Local Zernike Moment Representation for Facial Affect Recognition

Evangelos Sariyanidi[1]
e.sariyanidi@eecs.qmul.ac.uk
Hatice Gunes[1]
hatice@eecs.qmul.ac.uk
Muhittin Gökmen[2]
gokmen@itu.edu.tr
Andrea Cavallaro[1]
andrea.cavallaro@eecs.qmul.ac.uk

[1] School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, United Kingdom

[2] Department of Computer Engineering
Istanbul Technical University
Istanbul, Turkey

Abstract

In this paper, we propose to use local Zernike Moments (ZMs) for facial affect recognition and introduce a representation scheme based on performing non-linear encoding on ZMs via quantization. Local ZMs provide a useful and compact description of image discontinuities and texture. We demonstrate the use of this ZM-based representation for posed and discrete as well as naturalistic and continuous affect recognition on standard datasets, and show that ZM-based representations outperform well-established alternative approaches for *both* tasks. To the best of our knowledge, the performance we achieved on CK+ dataset is superior to all results reported to date.

1 Introduction

Affect recognition is a fundamental building block for personal robotics, novel human-computer interfaces and a variety of assistive technologies in healthcare [10, 37]. Faces have been the primary object of analysis for affect recognition, as they provide valuable information on affective states. Most existing solutions target the analysis of posed affective behaviour, *i.e*. the recognition of exaggeratedly acted facial expressions collected in controlled environments [15, 22]. Recently, researchers have started addressing the analysis of spontaneous affective behaviour in naturalistic settings [10]. This problem is more challenging as spontaneous emotions are manifested with subtler expressions. While discrete categorical labels (*e.g*. happiness, sadness) are suitable to model posed affective behaviour, they are limited in terms of describing naturalistic affective states of daily life [11]. Continuous affect dimensions (*e.g*. arousal, valence) provide a basis for representing affective states of a much wider range and scale.

Facial affect recognition is usually formulated as a machine learning problem including the extraction of facial features for representation followed by classification (or regression). An adequate facial representation is central for effective affect recognition as the classification performance is limited by the quality and relevance of the features used in the represen-

	Discrete		Continuous	
	Global	Local	Global	Local
P	[19],[40],[17]	[1],[38],[33],[36],[28],[34],[13], [OurWork]	—	—
N	[21],[14]	[9],[6],[24],[31],[6],[14]	[25]	[32],[14],[8],[25][30], [OurWork]

Table 1: Summary of appearance representations used for affect recognition. The representations are categorised by type (local vs. global), nature of data (naturalistic, N vs. posed, P) and affect modeling (discrete vs. continuous).

tation. Important sources of information for facial representation are image discontinuities, such as furrows and wrinkles [27, 39].

Representations used for facial affect recognition are often categorized as global and local representations. Global representations consist of features that contain information on the whole image (*e.g.* DCT/PCA coefficients), while local representation features contain information extracted from local neighbourhoods. Table 1 lists a number of facial affect recognition studies, and categorizes them by the selected facial representation. Global appearance representations are used to a lesser extent compared to local representations, which not only capture discontinuities manifested at high image frequencies, but may also provide information on global appearance by preserving the global topology of the local description units [2, 16]. Although Gabor-based representations are widely used in a variety of affect recognition tasks, their computational complexity [33, 38] motivated researchers towards utilising simpler features such as Local Binary Patterns (LBPs) [26]. LBPs, the *de facto* standard in the field, describe circular regions with integers computed through pair-wise pixel comparisons. Their efficient operation scale is usually limited to circular regions with a diameter of 3-5 pixels [13, 32, 33], as they neglect the pixels inside the circular region, and the range of LBP integers grows exponentially with the number of pair-wise comparisons. However, the optimal operation scale of LBPs may not necessarily be optimal for describing affect-related low-level appearance cues.

In this paper, we propose to use local Zernike Moments (ZMs) [29] for local appearance representation, an approach that describes image discontinuities at various scales and directions (see Figure 2-b). Unlike LBPs, local ZMs provide flexibility in terms of size and detail of local description without increased computational complexity. The contributions of this paper are 1) exploring for the first time in the literature the usability and usefulness of local ZM representations and their variants for affect recognition, 2) introducing a new local descriptor based on quantising local ZMs (Quantised Local ZMs — QLZMs), and 3) introducing a global representation scheme that relies on QLZMs. To demonstrate the efficiency of these representations, we used simple statistical models (*k*-Nearest Neighbours – *k*NN) and the commonly used Support Vector Machines (SVMs).

