recognition. The combined global alignment kernels exploited the low-level and high-level facial expression representations. Additionally, it borrowed the benefit of multiple kernel learning, which can obtain the optimal weights for combining two global alignment kernels. The optimal learned weights and multiple feature descriptors can make SVM-CGAK become more robust to the challenging environment in group-level emotion recognition.

It is observed that SVM-CGAK avoided from the heavy computation compared with group expression model and multi-modal frames, as it only contained one parameter. We throughly investigated the influence of the parameter σ to the global alignment kernel on the HAPPEI database. Two examples given in Figure 7 further confirmed that the global alignment kernel can better describe the distance between group-level images than averaging method. Based on the optimal designed σ, experiment results sufficiently demonstrated that SVM-CGAK was an efficient and effective way to estimate the happiness intensity of a group of people. In order to see the generalization ability of SVM-CGAK, we conducted three experiments on group-level happiness intensity estimation, group-level valence and arousal prediction, and group-level facial expression recognition. Compared with the state-of-the-art methods, SVM-CGAK surpassed most of the state-of-the-art methods on group-level emotion recognition.

Although group-level emotion recognition has received increasing attention in the computer vision community, as well, our proposed approaches achieved the considerable results in group-level emotion recognition, it still exists some improvements need to be discussed in future. Recently, Keyton and Heylen in [56] stated that the interaction of computer science and social science "will benefit when interdisciplinary collaborations make important contributions to both". It is recommended that we can collaborate social scientists on group-level emotion recognition in three following issues:

- Data collection: So far, the publicly available group-level emotion databases collected the images from the website and multimedia, lacking of the dynamic information of group-based emotion. Smith and Mackie in [57] emphasized over-time variability of group emotions is meaningful, as "people may react differently toward an outgroup member depending on their current emotion state" (p.2). As well, Pantic and Patras's research in [58] found that dynamic video contains more discriminative information to judge the change of emotion. Thus, through dynamics process, it can be a better way to observe the relationship and emotion change between member and team [57], [58]. Additionally, the current databases seriously lacks of the context or environment in which the team interacts, such as collecting from a sufficiently large number of naturalistic groups, and so on.

- Human annotation: Currently, the researchers of these group-level databases obtained annotation by asking independent observers to rate the images. It lacks of subjective evaluations, such as self-report measurement, is ignored. This may be caused by the data collection and difficulty obtaining the self-report measurement. However, without the solid annotation, it will make the false direction to the automatic group-level emotion recognition. Based on the solid annotation, computer scientists can develop high-level methods for analyzing data used by social scientists.
- More general emotion categories: The publicly available databases mostly focus on the basic emotion categories [11], [15], but ignore theorem supported by psychologist. In fact, group emotion is commonly defined as the moods, emotions, and dispositional affects of a group of people [5]. It is important to extend the basic emotion categories into more general emotion classes.

In our future work, collaborating with social scientist, we will consider more natural group emotion databases, including dynamic information and solid label annotation. Additionally, the workflow in social science and computer science exists some difference in analyzing group interactions [59]. We will consider the difference on workflow between social science and computer science for deeply analyzing group-level emotion, not only focusing on technical level. Moreover, we will consider an end-to-end neural network and multi-modal information for SVM-CGAK to obtain more effective performance in group-level emotion recognition. Specifically, as collaborating with social scientists, we hope that automatic group-level emotion recognition will assist researchers more efficient in analyzing the development of team/group in natural environment, get the win-win benefit to both communities, and bridge the gap between both fields.

REFERENCES

[1] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2852–2861.

[2] M. Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Image and Vision Computing*, vol. 44, no. 3, pp. 572–587, 2011.

[3] L. Rice, C. Wall, A. Fogel, and F. Shic, "Computer-assisted face processing instruction improves emotion recognition, mentalizing, and social skills in students with ASD," *Journal of Autism and Developmental Disorders*, vol. 45, no. 7, pp. 2176–2186, 2015.

[4] J. Levine and R. Moreland, "Progress in small group research," *Annual Review of Psychology*, vol. 41, pp. 585–634, 1990.

[5] S. Barsäde and D. Gibson, "Group emotion: A view from top and bottom," *Research on Managing in Group and Teams*, vol. 1, pp. 81–102, 1998.

