Riesz-based Volume Local Binary Pattern and A Novel Group Expression Model for Group Happiness Intensity Analysis

Xiaohua Huang[1]
huang.xiaohua@ee.oulu.fi
Abhinav Dhall[2,3]
abhinav.dhall@canberra.edu.au
Guoying Zhao[1]
gyzhao@ee.oulu.fi
Roland Goecke[2,3]
roland.goecke@ieee.org
Matti Pietikäinen[1]
mkp@ee.oulu.fi

[1] Center for Machine Vision Research
Department of Computer Science and Engineering
University of Oulu, Finland

[2] Human-Centred Technology Research Centre
University of Canberra, Australia

[3] IHCC Group
Research School of Computer Science
Australian National University, Australia

Abstract

Automatic emotion analysis and understanding has received much attention over the years in affective computing. Recently, there are increasing interests in inferring the emotional intensity of a group of people. For group emotional intensity analysis, feature extraction and group expression model are two critical issues. In this paper, we propose a new method to estimate the happiness intensity of a group of people in an image. Firstly, we combine the Riesz transform and the local binary pattern descriptor, named Riesz-based volume local binary pattern, which considers neighbouring changes not only in the spatial domain of a face but also along the different Riesz faces. Secondly, we exploit the continuous conditional random fields for constructing a new group expression model, which considers global and local attributes. Intensive experiments are performed on three challenging facial expression databases to evaluate the novel feature. Furthermore, experiments are conducted on the HAPPEI database to evaluate the new group expression model with the new feature. Our experimental results demonstrate the promising performance for group happiness intensity analysis.

1 Introduction

In recent years, millions of images and videos have been uploaded on the Internet (*e.g.* in YouTube and Flickr), enabling us to explore images from a social event, such as a family party. However, until recently, relatively little research has examined group emotion in an image. To advance affective computing research, it is indeed of interest to understand and model the affect exhibited by a group of people in images.

Recently, Dhall *et al.* [6] investigated the problem of happiness intensity analysis of a group of people in an image using facial expression analysis. This is one of the earliest

Pages 34.1-34.13
DOI: https://dx.doi.org/10.5244/C.29.34

attempts investigating the perception of emotion of a group of people in images. In [7], based on a survey conducted, the same authors argued that the perceived mood of a group of people in images is based on global and local attributes. For estimating happiness intensity, a group expression model (**GEM**) was proposed based on topic modelling and manually defined attributes for combining the global and local attributes. The Histogram of Oriented Gradient (**HOG**) was employed as a face descriptor for local attribute. However, in our observation in this paper, the group expression recognition performance is affected by the choice of facial expression descriptors. Additionally, a GEM based on topic modelling is limited by the visual vocabulary; and many parameters need to be considered during the training of the topic model and the construction of the dictionary. Thus, it is necessary to exploit more robust facial expression features and an efficient GEM for group happiness intensity analysis, which can model the structure of a group.

Facial expression descriptors can be broadly defined as geometric or appearance based. Geometric-based features represent the face geometry, such as the shapes and locations of facial landmarks. Instead, appearance-based features describe the skin texture of faces. However, geometric-based features are sometimes sensitive to illumination variation, pose change and error in fiducial points detection. Appearance features, as opposed to geometric features, have certain advantages in that they are more stable to such global changes as illumination and inaccurate alignment. Gabor wavelets [21] are used to capture the local structure corresponding to spatial frequency, spatial localisation and orientation selective. They have been demonstrated to be discriminative and robust to illumination changes. Another face descriptor, namely local binary pattern (**LBP**) [23], is a simple and efficient manner to represent faces, which is also robust to global changes. Recently, Zhang *et al.* [38] combined the Gabor and LBP descriptors (**LGBP**) to improve the face recognition performance. However, Gabor filters suffer from two problems: (1) they are not optimal if the broad spectral information with maximal spatial localization need to be sought and (2) their maximum bandwidth is restricted to about one octave [33].

Recent studies [10, 17, 25, 34, 36, 37] have demonstrated the local image information can be well characterized in a unified theoretic framework, namely the Riesz transform. Felsberg and Sommer proposed to use the Riesz transform for image processing [10]. In their work, they proposed the monogenic signal based on the 1st-order Riesz transform, which extends the classical analytic signal to a 2D domain. Additionally, a monogenic signal can well address the limitations of Gabor filters (i.e., not optimal and the restricted maximum bandwidth) because of the utilization of a log-Gabor filter. Until recently, Riesz transform has attracted much interest from researchers in the field for texture classification [37] and face analysis [17, 34]. However, from the intrinsic dimension theorem [35], the 1st-order Riesz transform is designed just for an intrinsic 1D signal, but losing some important complex structures, such as corners. In order to characterize the intrinsic 2D local structures, the higher-order Riesz transforms have been developed for biometric recognition [36] and texture recognition [25]. Zhang *et al.* [36] proposed to utilize the 1st-order and 2nd-order Riesz transforms to encode the local patterns of biometric images. Thus, it begs the question if the higher-order Riesz transforms can also provide rich information for face representation. Motivated by [25, 36], we propose a new method based on the higher-order Riesz transform and local binary patterns for characterizing facial expressions.