The remainder of this paper is organised as follows. Section 2 introduces ZMs, local ZMs and describes QLZM-based face representation. Section 3 describes the experiments and the results obtained; and Section 4 concludes the paper.

2 QLZM-based Face Representation

Low-level representations are often designed as frameworks consisting of three layers [2, 12]: Low-level feature extraction, non-linear encoding and pooling. Non-linear encoding aims at enhancing the relevance of low-level features by increasing their robustness against image noise [12]. Pooling describes small spatial neighbourhoods as single entities, ignoring

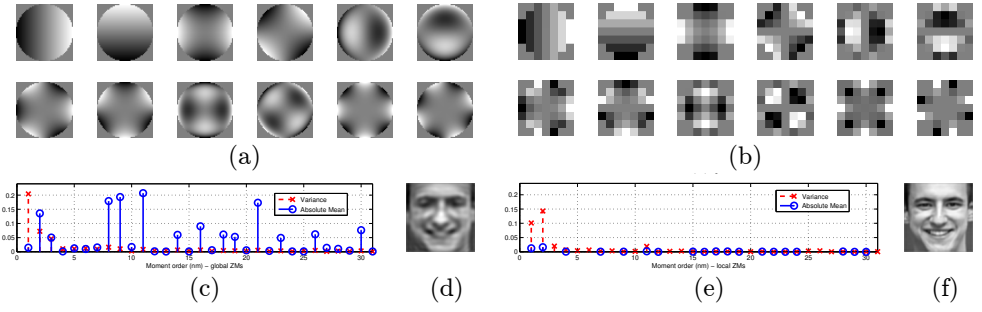


Figure 1: Example of reconstruction using global and local ZMs. (a), (b) ZM bases; (c), (e) the distribution of ZM coefficients on a set of face images; (d), (f) image reconstruction from global and local ZMs for n=35 for global and n=3 for local ZMs.

the precise location of the encoded features. The functionality of pooling is to increase robustness against small geometric inconsistencies. Our approach extracts low-level features using local ZM computation, performs non-linear encoding using quantization and pools encoded features over local histograms.

2.1 Local Zernike Moments

Let $I(x,y)$ be the input image of size $X \times Y$. ZMs are computed by decomposing $I(x,y)$ onto ZM basis matrices, a set of complex matrices that are orthogonal on the unit disk. Let the basis matrices be denoted with $V_{nm}(\cdot)$ and defined through the radial polynomials $R_{nm}(\cdot)$ as:

$$V_{nm}(\rho,\theta) = R_{nm}(\rho)e^{\mathbb{i}m\theta}, \tag{1}$$

where each radial polynomial $R_{nm}(\cdot)$ is defined as:

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^s \rho^{n-2s}(n-s)!}{s!(\frac{n+|m|}{2}-s)!(\frac{n-|m|}{2}-s)!}. \tag{2}$$

Here ρ and θ are the radial coordinates, n is the order of the polynomial that controls the number of coefficients and m is the number of iterations [33], which can be set to any value so that $|m| < n$ and $n-|m|$ is even. Let $\bar{x}$ and $\bar{y}$ be the coordinates mapped to the range $[-1,+1]$, $\rho_{xy} = \sqrt{\bar{x}^2+\bar{y}^2}$, and $\theta_{xy} = \tan^{-1}\frac{\bar{y}}{\bar{x}}$. A ZM coefficient of $I(x,y)$, Z^I_{nm}, consists of a real and an imaginary component and can be computed as follows:

$$Z^I_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} I(x,y) V^{\star}_{nm}(\rho_{xy},\theta_{xy}) \Delta\bar{x}\Delta\bar{y}. \tag{3}$$

Note that the basis matrices $V_{nm}(\cdot)$ are generic and do not depend on the input image. ZM coefficients can be used for image reconstruction, $\hat{I}(x,y)$, through inverse ZM transformation:

$$\hat{I}(x,y) = \sum_{p=0}^{K-1} \sum_{q=0}^{K-1} Z^I_{n_p m_q} V_{n_p m_q}(\rho_{xy},\theta_{xy}). \tag{4}$$

A large number of ZM coefficients is needed to reconstruct (face) images accurately since ZM basis matrices lack localisaton information. For this reason we will consider local ZMs [29] to describe image discontinuities.