[6] J. Kelly and S. Barsäde, "Mood and emotions in small groups and work teams," *Organizational behavior and human decision processes*, vol. 86, no. 1, pp. 99–130, 2001.

[7] R. Reiter-Palmon, T. Sinha, J. Gevers, J. Odobez, and G. Volpe, "Theories and models of team and groups," *Small Group Research*, vol. 48, no. 5, pp. 544–567, 2017.

[8] J. Yang and K. Mossholder, "Decoupling task and relationship conflict: the role of intragroup emotional processing," *Journal of Organizational Behavior*, vol. 25, no. 5, pp. 589–605, 2004.

[9] J. Penalver, M. Salanova, I. Martinez, and W. Schaufeli, "Happy-productive groups: how positive affect links to performance through social resources," *The Journal of Positive Psychology*, vol. 14, no. 3, pp. 377–392, 2019.

[10] W. Mou, O. Celiktutan, and H. Gunes, "Group-level arousal and valence recognition from static images: face, body and context," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2015, pp. 1–6.

[11] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: analysing the affect of a group of people in images," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015, pp. 1–6.

[12] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, "Finding happiest moments in a social context," in *Asian Conference on Computer Vision*, 2013, pp. 613–626.

[13] J. Malmberg, S. Järvelä, J. Holappa, E. Haataja, X. Huang, and A. Siipo, "Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning?" *Computers in Human Behavior*, vol. -, 2018.

[14] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *IEEE International Conference on Computer Vision*, 2011, pp. 2106–2112.

[15] A. Dhall, R. Goecke, and T. Gedeon, "Automatic group happiness intensity analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 13–26, 2015.

[16] J. Hernandez, M. Hoque, W. Drevo, and R. Picard, "Mood meter: counting smiles in the wild," in *ACM Conference on Ubiquitous Computing*, 2012, pp. 301–310.

[17] A. Gallagher and T. Chen, "Understanding images of groups of people," in *IEEE International Conference on Computer Vision*, 2009, pp. 256–263.

[18] X. Huang, A. Dhall, G. Zhao, R. Goecke, and M. Pietikäinen, "Riesz-based volume local binary pattern and a novel group expression model for group happiness intensity analysis," in *BMVC*, 2015, pp. 1–9.

[19] A. Cerekovic, "A deep look into group happiness prediction from images," in *ICMI*, 2016, pp. 437–444.

[20] J. Li, S. Roy, J. Feng, and T. Sim, "Happiness level prediction with sequential inputs via multiple regressions," in *ICMI*, 2016, pp. 487–493.

[21] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu, "LSTM for dynamic emotion and group emotion recognition in the wild," in *ICMI*, 2016, pp. 451–457.

[22] B. Balaji and V. Oruganti, "Multi-level feature fusion for group-level emotion recognition," in *International Conference on Multimodal Interaction*, 2017, pp. 583–586.

[23] A. Abbas and S. Chalup, "Group emotion recognition in the wild by combining deep neural networks for facial expression classification and scene-context analysis," in *International Conference on Multimodal Interaction*, 2017, pp. 561–568.

[24] A. Dhall and R. Goecke, "Group expression intensity estimation in videos via gaussian processes," in *IEEE International Conference on Pattern Recognition*, 2012, pp. 3525–3528.

[25] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao, "Group-level emotion recognition with individual facial emotion CNNs and global images based CNNs," in *International Conference on Multimodal Interaction*, 2017, pp. 549–552.

[26] X. Huang, A. Dhall, R. Goecke, M. Pietikäinen, and G. Zhao, "Multimodal framework for analyzing the affect of a group of people," *IEEE Transactions on Multmedia*, vol. 20, no. 10, pp. 2706–2721, 2018.

[27] H. Shimodaira, K. Noma, M. Nakai, and S. Sagayama, "Dynamic time-alignment kernel in support vector machine," in *NIPS*, 2002, pp. 1–8.

[28] C. Bahlmann, B. Haasdonk, and H. Burkhardt, "On-line handwriting recognition with support vector machines: A kernel approach," in *Frontiers in Handwriting Recognition*, 2002, pp. 49–54.