Images from social events generally contain a group of people. The analysis of the mood of a group of people in images has various applications such as image album creation, event summarization, key-frame selection and recommendation. Social psychology studies suggest that group emotion can be conceptualised in different ways. Generally, group emotion can be represented by pairing the bottom-up and top-down approaches [2, 19].

In the bottom-up methods, the subjects' attributes are employed to infer information at the group level. Hernandez *et al.* [15] proposed a bottom-up approach, in which an average over the smiles of multiple people was used for inferring the mood of the passerby. However, in reality, perceived group mood is not an averaging model [19]. In the case of top-down techniques, external factors to the group and members are considered, *e.g.*, the effect of the scene. In an interesting top-down approach, Gallagher and Chen [12] proposed contextual features based on the group structure for computing the age and gender of individuals. Another top-down approach, Stone *et al.* [31] proposed a Conditional Random Field (**CRF**) based on social relationship modelling between Facebook contacts for the problem of face recognition.

Dhall *et al.* [6]'s GEM models are an example of a hybrid approach, which considers both face-level expressions and neighbor effect. Their conducted a survey and argued that the mood of a group of people in images is based on top-down and bottom-up components: (1) the top-down component is referred to as global attributes such as the effect on the mood of a group member due to the neighbours; (2) the bottom-up component is referred to as low-level attributes such as the contribution of an individual's mood based on their facial expressions to the overall mood of the group. A GEM based on topic modelling and manually defined attributes are presented to combine the global and local attributes.

Recently, continuous conditional random fields (**CCRF**) have been proposed to model the content information of objects as well as the relation information between objects for global ranking problem [27]. Furthermore, Imbrsaite *et al.* [18] designed CCRF to model the affect continuously for dimensional emotion tracking. They demonstrated that CCRF is more suitable to continuous output variable modelling than CRF. Inspired by [7, 27], we propose to combine top-down and bottom-up components by using CCRF for a novel GEM.

The **key contributions** of this paper are as follows: (1) A new facial expression descriptor based on the Riesz transform is developed, which is robust to real-world situations including pose change and illumination variation; (2) A novel GEM approach is presented to effectively combine global and local attributes; and (3) The combination of the new feature and GEM approach is used to infer the happiness intensity of a group of people in an image.

To explain the concepts in our approach, the paper is organised as follows: In Section 2, we introduce Riesz-based volume local binary patterns as a facial expression descriptor. In Section 3, we provide a new GEM to estimate group happiness intensity. In Section 4, we present the results of examining the proposed feature and GEM for group happiness intensity analysis. Finally, we draw our conclusions in Section 5.

2 Riesz-based volume local binary patterns

In this section, we firstly give a brief review of the higher-order Riesz transform. Subsequently, we provide details of our proposed facial expression descriptor, named Riesz-based volume local binary patterns (**RVLBP**).

2.1 Higher-Order Riesz Transform

The Riesz transform [30] is a natural generalization of the Hilbert transform. Riesz transform based image analysis has been utilized in numerous fields [17, 25, 33, 36], but prior to applying Riesz transform to image analysis, it is necessary to pre-filter the image with a suitable band-pass filter, because real images commonly contain a wide range of frequencies. So far, many candidates on the band-pass filter have been proposed in the literature. In

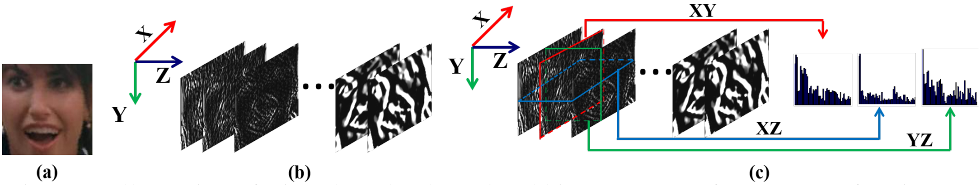

Figure 1: Illustration of Riesz-based volume local binary patterns for R_x: (a) A face image; (b) The 1st-order Riesz face R_x and its volume, where Z represents the axis of scale and orientation of Riesz transform; and (c) feature extraction for the R_x component.

this paper, we choose the commonly utilized multi-orientation and multi-scale log-Gabor filter [11], which is defined in the frequency domain as:

$$G(\omega,\theta) = exp(\frac{log(\frac{\omega}{\omega_0})^2}{2(log(\frac{\sigma_\omega}{\omega_0}))^2})exp(\frac{-(\theta-\theta_0)^2}{2\sigma_\theta^2}), \tag{1}$$

where ω_0 is the centre frequency, σ_ω is the width parameter for the frequency, θ_0 is the centre orientation, and σ_θ is the width parameter of the orientation.