Local ZMs are computed from $N \times N$ local blocks, I_N, rather than the entire image $I(x,y)$. A small number of ZM coefficients is sufficient to reconstruct I_N accurately. Figure 1 shows ZM matrices, the distribution of ZM coefficients and exemplar reconstructed face images comparatively for global and local ZMs. Note how the smooth variation in global ZM matrices becomes sharp in local matrices. ZM coefficients are scattered in a wide range when computed globally but concentrated in a short range when computed locally. This fact is reflected on the quality of the reconstructed images as well. Discontinuities that cannot be captured with global ZM coefficients are efficiently described with local coefficients even with a much smaller number of coefficients ($n = 35$ for global ZMs vs. $n = 3$ for local ZMs).

2.2 Non-linear Encoding

We perform non-linear encoding on complex-valued local ZMs through binary quantization. Specifically, we convert the real and imaginary part of each ZM coefficient into binary values through the signum$(\cdot)$ functions. Such coarse quantization increases compression and allows us to code each local block with a single integer.

Let $\mathbf{Z}^{I_N} = \begin{bmatrix} Z^{I_N}_{p_1q_1} & \dots & Z^{I_N}_{p_Kq_K} \end{bmatrix}_{1\times K}$ be a vector of K complex ZMs of I_N, and the complex notation of each coefficient be $Z^{I_N}_{pq} = Z^{I_N,\Re}_{pq} + \mathtt{i} Z^{I_N,\Im}_{pq}$. We compute Q^{I_N}, the vector of quantised local ZM coefficients as follows:

$$\mathbf{Q}^{I_N} = \begin{bmatrix} Q^{I_N,\Re}_{p_1q_1} & Q^{I_N,\Im}_{p_1q_1} & \dots & Q^{I_N,\Re}_{p_Kq_K} & Q^{I_N,\Im}_{p_Kq_K} \end{bmatrix}_{1\times 2K}, \tag{5}$$

where $Q^{I_N,\Re}_{p_iq_i} = \text{signum}(Z^{I_N,\Re}_{p_iq_i})$ and $Q^{I_N,\Im}_{p_iq_i} = \text{signum}(Z^{I_N,\Im}_{p_iq_i})$. However, the basis matrices V_{nm} must be zero-mean to ensure that the output of sgn$(\cdot)$ applied to coefficients computed through (3) is not biased. For any $m \neq 0$, this can be easily shown by computing the integral of V_{nm} over ρ and θ. For the continuous case ($\theta \in \Theta, \rho \in \mathrm{P}; \Theta = [-\pi, \pi], \mathrm{P} = [0,1]$), it can be shown that:

$$\begin{aligned} \iint_{\Theta,\mathrm{P}} V_{nm}(\rho,\theta)\mathrm{d}\rho\mathrm{d}\theta &= \int_{\Theta} e^{\mathtt{i}m\theta} \overbrace{\int_{\mathrm{P}} R_{nm}(\rho)\mathrm{d}\rho}^{C(\mathrm{P})}\mathrm{d}\theta = C(\mathrm{P}) \int_{-\pi}^{\pi} e^{\mathtt{i}m\theta}\mathrm{d}\theta \\ &= \frac{C(\mathrm{P})}{\mathtt{i}m}\left[e^{\mathtt{i}m\theta}\right]_{-\pi}^{\pi} = \frac{C(\mathrm{P})}{\mathtt{i}m}\left[(\cos\theta + \mathtt{i}\sin\theta)^m\right]_{-\pi}^{\pi} = 0. \end{aligned}$$

On the other hand, for $m = 0$ it can be shown that $\iint_{\Theta,\mathrm{P}} V_{nm}(\rho,\theta)\mathrm{d}\rho\mathrm{d}\theta = 2\pi C(\mathrm{P})$, *i.e.* the mean of basis matrices is not zero for $C(\mathrm{P}) \neq 0$. Therefore, we neglect the ZM coefficients with $m = 0$ while extracting local ZMs. Following the general rule of ZMs ($|m| < n$ and $n - |m|$), we select local ZM coefficients such as $\mathbf{Z}^{I_N} = \begin{bmatrix} Z^{I_N}_{11} & Z^{I_N}_{22} & Z^{I_N}_{31} & Z^{I_N}_{33} & \dots \end{bmatrix}_{1\times K}$ and the QLZM vector becomes $\mathbf{Q}^{I_N} = \begin{bmatrix} Q^{I_N,\Re}_{11} & Q^{I_N,\Im}_{11} & Q^{I_N,\Re}_{22} & Q^{I_N,\Im}_{22} & \dots \end{bmatrix}_{1\times 2K}$. The number of moment coefficients, K, can be considered as a function of n and is computed as:

$$K(n) = \begin{cases} \frac{n(n+2)}{4} & \text{if } n \text{ is even} \\ \frac{(n+1)^2}{4} & \text{if } n \text{ is odd.} \end{cases} \tag{6}$$