[29] A. Gaidon, Z. Harchaoui, and C. Schmid, "A time series kernel for action recognition," in *BMVC*, 2011, pp. 1–11.

[30] L. Brun, G. Percannella, A. Saggese, and M. Vento, "Action recognition by using kernels on aclets sequences," *Computer Vision and Image Understanding*, vol. 144, pp. 3–13, 2016.

[31] J. Deng and C. Leung, "Time warping for music retrieval using time series modeling for music emotions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 6, no. 2, pp. 137–151, 2015.

[32] M. Cuturi, J. Vert, O. Birkenes, and T. Matsui, "A kernel for time series based on global alignments," in *ICASSP*, 2007, pp. 413–416.

[33] M. Cuturi, "Fast global alignment kernels," in *ICML*, 2011, pp. 1–8.

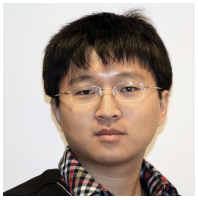
[34] A. Lorincz, L. Jeni, Z. Szabo, J. Cohn, and T. Kanade, "Emotional expression classification using time-series kernels," in *CVPR Workshop*, 2013, pp. 889–895.

[35] ——, "Spatio-temporal event classification using series kernel based structured sparsity," in *ECCV*, 2014, pp. 135–150.

[36] P. Pavlidis, J. Weston, J. Cai, and W. Grundy, "Gene functional classification from heterogeneous data," in *International Conference on Computational Molecular Biology*, 2001, pp. 242–248.

[37] A. Ben-Hur and W. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, pp. 38–46, May 2005.

[38] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedon, "From individual to group-level emotion recognition: Emotiw 5.0," in *ACM ICMI*, 2017, pp. 524–528.
[39] R. Prim, "Shortest connection networks and some generalizations," *Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957.
[40] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary pattern," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
[41] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proceeding of Image and Signal Processing*, 2008, pp. 236–243.
[42] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video," in *BMVC*, 2006, pp. 899–908.
[43] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 38–50, 2018.
[44] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *International Conference on Multimodal Interaction*, 2013, pp. 517–524.
[45] H. Ng, V. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *ICMI*, 2015, pp. 443–449.
[46] A. Lopes, E. Aguiar, A. Souza, and T. Oliviera-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.
[47] M. Gönen and E. Alpaydn, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.
[48] ——, "Localized algorithms for multiple kernel learning," *Pattern Recognition*, vol. 46, pp. 795–807, 2013.
[49] B. McCune and J. Grace, *Analysis of Ecological Communities*. MjM Software Design, 2002.
[50] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.
[51] M. Gönen and E. Alpaydin, "Localized multiple kernel regression," in *International Conference on Pattern Recognition*, 2010, pp. 1425–1428.
[52] J. Santarcangelo and X. Zhang, "Arousal content representation of sport videos using dynamic prediction hidden markov models," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 1049–1053.
[53] V. Imbrsaitė, T. Baltrušaitis, and P. Robinson, "Emotion tracking in music using continuous conditional random fields and relative feature representation," in *IEEE International Conference on Multimedia and Expo Workshops*, 2013, pp. 1–6.
[54] J. Wei, E. Pei, D. Jiang, H. Sahli, L. Xie, and Z. Fu, "Multimodal continuous affect recognition based on lstm and multiple kernel learning," in *Asian-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2014, pp. 1–4.
[55] A. Rassadin, A. Gruzdev, and A. Savchenko, "Group-level emotion recognition using transfer learning from face identification," in *International Conference on Multimodal Interaction*, 2017, pp. 544–548.
[56] J. Keyton and D. Heylen, "Pushing interdisciplinary in the study of groups and teams," *Small Group Research*, vol. 48, pp. 621–630, 2017.
[57] E. Smith and D. Mackie, "Dynamics of group-based emotions: insights from intergroup emotions theory," *Emotion Review*, vol. 7, pp. 349–354, 2015.
[58] M. Pantic and I. 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*, vol. 36, pp. 433–449, 2006.
[59] J. Allen, C. Fisher, M. Chetouani, M. Chiu, H. Gunes, M. Mehu, and H. Hung, "Comparing social science and computer science workflow processes for studying group interactions," *Small Group Research*, vol. 48, pp. 568–590, 2017.