Along with the log-Gabor filter, the 1st-order Riesz transform in the nD spatial domain can be expressed as:

$$R_j(X) = g(\omega,\theta) * c_n \frac{\mathbf{X}_j}{|\mathbf{X}|^{n+1}}, \tag{2}$$

where $c_n = \Gamma[(n+1)/2]/\pi^{(n+1)/2}$, $\mathbf{X} = [x_1, x_2, \ldots, x_n]$, $j = 1, \ldots, n$, and $g(\omega,\theta)$ is the spatial expression of log-Gabor $G(\omega,\theta)$.

The intrinsic dimension is used to describe a local image structure [35, 36]. In a 2D image, the local structure can be classified into numerous regions of i0D, i1D and i2D structures. For example, constant areas are of intrinsic dimension zero (i0D), while straight lines and edges are of intrinsic dimension one (i1D). According to the intrinsic dimension, for 2D images, the 1st-order Riesz transform enables the rotationally invariant analysis of the i1D structure; the 2nd-order Riesz transform can characterize i2D image structures such as corners and texture [35]. Therefore, the 1st-order and 2nd-order Riesz transforms are employed to describe the structure of a facial expression image. According to Eq. 2, in the case of a 2D image, $\mathbf{X} = (x, y)$, the 1st-order and 2nd-order Riesz transforms are expressed as:

$$h_x(\mathbf{X}) = g(\omega,\theta) * \frac{x}{2\pi|\mathbf{X}|^3}, \quad h_y(\mathbf{X}) = g(\omega,\theta) * \frac{y}{2\pi|\mathbf{X}|^3}, \tag{3}$$

$$h_{xx}(\mathbf{X}) = h_x\{h_x\}(\mathbf{X}) = h_x(\mathbf{X}) * h_x(\mathbf{X}), \tag{4}$$

$$h_{yy}(\mathbf{X}) = h_y\{h_y\}(\mathbf{X}) = h_y(\mathbf{X}) * h_y(\mathbf{X}), \tag{5}$$

$$h_{xy}(\mathbf{X}) = h_x\{h_y\}(\mathbf{X}) = h_x(\mathbf{X}) * h_y(\mathbf{X}). \tag{6}$$

In this paper, we use log-Gabor filters at three scales and four orientations, thus deriving the 1st-order and 2nd-order Riesz components by convolving face images with log-Gabor filters. For an image $I(x,y)$, we will obtain new Riesz faces for $h_x(\mathbf{X})$ as:

$$R_x = I(x,y) * [h_x(\mathbf{X})_1^1, \ \ldots, h_x(\mathbf{X})_3^4]. \tag{7}$$

For R_x, these Riesz face images can then be resembled to form the Riesz volume. Fig. 1(a-b) shows an example of a face image with its corresponding volume of Riesz face R_x. The same procedure of obtaining R_x is applied to $h_y(\mathbf{X})$, $h_{xx}(\mathbf{X})$, $h_{yy}(\mathbf{X})$ and $h_{xy}(\mathbf{X})$ for obtaining R_y, R_{xx}, R_{yy}, R_{xy}, respectively.

2.2 Riesz-based Volume Local Binary Patterns (RVLBP)

Local binary pattern has been demonstrated as a powerful and efficient local descriptor for micro-features of images in face analysis and texture classification. Recently, the combination of a Riesz-based method and LBP has been shown to be an effective way for many applications [17, 25, 34]. Moreover, Lei *et al.* [20] developed an approach to exploit the neighbouring relationship in the spatial domain and various Gabor faces. Inspired by this, we propose to explore discriminative information from the 1st-order and 2nd-order Riesz faces.

For a face image, the derived Riesz faces can be formulated as the 3D volume as illustrated in Fig. 1(b), where the three axes X, Y and Z denote width and height of face image and different types of Riesz filters, respectively. Local binary pattern from three orthogonal planes [16, 39] applies LBP separately on three orthogonal planes, which intersect in the centre pixel. All histograms can describe effectively appearance and motion from an image sequence. Motivated by this, we conduct LBP analysis on three orthogonal planes of Riesz volume, exploring not only spatial pixel correlation in a Riesz face but also the relationship among Riesz faces along frequency and orientation directions.

In this work, we first employ the LBP operator on XY, XZ and YZ of volume-based Riesz faces, respectively. Secondly, we combine the results of these planes to represents faces. For each plane, its histogram is computed as

$$H_j(l) = \sum_{x,y} L(f_j(x,y) = l), l = 0, 1, \dots, Q_j - 1, \tag{8}$$

where $L(x)$ is a logical function with $L(x) = 1$ if x is true and $L(x) = 0$ otherwise; $f_j(.)$ expresses the LBP codes in the j-th plane ($j = 0 : XY; 1 : XZ; 2 : YZ$), and Q_j is the bin number of the LBP code in the j-th plane. Finally, the histograms H_{XY}, H_{XZ} and H_{YZ} are concatenated into one feature vector H. The procedure is shown in Fig. 1(c).