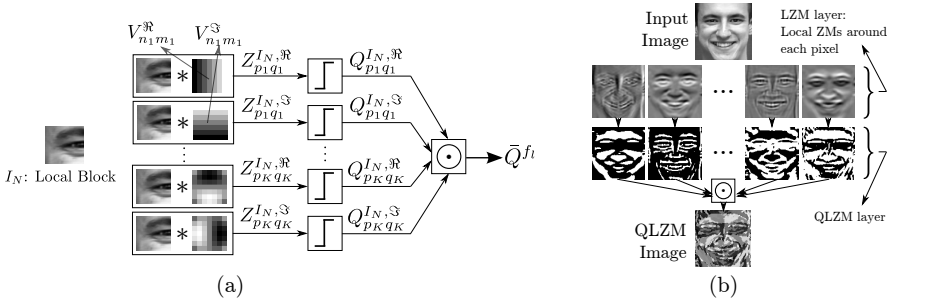


Figure 2: Computation of QLZMs and corresponding image transformation. (a) QLZMs computed from a local block. (b) QLZM image computed from an input image.

We convert the vector $\mathbf{Q}^{I_N}$ to a $2K$-bit decimal integer, QLZM integer $\bar{Q}^{I_N}$ such as $\bar{Q}^{I_N} = (Q_{11}^{I_N,\Re} Q_{11}^{I_N,\Im} Q_{22}^{I_N,\Re} Q_{22}^{I_N,\Im} \ldots)_{10}$.

A QLZM integer $\bar{Q}^{I_N}$ describes an image block. Our global representation scheme computes QLZM integers across $I(x,y)$ and codes them in the QLZM image $I_q(x_q,y_q)$. Let $I_N^{x,y}$ be a local block anchored at (x,y), and $\bar{Q}_{(x,y)}^{I_N}$ the QLZM integer of $I_N^{x,y}$. $I_q(x_q,y_q)$ is computed as $I_q(x_q,y_q) = Q_{(x_q\cdot\Delta x_q, y_q\cdot\Delta y_q)}^{I_N}$. We either set $\Delta x_q, \Delta y_q = 1$ to compute $I_q(x_q,y_q)$ from overlapping blocks, or $\Delta x_q, \Delta y_q = N$ to compute it from non-overlapping blocks. The size of I_q becomes $X \times Y$ for the former and $\frac{X}{N} \times \frac{Y}{N}$ for the latter case. These two approaches offer a trade-off between level of detail and compactness.

Figure 2 illustrates the process of extracting a single QLZM integer, the computation of QLZM image I_q, and how QLZMs encode discontinuities at different scales and orientations. The information provided by different ZM coefficients does not overlap as ZM bases are orthogonal [65]. LBPs and their variants describe texture via pairwise pixel comparisons within a local neighbourhood, and each binary value in an LBP pattern is the outcome of a pairwise comparison. Local ZMs describe image blocks as a whole, and each binary value in a QLZM pattern describes the variation within the block at a unique scale and orientation.

2.3 Pooling

Our global representation scheme pools encoded features over local histograms. However, a problem with local histograms is that in the presence of small geometric variations, features along the borders may fall out of the local histogram. To deal with this, we downweight the features along the borders by applying a Gaussian window peaked at the center of each subregion — similar strategies are employed in a number of histogram-based representations [7, 20, 29]. To account for the downweighted features, we apply a second (inner) partitioning, where a higher emphasis is placed on features downweighted at the first (outer) partitioning.

The outer partitioning scheme divides $I_q(x_q,y_q)$ uniformly in $M \times M$ subregions, $I_q^{i,j}$, with $i,j \in \{1,2,\ldots,M\}$. Let us denote the size of each $I_q^{i,j}$ with $D_W \times D_H$ and its local histogram with $h_q^{i,j}$. The inner partitioning scheme divides I_q to $(M-1)\times(M-1)$ subregions, starting from the point $(\frac{D_W}{2}, \frac{D_H}{2})$. Let the subregions extracted from the inner partitioning scheme be denoted with $I_{q'}^{i,j}$, $i,j \in \{1,2,\ldots,M-1\}$ and their histogram with $h_{q'}^{i,j}$. The final represen-

