Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition

Yuan An

In the last article, facial registration and spatial representation is introduced [7]. Today, spatio-temporal representations is talked about in brief.

Spatio-temporal Representations

Spatio-temporal representations consider a range of frames within a temporal window as a single entity, and enable modeling temporal variation in order to represent subtle expressions more efficiently.

They can discriminate the expressions that look similar in space (*e.g.* closing eyes and eye blinking [4, 3]), and facilitate the incorporation of domain knowledge from psychology. This domain knowledge relates the muscular activity with higher level tasks, such as distinguishing between posed and spontaneous affective behavior or recognition of temporal phases.

Geometric Features from Tracked Facial Points

This representation is designed to incorporating the knowledge from cognitive science to analyze temporal variation and the corresponding muscular activity. It has been used for the recognition of AUs with their temporal phase [9], and the discrimination of spontaneous between posed smiles and brow actions.

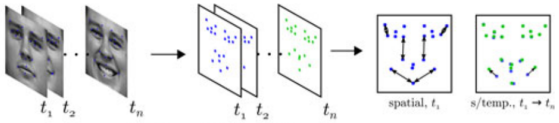

Figure 1. Geometric features from tracked feature points

The representation describes the facial shape and activity by means of fiducial point [9]. To this end, it uses the raw location of each point, the length and angle of the lines obtained by connecting all points pairwise in space, and the differences obtained by comparing these features with respect to their value in a neutral face. Some of these features describe componential information such as the opening of the mouth, as well as configural information such as the distance between the corner of the eye and the nose (see Figure. 1). Other features aim at capturing temporal variation.

The representation is sensitive to registration errors as its features are mostly extracted from raw or differential point coordinates. Although the representation describes temporal variation, it may not capture subtle expressions as it is extracted from a small number of facial points and depends on accurate point registration.

Low-level Features from Orthogonal Plane

Extracting features from three orthogonal planes (TOP) is a popular approach towards extending low-level spatial appearance representations to the spatio-temporal domain (see Figure 2 and 3). This paradigm originally emerged when extending LBP to LBP-TOP. LBP-TOP is applied for basic emotion recognition [15, 14] and AU recognition. Following this method, LPQ is extended to LPQ-TOP and used for AU and temporal segment recognition [2, 1].

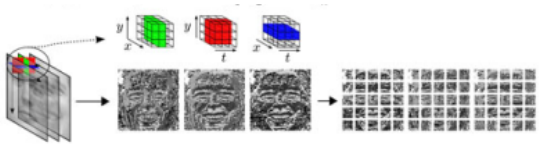

Figure 2. LBP-TOP, and the TOP paradigm

Figure 3. LPQ-TOP

Convolution with Smooth Filter

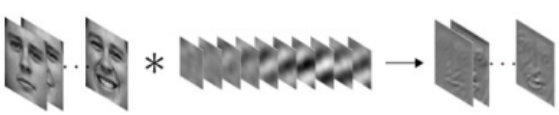

Figure 4. Spatio-temporal ICA filtering, the out put on an exemplar spatio-temporal filter.

An alternative approach for representing the temporal variation in texture with low-level features is applying

convolution with smooth spatio-temporal filters (see Figure 4). Two such approaches are spatio-temporal Gabor filtering [10] and spatio-temporal independent component (IC) filtering [6]. Both approaches target explicitly the recognition of subtle expressions.

Gabor and IC filters are localized in space and time. At the spatial level, the output of the filtering encodes componential information. The main difference between the Gabor and IC filters is that the parameters of Gabor filters are adjusted manually, while IC filters are obtained automatically in the process of unsupervised Independent Component Analysis.

Spatio-temporal Haar Representations

There are two representations using the well-established Haar features for spatio-temporal representation. They are the dynamic Haar features [13] and similarity features [11, 12]. The former is a straightforward temporal extension of the Haar features, whereas the latter tailors an overall representation scheme for affect recognition.

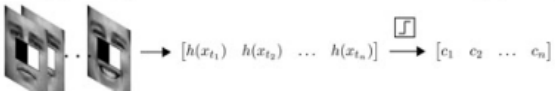

Figure 5. Dynamic Haar representation

As illustrated in Figure 5, each dynamic Haar feature encodes the temporal variation in an image sequence with a pattern of binary values, where each binary value is obtained by thresholding the output of the Haar feature in the corresponding frame.

Free-form Deformation Representation

The free-form deformation representation extends free-form deformation, which is essentially a registration technique, into a representation that extracts features in the process of registration by computing the pixels' spatial and temporal displacement (see Figure 6). This representation is used for AU recognition with temporal segments [5].

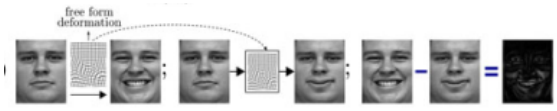

Figure 6. Free-form deformation representation, illustration of free-form deformation.

Temporal Bag-of-Words Representation

The temporal BoW representation is specific to AU detection and can be best explained by describing how the problem is formulated by its authors. Simon *et al.* [8] assume that an AU is an event that exists in a given image sequence. The problem is then formulated as identifying the boundaries of the existing AU event. The approach was also generalized for multiple AUs.

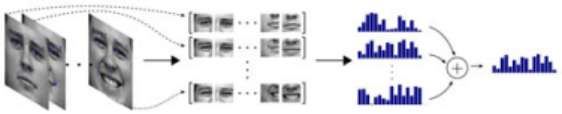

Figure 7. Temporal BoW.

Temporal BoW represents an arbitrary subset of the given image sequence with a single histogram which is computed as follows (see Figure 7):

1. Each fream in the subset is represented using the part-based SIFT representation and compressed with principal component analysis to obtain a frame-wise vector.

2. Each frame-wise vector is encoded using the BoW paradigm that measures similarity by means of multiple vectors via soft clustering.

3. All encoded frame-wise vectors are collected in a histogram.

References

[1] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Transactions on Cybernetics*, 44(2):161–174, 2014. 1

[2] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *IEEE Conference on Automatic Face and Gesture Recognition*, pages 314–321, 2011. 1

[3] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368–377, 2012. 1

[4] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1940–1954, 2010. 1

[5] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1940–1954, 2010. 2

[6] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett, and G. Littlewort. Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. *Neurocomputing*, 93(2):126–132, 2012. 2

[7] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, 2015. 1

[8] T. Simon, M. H. Nguyen, F. D. L. Torre, and J. F. Cohn. Action unit detection with segment-based svms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2737–2744, 2010. 2

[9] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Cybernetics*, 42(1):28–43, 2012. 1

[10] T. Wu, M. S. Bartlett, and J. R. Movellan. Facial expression recognition using gabor motion energy filters. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 42–47, 2010. 2

[11] P. Yang, Q. Liu, and D. Metaxas. Similarity features for facial event analysis. In *European Conference on Computer Vision*, pages 685–696, 2008. 2

[12] P. Yang, Q. Liu, and D. Metaxas. Dynamic soft encoded patterns for facial event analysis. *Computer Vision and Image Understanding*, 115(3):456 – 465, 2011. 2

[13] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007. 2

[14] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. 1

[15] G. Zhao and M. Pietikäinen. Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. *Pattern Recognition Letters*, 30(12):1117–1127, 2009. 1