Following Section 2.1, R_x, R_y, R_{xx}, R_{yy}, R_{xy} are obtained. For each component R_m ($m = \{x, y, xx, xy, yy\}$), H_m is obtained by using the above-mentioned procedure. These five histograms $H_x, H_y, H_{xx}, H_{xy}, H_{yy}$ are concatenated into one feature vector F for representing the face. This feature vector incorporates the spatial information and the co-occurrence statistics on frequency and orientation domains of Riesz transform, thus is more effective for face representation.

Our new feature contains the 1st-order and 2nd-order Riesz components, thus our features are of high dimensionality. Using high dimensional features makes the descriptor matching process slow, and always includes the risk of over-fitting. In order to obtain a more discriminant and low-dimensional facial representation, we use Locality Preserving Projection [14] to project the features to a low-dimensional space. In the case of face analysis, principal components having high variance might not necessarily contain the discriminant information, so we process all principal components using square-roots of their corresponding eigenvalues. This action can reduce the influence of leading principal components while increase the influence of trailing components. Once we can obtain feature space $\mathbf{U}$, the low-dimensional features can be formulated as $\widetilde{F} = \mathbf{U}'F$.

3 A Novel Group Expression Model

A group expression model aims to explore the relationship between faces and intensity label in a group image. Recently, continuous conditional random fields has been proposed to model the relation between objects and ranking score [27] and the affect continuously in

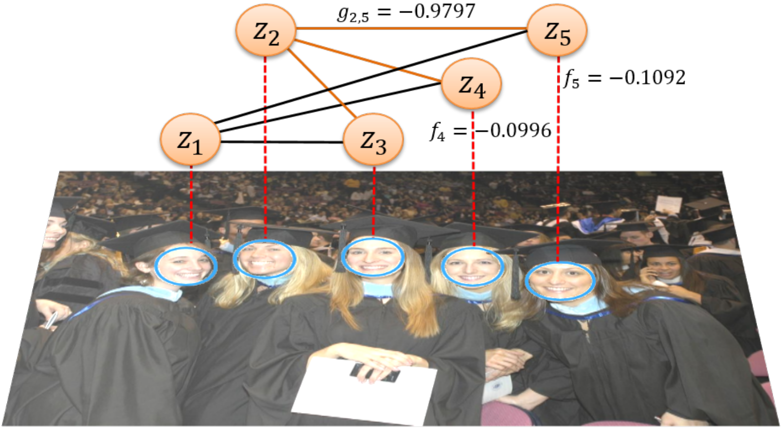

Figure 2: GEM$_{CCRF}$: The blue circle on a facial image represents the extracted content (including local and global attributes), z_i is an happiness intensity label, an edge (a solid line) between z_i and z_j, *e.g.* $g_{2,5}$, means the dependency between intensities of two faces, an edge (a dash line), *e.g.* f_5, represents the dependency of an intensity label on its content.

emotion tracking [18]. It is found that CCRF is suitable to model the relationship between faces and intensity label, since faces and intensity label can be equivalent to objects and ranking score, respectively. On the other hand, Dhall *et al.* [7] emphasised that the top-down component (i.e. global attributes) and the bottom-up component (i.e. local attributes) are very important for a group expression model. Inspired by [7, 18, 27], we formulate a new group expression model based on CCRF (GEM$_{CCRF}$) for combining global and local attributes.

According to [7], a fully connected graph $\mathcal{G} = (V,E)$ is constructed to map the global structure of faces in a group G. Here, V_i represents the i-th face in the group and an edge $E_{i,j}$ represents the link between two faces V_i and V_j. The minimal spanning tree algorithm [26] is employed to obtain the graph $\mathcal{G}$, which can provide the location and minimally connected neighbours of a face. Finally, the global attribute can be expressed by the relative size and relative distance as follows:

(1) **Relative size**: For a face V_i in the group, its size is taken as the distance between the locations l_i and r_i of the left and right eyes, $d_i = \parallel l_i - r_i \parallel$, where l_i and r_i are obtained from facial landmarks by using [9]. The relative face size S_i of V_i is given by $S_i = \frac{d_i}{d_i + \sum_{j=1}^{n} \frac{d_j}{n}}$, where n is the number of connected neighbours of V_i.

(2) **Relative distance**: Based on the nose tip locations p_i of all faces in a group G, the centroid c_g of G is computed. The relative distance δ_i of the i-th face is described as $\delta_i = \parallel p_i - c_g \parallel$, and δ_i is further normalised based on the mean relative distance.

On the other hand, a local attribute is also important for a group expression model. In this work, the local attribute contains the local features $\widetilde{F}$ for faces, for example, extracted by RVLBP (in Section 2). Global and local attributes are concatenated into $\mathcal{F} = [\widetilde{F}, S, \delta]$, providing sufficient and useful information for the CCRF.