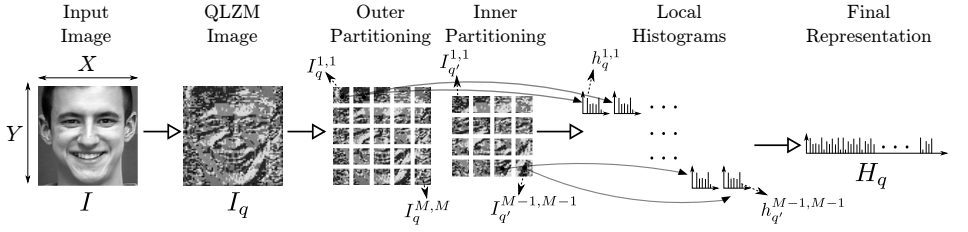


Figure 3: Illustration of the entire QLZM based facial representation framework.

tation is denoted with H_q and obtained by concatenating all local histograms $h_q^{(i,j)}$ and $h_{q'}^{(i,j)}$ (Figure 3). The length of the representation vector depends on two parameters: the number of moment coefficients (K) and the size of the grid (M). The size of each local histogram is 2^{2K}, and the length of the final vector can be computed by $2^{2K} \cdot [M^2 + (M-1)^2]$, where K can be computed using (6).

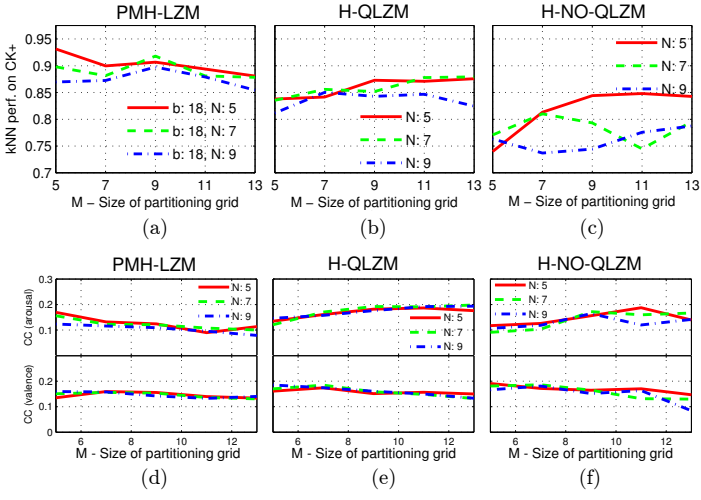
3 Experimental Evaluation

We evaluated the proposed scheme for posed and discrete facial expression recognition, as well as for spontaneous, dimensional and continuous affect recognition. We compared three local ZM representations: 1) The (non-quantised) LZM representation [29] that relies on local phase-magnitude histograms (PMHs), 2) histograms of QLZMs extracted from overlapping regions (H-QLZM), and 3) histograms of QLZMs extracted from non-overlapping regions (H-NO-QLZM). We used simple classifiers/regressors (kNN), and additionally reported performance with SVMs [8], as the results of almost all representations in the literature are reported using this machine learning technique.

Parameters that need to be determined for all local ZM representations are the moment order n, local block size N and the size of partitioning grid M. The choice of n is influenced by the size of local description N. We experimented with small N values, where local characteristics can be captured with $n = 2$ (see Figure 1). To determine the value of parameters N, M and demonstrate the sensitivity of the representation to these parameters, we carried out sensitivity analysis on posed and naturalistic data (Figure 4) — we analysed parameter sensitivity in small subsets of relevant datasets. N and M are determined separately for facial expression and spontaneous affect recognition experiments, and their values are listed in the relevant experiment sections. The non-quantised local ZM representation requires an additional parameter to be set, the number of bins for PMHs, which is set to $b = 18$. The length of representation vectors for all representations used in the sensitivity analysis are shown in Figure 4-d — sizes of H-QLZM and H-NO-QLZM representations are identical for the same M values. Overall, the sensitivity analysis shows that local ZM representations are not very sensitive to parameter changes, and small M values (*e.g.* 5, 7) can be chosen to keep the dimensionality relatively low.

