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

Local representations became popular for facial affect recognition as they efficiently capture the image discontinuities, which play an important role for interpreting facial actions. We propose to use Local Zernike Moments (ZMs) [4] due to their useful and compact description of the image discontinuities and texture. Their main advantage in comparison to well-established alternatives such as Local Binary Patterns (LBPs) [5], is their flexibility in terms of the size and level of detail of the local description. We introduce a local ZM-based representation which involves a non-linear encoding layer (quantisation). The functionality of this layer is mapping similar facial configurations together and increasing compactness. We demonstrate the use of the local ZM-based representation for posed and naturalistic affect recognition on standard datasets, and show its superiority to alternative approaches for *both* tasks.

Contemporary representations are often designed as frameworks consisting of three layers [2]: (Local) feature extraction, non-linear encoding and pooling. Non-linear encoding aims at enhancing the relevance of local features by increasing their robustness against image noise. Pooling describes small spatial neighbourhoods as single entities, ignoring the precise location of the encoded features, and increasing the tolerance against small geometric inconsistencies. In what follows, we describe the proposed local ZM-based representation scheme in terms of this three-layered framework.

Feature Extraction – Local Zernike Moments: The computation of (complex) ZMs can be considered equivalent to representing an image in an alternative space. As shown in Figure 1-a, an image is decomposed onto a set of basis matrices (ZM bases), which are useful for describing the variation at different directions and scales. ZM bases are orthogonal, therefore there is no overlap in the information conveyed by each feature (ZM coefficient). ZMs are usually computed for the entire image, however in this case, ZMs cannot capture the local variation due to ZM bases lacking localisation [3]. In contrary, when computed around local neighbourhoods across the image, they become an efficient tool for describing the image discontinuities which are essential to interpreting facial activity.

Non-linear Encoding – Quantisation: We perform quantisation via converting local features into binary values. Such coarse quantisation increases compactness and allows us to code each local block only with a single integer. Figure 1-b illustrates the process of obtaining the Quantised Local ZM (QLZM) image. Firstly, local ZM coefficients are computed across the input image (LZM layer) — each image in the LZM layer (LZM image) contains the features that are extracted through a particular ZM basis. Next, each LZM image is converted into a binary image by quantising each pixel via the signum$(\cdot)$ function. Finally, the QLZM image is obtained by combining all of the binary images. Specifically, each pixel in a particular location of the QLZM image is an integer (QLZM integer), computed by concatenating all of the binary values in the corresponding location of all binary images. The QLZM image is similar to an LBP-transformed image, in the sense that it contains integers of a limited range. Yet, the physical meaning of the information encoded by each integer is quite different. LBP integers describe a circular block by considering only the values along the border, neglecting the pixels that remain inside the block. Therefore, the efficient operation scale of LBPs is usually limited to 3-5 pixels [1, 5]. QLZM integers, on the other hand, describe blocks as a whole, and provide flexibility in terms of operation scale without major loss of information.

Pooling – Histograms: Our representation scheme pools encoded features over local histograms. Figure 1-c illustrates the overall pipeline of the proposed representation scheme. Firstly, the QLZM image is computed through the process that is illustrated in detail in Figure 1-b. Next,

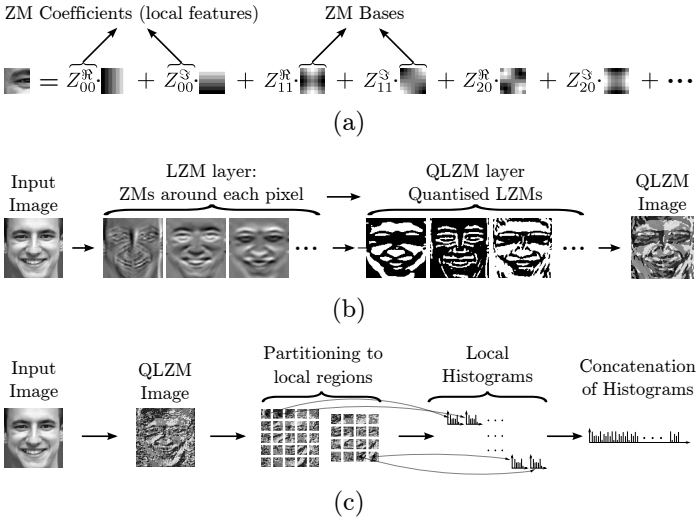

Figure 1: Illustration of the proposed representation scheme. (a) Local ZM features (ZM coefficients) and the ZM bases. (b) Illustration of the process of computing the QLZM image. (c) Illustration of the overall representation pipeline.

the QLZM image is divided into uniform regions, and the local histograms that count the QLZM integers are computed for each region. The final representation is obtained by concatenating all local histograms.

Our experimental results show that the local ZM based representation is superior to its well-established alternatives. By experimenting both on datasets of posed and spontaneous affective behaviour, we show that the suitable level of detail to encode differs for these two different contexts, and the contribution of quantisation is more pronounced on spontaneous data. While fine-grained representation is suitable for the idealised context of posed affective behaviour, coarser representation via quantisation is more suitable for spontaneous behaviour. This is likely to be due to the better generalisation capability that quantisation provides. This improvement suggests that the full potential of alternative non-linear encoding schemes such as quantisation with finer resolutions or non-linear functions with continuous output should be explored further.

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12): 2037–2041, 2006.

[2] 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.

[3] R. Rubinstein, A.M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[4] 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.

[5] 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.