The CCRF model is defined as a conditional probability distribution over ranking scores of objects conditioned on the objects. We suppose that a group image contains n faces, which is represented by $\mathcal{F}_i, i = 1, \ldots, n$. According to [18], we transform $\mathcal{F}_i$ into t_i by using SVR [3], which represents the relevance factor of one subject for the CCRF. Therefore, the

CCRF model for a group image is a conditional probability distribution with the probability density function:

$$P(z \mid \mathbf{T}) = \frac{\exp(\Pi)}{\int \exp(\Pi) d\mathbf{z}}, \tag{9}$$

$$\Pi = \sum_{i=1}^{n} \sum_{k=1}^{m} \mu_k f_k(z_i, \mathbf{T}_{i,k}) + \sum_{i,j} \nu g(z_i, z_j, \mathbf{T}), \tag{10}$$

where $\mathbf{T} = \{t_1, \ldots, t_n\} \in \mathcal{R}^m$ is the set of input feature vector, $\mathbf{Z} = \{z_1, \ldots, z_n\}$ is the intensity label of faces in a group image, $f(.)$ and $g(.)$ are vertex and edge features, respectively.

Following the convention of [27], we employ $f_k(z_i, \mathbf{T}_{i,k}) = -(z_i - \mathbf{T}_{i,k})^2$ for representing the dependency between $\mathbf{T}_{i,k}$ and z_i. The larger f_k is, the more dependent $\mathbf{T}_{i,k}$ and z_i are. We also used $g(z_i, z_j, \mathbf{T}) = -\frac{1}{2} W_{i,j}(z_i - z_j)^2$ for expressing the dependency between intensities z_i and z_j, where $W_{i,j}$ is the weight between two faces V_i and V_j, defined as $\exp(\frac{\|\mathcal{F}_i - \mathcal{F}_j\|}{\sigma})$. The edge feature g represents how related the intensity label of two faces V_i and V_j is. The CCRF model for group happiness intensity analysis is depicted in Fig. 2.

It is very important to estimate the parameters μ and ν in Eq. 10. Following the work of [18], we pick μ and ν that maximise the conditional log-likelihood as:

$$(\mu^*, \nu^*) = \arg\max_{\mu,\nu} \sum_{q=1}^{M} \log P(\mathbf{z}^{(q)} | \mathbf{T}^{(q)}), \tag{11}$$

where M expresses the number of group images. Since this problem is convex, the optimization procedure can be solved by using stochastic gradient ascent.

Since the CCRF model can be seen as a multivariate Gaussian, it is straightforward to infer happiness intensities $\mathbf{z}$ for all faces in a group using maximizing $P(\mathbf{z}|\mathbf{T})$. Given a group image, containing n faces, the intensity can be estimated by using $\frac{1}{n} \sum_{i=1}^{n} \arg\max_{z_i} P(z_i | \mathbf{T}_{test})$.

4 Experiment

As previously mentioned, our proposed method includes a new facial expression descriptor (RVLBP) and a group expression model (GEM$_{CCRF}$). Since the HAPPEI database [7] is the only database for group happiness intensity analysis, we apply the following strategy for analyzing our proposed methods. Firstly, we examine the performance of the RVLBP descriptor on three challenging face databases [5, 7, 32] in Section 4.1. Secondly, we conduct the experiment on the HAPPEI database [7] for evaluating GEM$_{CCRF}$ in Section 4.2.

4.1 Feature Evaluation

In order to evaluate the performance of the RVLBP descriptor, we conduct experiments on the SFEW [5], GENKI-4K [32] and HAPPEI [7] databases, for the task of facial expression recognition in the wild, smile detection and happiness intensity estimation, respectively.

(1) Based on [5], the experiments on the SFEW database follow the strictly person independent strategy protocol (i.e. the train and test set have no common subjects). An Support Vector Machine (**SVM**) with linear kernel [3] is utilized to classify seven facial expressions (*anger*, *disgust*, *fear*, *happiness*, *neutral*, *sadness* and *surprise*). The final result is the average of the two sets.

(2) For the GENKI-4K database, a 5-fold cross validation protocol is followed. We compare the performance by using the Area Under the Curve (**AUC**) and the average accuracy metrics. An SVM with linear kernel [3] is utilized for classyfying *smile* vs *no smile*.

(3) For the HAPPEI database, 2,000 images are chosen in our case, including a total of 7,490 faces. All faces are labeled into six mood intensity levels (0-5). We perform the

Table 1: Performance comparison on SFEW, GENKI-4K and HAPPEI databases, where * represents our implementation of an algorithm on the database.