3.1 Discrete and Categorical Facial Expression Recognition

The Cohn-Kanade dataset [15] is the most widely used dataset for evaluating automatic facial expression recognizers for discrete and categorical emotion recognition, and we used its



Vector lengths of representations *w.r.t.* M

M	PMH-LZM	H-QLZM
5	1476	656
7	3060	1360
9	5220	2320
11	7956	3536
13	11268	5008

(g)

Figure 4: Sensitivity analysis of (a–c) PMH-LZM, H-QLZM and H-NO-QLZM representations on a subset of CK+ dataset, (d–f) PMH-LZM, H-QLZM and H-NO-QLZM representations on a subset of AVEC data. (g) Corresponding grid size (M) and vector sizes.

most recent version, the Extended CK (CK+) dataset [22]. This version includes 327 image sequences of 118 subjects, labeled by experts with one of the six basic emotion categories (anger, disgust, fear, happy, sadness, surprise) and a non-basic emotion category (contempt). Similarly to the majority of the techniques in the literature, we aim at classifying the peak (apex) frame of the sequences. We followed the standard leave-one-subject-out cross validation protocol [22] and reported the unweighted average classification accuracy and standard error measure over 118 folds. We evaluated our approach in comparison with results reported in recent publications (year 2012) that use a variety of representations such as AAM-based appearance representation (CAPP) [4], Gabor and LBP representations [34], Local Directional Numbers (LDN) [28] and a Bag-of-Words (BoW) representation [34].

Pre-processing — Faces are registered to align the centers of two eyes (computed by taking the average of eye-related AAM landmarks provided as part of the dataset), and downsampled to 150×150.

Representation Parameters — Parameter pairs (M, N) are set to (5,5), (7,7) and (9,5) respectively for PMH-LZM, H-QLZM and H-NO-QLZM representations.

Classification — For kNN classification we used L_1 distance and reported results for three different k values (5,7,9). For SVM, we used linear and radial basis function (RBF) kernels, and trained one-vs-one classifiers (with probabilistic output) for each expression. Classification was obtained based on the probability of SVMs.

3.2 Dimensional and Continuous Affect Recognition

We used the well-known Audio/Visual Emotion Recognition Challenge (AVEC'12) data and evaluation protocol to evaluate the performance of local ZM representations in naturalistic settings. The AVEC'12 challenge uses the SEMAINE database [23] that consists of videos recorded while subjects are having conversations with artificial listeners. The affective state of the subjects is annotated along multiple continuous affect dimensions [32]. AVEC 2012

Performance of local ZM-based methods

Represent.	Classifier				
	*k*NN			SVM	
	k=5	*k*=7	*k*=9	Lin.	RBF
PMH-LZM	92.1 ±1.6	**93.1 ±1.4**	90.7 ±1.8	94.4 ±1.9	95.5 ±1.8
H-QLZM	85.3 ±2.3	87.8 ±2.2	86.4 ±2.3	94.2 ±1.9	**96.1 ±1.6**
H-NO-QLZM	82.6 ±2.4	84.5 ±2.4	80.9 ±2.5	90.9 ±2.0	92.0 ±1.8

Performance of other methods

Represent.	Class.	Rec. Rate (%)
CAPP [9]	SVM Lin.	70.1 §
	Uni-Hyp.	89.4 §
LDN [25]	SVM Lin.	89.3±0.6 *
	SVM RBF	89.3±0.7 *
	SVM Pol.	81.7±0.7 *
Gabor [34]	SVM Lin.	91.8±2.0 †
LBP [34]	SVM Pol.	82.4±2.3 †
BoW [42]	SVM Lin.	**95.9±1.4** †
H-QLZM	SVM RBF	**96.1±1.6** †† (96.2§)

(a)

Represent.	Regressor							
	*k*NN						SVM	
	k=35		*k*=45		*k*=55		HI	
	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
PMH-LZM	0.155	0.132	0.151	0.132	0.151	0.130	0.149	0.130
H-QLZM	**0.167**	0.174	0.161	0.175	0.161	0.175	**0.197**	0.221
H-NO-QLZM	0.152	0.194	0.156	0.198	0.156	**0.203**	0.195	**0.229**
LBP (Baseline)	–	–	–	–	–	–	0.151	0.207

(b)

Table 2: Performance of local ZM based approaches. (a) Discrete categorical expression recognition performance. (b) Continuous dimensional affect recognition performance. *Evaluation metric not stated. †Average recognition rate (weighted or not is not clear) and standard error. ††Unweighted average recognition rate and standard error over 118 folds. §Trace of confusion matrix (mean value of the elements on the main diagonal).

included two sub-challenges, the Fully Continuous Sub-Challenge (FCSC) and Word-level Sub-Challenge. Since our interest is vision-based recognition, we perform experiments using the FCSC protocol. We compare our results with the baseline results obtained using visual features (LBP histogram representation). Although other AVEC participants reported results using this protocol, we do not compare our technique to theirs — they utilised ensembles of audio/visual features without reporting separate results for each modality, or relied on sophisticated components (classifiers, feature selection schemes) that are not trivial to reproduce. We emulate the experimental setup of the baseline by performing similar preprocessing steps and using similar machine learning techniques. We report results for valence and arousal, the most widely used affect dimensions in the literature [10]. As defined in the challenge protocol, we report performance using the cross correlation (CC) metric (Pearson's correlation).