Xiaohua Huang received the B.S. degree in communication engineering from Huaqiao University, Quanzhou, China in 2006. He received his Ph.D degree in Computer Science and Engineering from University of Oulu, Oulu, Finland in 2014. He was a research assistant in Southeast University since 2006. He had been a scientist researcher in the Center for Machine Vision and Signal Analysis at University of Oulu in 2015-2018. He is now independent researcher at University of Oulu. He has authored or co-authored more than 20 papers in journals and conferences, and has served as a reviewer for journals and conferences. His current research interests include facial expression recognition, micro-expression analysis, group-level emotion recognition, multi-modal emotion recognition and texture classification. He is a member of the IEEE.

Abhinav Dhall received the PhD degree in computer science from the Australian National University in 2014. He is currently an Assistant Professor of Computer Science & Engineering at the Indian Institute of Technology, Ropar. Prior to this position he did postdocs at the University of Waterloo and the University of Canberra. He was also an adjunct research fellow at the Australian National University. He was awarded the Best Doctoral Paper Award at ACM International Conference on Multimodal Interaction 2013, Best Student Paper Honourable mention at IEEE International Conference on Automatic Face and Gesture Recognition 2013 and Best Paper Nomination at IEEE International Conference on Multimedia and Expo 2012. His research interests are in computer vision for affective computing and assistive technology. He is a member of the IEEE.

Roland Goecke is Professor of Affective Computing at the University of Canberra and an adjunct senior research fellow at the Australian National University. He is the Director of the Human-Centred Technology Research Centre and leads the Vision and Sensing Group, University of Canberra. He received his Masters degree in Computer Science from the University of Rostock, Germany, in 1998 and his PhD in Computer Science from the Australian National University, Canberra, Australia, in 2004. His research interests are in affective computing, pattern recognition, computer vision, human-computer interaction and multimodal signal processing.

Matti Pietikäinen received his Doctor of Science in Technology degree from the University of Oulu, Finland. He is currently a professor, Scientific Director of Infotech Oulu and Director of Center for Machine Vision Research at the University of Oulu. From 1980 to 1981 and from 1984 to 1985, he visited the Computer Vision Laboratory at the University of Maryland. He has made pioneering contributions, e.g. to local binary pattern (LBP) methodology, texture-based image and video analysis, and facial image analysis. He has authored over 340 refereed papers in international journals, books and conferences. He was Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and Pattern Recognition journals, and currently serves as Associate Editor of Image and Vision Computing and IEEE Transactions on Forensics and Security journals. He was President of the Pattern Recognition Society of Finland from 1989 to 1992, and was named its Honorary Member in 2014. From 1989 to 2007 he served as Member of the Governing Board of International Association for Pattern Recognition (IAPR), and became one of the founding fellows of the IAPR in 1994. He is IEEE Fellow for contributions to texture and facial image analysis for machine vision. In 2014, his research on LBP-based face description was awarded the Koenderink Prize for Fundamental Contributions in Computer Vision.

Guoying Zhao received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently an Associate Professor with the Center for Machine Vision Research, University of Oulu, Finland, where she has been a researcher since 2005. In 2011, she was selected to the highly competitive Academy Research Fellow position. She has authored or co-authored more than 110 papers in journals and conferences, and has served as a reviewer for many journals and conferences. She has lectured tutorials at ICPR 2006, ICCV 2009, and SCIA 2013, and authored/edited three books and two special issues in journals. Dr. Zhao was a Co-Chair of the International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA) at ECCV2008, ICCV2009, and CVPR2011, ECCV2014 workshop on Spontaneous Facial Behavior Analysis: Long term continuous analysis of facial expressions and micro-expressions and ACCV 2014 workshop on RoLoD: Robust local descriptors for computer vision and of a special session for IEEE International Conference on Automatic Face and Gesture Recognition 2013 (FG13). She is IEEE Senior Member. Her current research interests include image and video descriptors, gait analysis, dynamic-texture recognition, facial-expression recognition, human motion analysis, and person identification.