Methods	**SFEW**	**GENKI-4K**		**HAPPEI**	Execution time (second)
	Accuracy	AUC	Accuracy	MAE	
LBP* [23]	21.56%	0.9132	91.35%	0.8013	0.0768
HOG* [4]	20.54%	0.8629	86.32%	0.8509	0.1735
LPQ* [24]	26.17%	0.8856	88.67%	0.7808	0.3775
CLBP* [13]	22.85%	0.9069	90.72%	0.8198	0.0368
MonoLBP* [33]	25.58%	0.8814	88.22%	0.9198	0.0898
GV-LBP-TOP* [20]	24.26%	0.8882	88.92%	0.7870	8.48
HOG+LPQ [5]	19%	-	-	-	-
AUDN [22]	26.14%	-	-	-	-
HOG+ELM [1]	-	**0.946**	88.2%	-	-
BPD [29]	-	-	89.7%	-	-
Proposed RVLBP	**29.84%**	0.9248	**92.52%**	**0.7688**	0.5639

experiments by using a 4-fold cross validation protocol, in which faces from 1,500 images are chosen as training and the rest for testing, repeating 4 times. Kernel Partial Least Square regression [28] is employed to estimate the intensity of happiness. The Mean Absolute Error (**MAE**) is calculated for all comparisons.

For localising the facial parts in the images of the three databases, we use the work in [9] to detect the nine facial landmarks, which describes the center point of the nose, the left and right corners of both eyes, the left and right corners of the nostrils, and the left and right corners of the mouth. For aligning the faces, an affine transform is applied. The faces are cropped to 128×128 pixel size. For a fair comparison, we compare RVLBP with LBP [23], HOG [4], Local Phase Quantization (LPQ) [24], Completed LBP (CLBP) [13], Monogenic LBP (MonoLBP) [33], and GV-LBP-TOP [20] on three databases. For all methods except HOG, we divide a facial image into 8×8 blocks. For LBP, CLBP, MonoLBP and GV-LBP-TOP, radius and the number of neighbours are set to 3 and 8, respectively. For GV-LBP-TOP, 40 Gabor faces are used. For HOG, pyramid level and bin count are set to 3 and 16, respectively. Tab. 1 shows the results of all approaches and the state of art on the three databases.

As seen in Tab. 1, for the SFEW database, the LPQ descriptor achieves a classification accuracy rate of 26.17% on the baseline algorithms, while our proposed method yields an accuracy of 29.84%, which is highest among all the methods. This can be explained due to the application of a higher-order Riesz transform, which increases the discriminative power of the final feature descriptor. Furthermore, we also compare the results of the state of the art [22], in which they produced 26.14% by using an AU-aware deep network. It is seen that our method outperforms AUDN.

Subsequently, for the GENKI-4K database, we can see from Tab. 1 that LBP performs better than HOG, MonoLBP and LPQ on average AUC and accuracy. Compared with LBP, RVLBP performs better by an absolute margin of 1.17% and 0.0116 for accuracy and AUC, respectively. We further compare our descriptor RVLBP with two other descriptors, the results compared in this paper are taken from [1, 29]. It can be seen that RVLP performs better in terms of accuracy while a little less than [1] in terms of AUC. Lastly, for the HAPPEI database, as seen from Tab. 1, LPQ achieves the lowest MAE of 0.7808 among the baseline methods. Comparing with LPQ, the MAE of RVLBP is decreased to 0.7688. This demonstrates that RVLBP achieves promising performance on the task of facial expression intensity

Table 2: Comparison of various GEM approaches with three features (Mean absolute error).

Models	Feature			
	HOG	LBP	LPQ	RVLBP
GEM$_{avg}$	0.6062	0.5824	0.5738	0.5622
GEM$_{w}$	0.6069	0.5695	0.5665	0.5469
GEM$_{LDA}$	0.6019	0.5542	0.5591	0.5407
GEM$_{CCRF}$	0.5815	0.5347	0.5399	**0.5292**

estimation as well.

Furthermore, we compare these facial descriptors on the basis of their computational complexity. Matlab R2013 based code is computed on an Intel i5-2400 processor at 3.10 GHz. The execution times are reported in Tab. 1. It is found that the proposed descriptor costs more than other descriptors except GV-LBP-TOP. The reason for longer execution time for RVLBP is the computation time spent in the pre-processing of an image based on the Riesz transform. However, this computation complexity for RVLBP is still acceptable for real-time applications and can be improved using trivial parallelization.

Overall, our RVLBP features result in better accuracy than LBP and its variants on three databases. We argue that this is due to the ability of the Riesz transform in providing an intrinsic dimensional structure to the features.

4.2 Comparison of GEM

It is noted that few studies have investigated group-level affect recognition. In this section, our GEM$_{CCRF}$ is evaluated on the HAPPEI database [7], in which the images are annotated with the group level mood intensity ('neutral' to 'thrilled'). In the experiments, we employ a 4-fold-cross-validation protocol, where 1,500 images are used for training and 500 for testing, repeating 4 times. For comparing our GEM model, we choose classical GEM (GEM$_{avg}$), weighted GEM (GEM$_{w}$) and Latent Dirichlet Allocation based GEM (GEM$_{LDA}$). For more details, refer to [7]. Among them, the global parameters α and β are 0.1 and 0.9, respectively. According to Tab. 1 and [7], we choose HOG, LBP, LPQ and RVLBP for all GEM models.