Pre-processing — Face rectangles and eye locations are provided as part of the challenge data. Despite the imperfection of the localisations provided, for a fair comparison with the baseline, we used these features to register the faces. We resized the frames to make the distance between the eye centers equal to 100 pixels, and cropped faces to be 200×200 pixels, and aggregated the histograms over 50 frames to reduce computational overhead.

Representation Parameters — Parameter pairs (M, N) are set to (7,7), (7,7) and (5,7) respectively for PMH-LZM, H-QLZM and H-NO-QLZM representations.

Regression — For *k*NN regression we computed the average of k neighbours by weighting them according to the similarity ($1/(L_1$ distance)), and reported results for three different k values (25,35,45). For SVM regression, we trained separate SVMs with histogram intersection (HI) kernels for each dimension and performed subject-independent cross validation on the training subset to optimize the parameters of the SVR models.

3.3 Discussion

Table 2-a shows that for discrete and categorical facial expression recognition on CK+ dataset, the best performance is attained with H-QLZM representation used in conjunction with RBF SVM. The results we attained outperform all previous results reported on this dataset to date. Our results are also better than the BoW-based representation [34], which requires a training stage for dictionary learning and makes use of features at multiple scales through a pyramid matching structure [18]. In spite of CK+ being an unbalanced dataset (*e.g.* 18 instances of contempt vs. 83 instances of surprise), simple *k*NN classifiers lead to better performance than several sophisticated classifiers (*e.g.* [4]), demonstrating the importance of appropriate facial representation.

Table 2-b shows the results on AVEC benchmark data. These results suggest that QLZM-based representation outperforms PMH-LZM for continuous affect recognition on naturalistic data. The performance of PMH-LZM is also below the baseline LBP representation. With SVM regression, QLZM-based representation outperforms the baseline LBP representation on both affect dimensions, and even the simple *k*NN regression leads to better results on the arousal dimension. While the coarse non-overlapping representation (H-NO-QLZM) was consistently outperformed by other representations on posed data, it performs well on naturalistic data, yielding the best performance on the valence dimension. Interestingly for the arousal dimension highest performance is obtained using overlapping representation (H-QLZM) while for the valence dimension using non-overlapping representation (H-NO-QLZM). On the other hand, the finest representation (PMH-LZM), which yields the best *k*NN results on posed expressions, performs consistently worse than QLZM representations as well as the baseline system on naturalistic data.

The results demonstrate that the amount of detail needed in the facial representation for affect recognition in posed settings may be quite different for recognition in naturalistic settings. The idealized context of posed affective behaviour favours finer representations, which capture more details. The naturalistic settings require coarser representations. We show that quantization has a positive effect for the task of naturalistic affect recognition, which is likely to be due to its better generalization capability via clustering similar features.

Local ZM-based representations are computationally efficient, *i.e.*, average computation times of PMH-LZM, H-QLZM and H-NO-QLZM on images of 200×200 pixels are respectively 37, 18 and 1 milisecond(s) (on an average desktop computer with an Intel i5 processor without parallelization). Each local ZM coefficient is computed as the component-wise inner product (Frobenius product) of the image block with the corresponding ZM basis matrix, and the quantization that takes place for QLZMs is computationally trivial. The source code of representations is made available for research purposes[1].

4 Conclusion

We proposed to use Local Zernike Moments (LZMs) and introduced Quantised Local Zernike Moments (QLZMs) for categorical and discrete as well as spontaneous, dimensional and continuous affect recognition. We showed that LZM and QLZM representations are useful, but the most appropriate representation for affect recognition in posed and naturalistic settings may differ (unlike the common practice: [33] vs. [30, 32]).

[1] http://cis.eecs.qmul.ac.uk/software.html

The promising experimental results attained by QLZMs is mainly due to the application of binary quantization for non-linear encoding. This finding suggests that the full potential of alternative non-linear encoding schemes (e.g., quantization with finer resolutions or using non-linear functions with continuous output) for automatic affect recognition should be explored further.

5 Acknowledgments

The work of E. Sariyanidi and H. Gunes is partially supported by the EPSRC MAPTRAITS Project (Grant Ref: EP/K017500/1) & the British Council UK-Turkey Higher Education Partnership Programme Project (Grant Ref: TR/012012/KP40). The work of M. Gökmen is partially supported by The Scientific and Technological Research Council of Turkey with the grant number 112E201.