Tab. 2 shows the mean absolute errors of all GEM approaches. As seen from this table, RVLBP obtains promising results on all GEM models. While we use the RVLBP feature on GEM$_{avg}$, comparing with HOG, LBP and LPQ, the MAE is decreased by 0.044, 0.0202 and 0.0116, respectively. Similar results are obtained for the two other GEM models. Moreover, we find that RVLBP achieves the best performance on all GEM models, followed closely by LPQ and LBP, and more distantly by HOG. This is the same to our findings previously mentioned in Section 4.1. These results demonstrate that a suitable feature can boost the performance for all GEM models.

Based on RVLBP, we further examine the performance of various GEM models. The results are reported as 0.5622, 0.5469, 0.5407, and 0.5292 for GEM$_{avg}$, GEM$_{w}$, GEM$_{LDA}$ and GEM$_{CCRF}$, respectively. For GEM$_{w}$, comparing with GEM$_{avg}$, it is found that the global attributes can improve the performance for happiness intensity estimation. Moreover, it is found that GEM$_{LDA}$ considerably makes MAE lower than GEM$_{w}$, since GEM$_{LDA}$ further considers the local attributes in the topic model. It demonstrates that the combination of local and global attributes can improve the performance of group happiness intensity estimation.

It is observed that GEM$_{LDA}$ and our proposed model both consider the local and global attributes, but our model (GEM$_{CCRF}$) decreased the MAE to 0.5292, while GEM$_{LDA}$ to

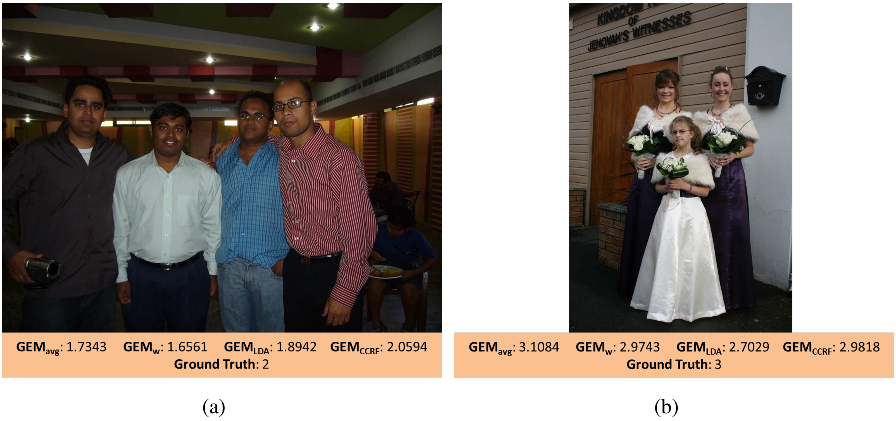
GEM$_{avg}$: 1.7343 GEM$_{w}$: 1.6561 GEM$_{LDA}$: 1.8942 GEM$_{CCRF}$: 2.0594
Ground Truth: 2
GEM$_{avg}$: 3.1084 GEM$_{w}$: 2.9743 GEM$_{LDA}$: 2.7029 GEM$_{CCRF}$: 2.9818
Ground Truth: 3
(a)
(b)

Figure 3: Two group images from the HAPPEI database, and their estimated happiness intensities using GEM$_{avg}$, GEM$_{w}$, GEM$_{LDA}$ and GEM$_{CCRF}$. For each image, the texts at the bottom of the image indicate the estimated happiness intensity results and the ground truth, respectively

0.5407. It shows that our method outperforms GEM$_{LDA}$. This is explained by GEM$_{LDA}$ using k-means to produce the bag-of-word model, which can reduce the contribution of local features to GEM; while our model emphasizes the role of local attributes. The results demonstrate that using CCRF is an effective way to build the relationship among subjects in a group.

5 Conclusion

To advance affective computing research, it is indeed of interest to understand and model the affect exhibited by a group of people in images. In this paper, a novel framework has been presented to analyse affect of group of people in an image. Firstly, in order to increase the robustness of recognition, we investigate the higher-order Riesz transform, and employ LBP analysis on the volumes of the 1st-order and the 2nd-order Riesz components. Secondly, we exploit GEM based on continuous conditional random fields to combine the global and local attributes for estimating the group mood. The combination of feature descriptor and GEM is finally presented to infer the intensity of a group of people. Example of two inferred group images is given in Fig. 3.

We have conducted experiments on three databases to show that the proposed feature (RVLBP) considerably improves the performance of facial expression recognition, smile detection and happiness intensity estimation. Additionally, we evaluate the proposed GEM with RVLBP on the HAPPEI database for group happiness intensity estimation. The experiment results show that GEM$_{CCRF}$ results in predicting the perceived group mood more accurately. In addition, the feature indeed affects the performance of recognition for GEM models. In future, we will extend the method to different group-level emotions and perform experiments on the Group Affect database [8].

6 Acknowledgement

This work was supported by the Academy of Finland and Infotech Oulu.

References

[1] L. An, S. Yang, and B. Bhanu. Efficient smile detection by extreme learning machine. *Neurocomputing*, 149:354–363, 2015.

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

[3] C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27):1–27, 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[5] 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*, pages 2106–2112, 2011.

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

[7] A. Dhall, R. Goecke, and T. Gedeon. Automatic group happiness intensity analysis. *IEEE Transaction on Affective Computing*, 6(1):13–26, 2015.

[8] 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.

[9] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video. In *BMVC*, pages 899–908, 2006.

[10] M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processings*, 49(12):3136–3144, 2001.

[11] S. Fischer, F. Šroubek, L. Perrinet, R. Redondo, and G. Cristóbal. Self-invertible 2D log-Gobor wavelets. *International Journal of Computer Vision*, 75(2):231–246, 2007.

[12] A. Gallagher and T. Chen. Understanding images of groups of people. In *IEEE International Conference on Computer Vision*, pages 256–263, 2009.

[13] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

[14] X. He and P. Niyogi. Locality preserving projections. In *Neural Information Processing System*, 2003.

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

[16] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen. Towards a dynamic expression recognition system under facial occlusion. *Pattern Recognition Letters*, 33(16):2181–2191, May 2012.

[17] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen. Spatiotemporal local monogenic binary patterns for facial expression recognition. *IEEE Signal Processing Letters*, 19 (5):243–246, May 2012.

[18] 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*, pages 1–6, 2013.

[19] J. Kelly and S. Barsäde. Mood and emotions in small groups and work teams. *Organizational behavior and human decision processes*, 86(1):99–130, 2001.

[20] Z. Lei, S. Liao, M. Pietikäinen, and S.Z. Li. Face recognition by exploring information jointly in space, scale and orientation. *IEEE Transactions on Image Processing*, 20(1): 247–256, 2011.

[21] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002.

[22] M. Liu, S. Li, S. Shan, and X. Chen. AU-aware deep networks for facial expression recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6, 2013.

[23] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary pattern. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[24] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Proceeding of Image and Signal Processing*, pages 236–243, 2008.

[25] J. Pan and Y. Tang. Texture classification based on BIMF Monogenic signals. In *Asian Conference on Computer Vision*, pages 177–187, 2012.

[26] R. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401, 1957.

[27] T. Qin, T. Liu, X. Zhang, D. Wang, and H. Li. Global ranking using continuous conditional random fields. In *Neural Information Processing System*, 2008.

[28] R. Rosipal and L. Trejo. Kernel partial least square regression in reproducing kernel hilbert space. *Journal of Machine Learning Research*, 2:97–123, 2001.

[29] C. Shan. Smile detection by boosting pixel differences. *IEEE Transactions on Image Processing*, 21(1):431–436, 2012.

[30] E. Stein and G. Weiss. *Introduction to Fourier analysis on Euclidean spaces*. Princeton University Press, 1971.

[31] Z. Stone, T. Zickler, and T. Darell. Autotagging facebook: Social network context improves photo annotation. In *IEEE International Conference on Computer Vision*, pages 1–8, 2008.

[32] GENKI-4K Subset The MPLab GENKI Database. Available: http://mplab.ucsd.edu.

[33] M. Yang, L. Zhang, L. Zhang, and D. Zhang. Monogenic binary pattern (MBP): A novel feature extraction and representation model for face recognition. In *International Conference on Pattern Recognition*, pages 2683–2690, 2010.

[34] M. Yang, L. Zhang, S. Shiu, and D. Zhang. Monogenic binary coding: An efficient local feature extraction approach to face recognition. *IEEE Transactions on Information Forensics and Security*, 7(6):1738–1751, 2012.

[35] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two dimensional signals. *Vision Research*, 30(7):1111–1117, 1990.

[36] L. Zhang and H. Li. Encoding local image patterns using Riesz transform: With applications to palmpring and finger-knuckle-print recognition. *Image and Vision Computing*, 30(12):1043–1051, 2012.

[37] L. Zhang, D. Zhang, Z. Guo, and D. Zhang. Monogenic-LBP: A new approach for rotation invariant texture classification. In *IEEE International Conference on Image Processing*, pages 2677–2680, 2010.

[38] W. Zhang, S. Shan, W. Gao, and H. Zhang. Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. In *IEEE International Conference on Computer Vision*, pages 786–791, 2005.

[39] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary pattern with application to facial expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.