References

[1] Marian Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6), 2006.

[2] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the International Conference on Machine Learning*, pages 111–118, 2010.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

[4] Sien W. Chew, Simon Lucey, Patrick Lucey, Sridha Sridharan, and Jeff F. Cohn. Improved facial expression recognition via uni-hyperplane classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2554–2561, 2012.

[5] Albert Cruz, Bir Bhanu, and Songfan Yang. A psychologically-inspired match-score fusion mode for video-based facial expression recognition. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 341–350, 2011.

[6] Mohamed Dahmane and Jean Meunier. Continuous emotion recognition using Gabor energy filters. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 351–358, 2011.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[8] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using PHOG and LPQ features. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, pages 878–883, 2011.

[9] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, and Friedhelm Schwenker. Multiple classifier systems for the classification of audio-visual emotional states. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 359–368. 2011.

[10] Hatice Gunes and Björn Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31 (2):120 – 136, 2013.

[11] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddie Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition*, pages 827–834, 2011.

[12] Kevin Jarrett, Koray Kavukcuoglu, Marc' Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2146–2153, 2009.

[13] Bihan Jiang, Michel Valstar, Brais Martinez, and Maja Pantic. Dynamic appearance descriptor approach to facial actions temporal modelling. *IEEE Transactions of Systems, Man and Cybernetics – Part B*, 2013.

[14] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *International Symposium on Advances in Visual Computing*, pages 368–377. 2012.

[15] Takeo Kanade, Jeffrey Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[16] Jan J. Koenderink and Andrea J. Van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31:159–168, 1999.

[17] Seyed Mehdi Lajevardi and Zahir M. Hussain. Higher order orthogonal moments for invariant facial expression recognition. *Digital Signal Processing*, 20(6):1771 – 1779, 2010.

[18] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[19] Daw-Tung Lin. Facial expression classification using pca and hierarchical radial basis function network. *Journal of Information Science and Engineering*, 22(5):1033–1046, 2006.

[20] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[21] Patrick Lucey, Jeffery Cohn, Simon Lucey, Iain Matthews, S. Sridharan, and K.M. Prkachin. Automatically detecting pain using facial actions. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8, 2009.

[22] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 94 –101, 2010.

[23] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3 (1):5–17, 2012.

[24] Loris Nanni, Alessandra Lumini, and Sheryl Brahnam. Local binary patterns variants as texture descriptors for medical image analysis. *Artificial Intelligence in Medicine*, 49(2):117 – 125, 2010.

[25] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 501–508, 2012.

[26] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[27] Maja Pantic and Marian Stewart Bartlett. Machine analysis of facial expressions. pages 377–416. I-Tech Education and Publishing, Vienna, Austria, July 2007.

[28] A. Ramirez Rivera, J. Rojas Castillo, and O. Chae. Local directional number pattern for face analysis: Face and expression recognition. *IEEE Transactions on Image Processing*, PP(99):1, 2012.

[29] Evangelos Sariyanidi, Volkan Dagli, Salih Cihan Tek, Birkan Tunc, and Muhittin Gökmen. Local Zernike Moments: A new representation for face recognition. In *Proceedings of the IEEE International Conference on Image Processing*, pages 585–588, 2012.

[30] Arman Savran, Houwei Cao, Miraj Shah, Ani Nenkova, and Ragini Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 485–492, 2012.

[31] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. AVEC 2011 - the first international audio/visual emotion challenge. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. 2011.

[32] Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. AVEC 2012 - the continuous audio / visual emotion challenge. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 361–362, 2012.

[33] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803 – 816, 2009.

[34] Karan Sikka, Tingfan Wu, Josh Susskind, and Marian Bartlett. Exploring bag of words architectures in the facial expression domain. In *Proceedings of the European Conference on Computer Vision Workshops and Demonstrations*, pages 250–259. 2012.

[35] Michael Reed Teague. Image analysis via the general theory of moments. *J. Opt. Soc. Am*, 70(8):920–930, 1980.

[36] Michel F. Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Klaus Scherer. The first facial expression recognition and analysis challenge. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition*, pages 921 –926, 2011.

[37] Alessandro Vinciarelli, Maja Pantic, and Hervè Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743 – 1759, 2009.

[38] Jacob Whitehill and Christian W Omlin. Haar features for FACS AU recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 101–105, 2006.

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

[40] Ruicong Zhi and Qiuqi Ruan. A comparative study on region-based moments for facial expression recognition. In *Proceedings of Congress on Image and Signal Processing*, pages 600–604, 2008.