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

Yuan An

In the last article, the first two section of Sariyanidi *et al*. [10] review paper is introduced. Today, face registration and spatial representations is talking about in brief.

Face Registration

Face registration is a fundamental step for facial affect recognition. Depending on the output of the registration process, registration strategies can be categorized as whole face, part and point registration.

Whole Face Registration

The region of interest for most systems is the whole face. The technique used to register the whole face can be categorised as rigid and non-rigid. Rigid registration is generally performed by detecting facial landmarks and using their location to compute a global transformation (*e.g*. Euclidean, affine) that maps an input face to a prototypical (typical) face. Many systems use the two eye points or the eyes and nose or mouth [5, 7]. The transformation can also be computed from more points (*e.g*. 60-70) using techniques such as Active Appearance Models (AAM) [2]. Alternatively to landmark-based approaches, generic image registration techniques such as Robust FFT [13] or Lucas-Kanade approaches [1].

While rigid approaches register the face as a whole entity, non-rigid approaches enable registration locally and can suppress registration errors duo to facial activity. For instance, an express face (*e.g*. smiling face) can be warped into a neutral face. Techniques such as AAM are used for non-rigid registration by performing piece-wise affine transformations around each landmark [8].

Parts Registration

A number of appearance representation process face in terms of parts (*e.g*. eyes, mouth), and may require the spatial consistency of each part to be ensured explicitly. The number, size and location of the parts to be registered may vary (*e.g*. 2 large parts [12] or 36 small parts [15]). Similarly to whole face registration, a technique used frequently for parts registration is AMM–the parts are typically localized as fixed-sized patches around detected landmarks.

Point Registration

Points registration is needed for shape representations, for which registration involves the localization of fiducial point. Similarly to whole and parts registration, AMM is used widely for points registration. Points in a sequence can also be registered by localizing points using a point deector on the first frame and then racking them.

Spatial Representations

Spatial representations encode image sequences frame-by-frame. There exists a variety of appearance representations that encode low or high-level information. Low-level information is encoded with low-level histograms(see Figure 1), Gabor representations and data-driven representations such as those using bag-of-words (BoW). Higher level information is encoded using for example non-negative matrix factorization (NMF) or sparse coding. There exist hierarchical representations that consist of cascaded low and high-level representation layers. Several appearance representations are part-based. Shape representations are less common than appearance representations.

Figure 1. Low-level histograms.

Shape Representations

The most frequently used shape representation is the facial points representation, which describes a face by simply concatenating the x and y coordinates of a number of fiducial points. When the neutral face image is available, it can be used to reduce identity bias (see Figure 2a). This representation reflects registration errors straightforwardly as it is based on either raw or differential coordinate values. Illumination variations are not an issue since the intensity of the pixels is ignored.

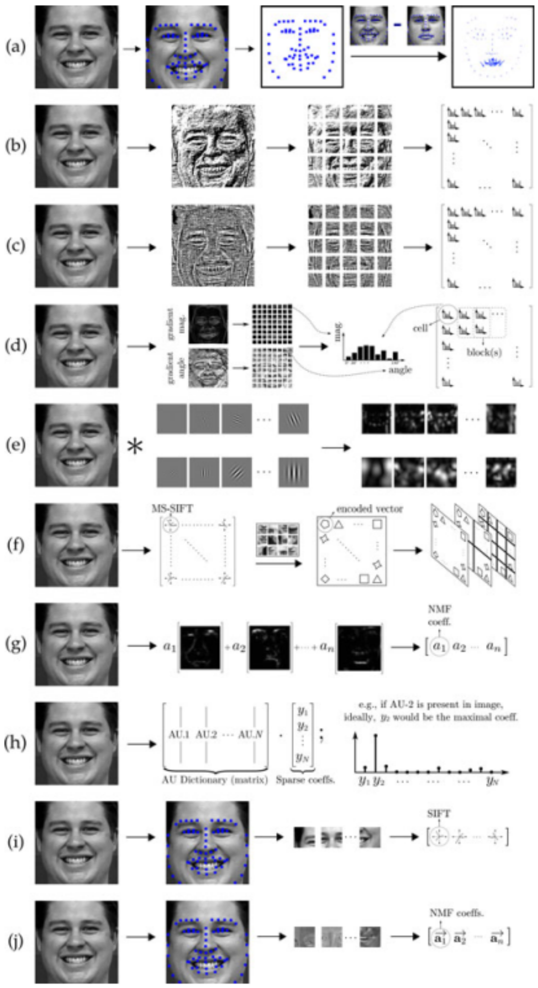

Figure 2. Spatial representations. (a) Facial points; (b) LBP histograms; (c) LPQ histograms (d) HoG; (e) Gabor-based representation; (f) dense BoW; (g) GP-NMF; (h) sparse conding; (i) part-based SIFT; (j) part-based NMF.

Low-level Histogram Representations

Low-level histogram representations (see Figure 2b, 2c, and 2d) first extract local features and encode them in a transformed image, then cluster the local features into uniform regions and finally pool the features of each region with local histograms. The representations are obtained by concatenating all local histograms.

Low-level features are robust to illumination variations to a degree, as they are extracted from small regions. Also, they are invariant to global illumination variations (i.e. grayscale shifts). Additionally, the histograms can be normalized (*e.g.* unit-norm normalization [4]) to increase the robustness of the overall representation.

Gabor Representation

Another representation based on low-level features is the Gabor representation, which is used by various systems including the winner of the FERA AU dection challenge and AVEC.

A Gabor representation is obtained by convolving the input image with a set of Gabor filters of various scales and orientations (see Figure 2e) [6, 14].

Bag-of-Words Representation

The BoW representation used in affect recognition [11] describes local neighborhoods by extracting local features (i.e. SIFT) densely from fixed locations and then measuring the similarity of each of these features with a set of features in a dataset using locality constrained linear coding. The representation inherits the robustness of SIFT features against illumination variations and small registration error. The representation use spatial pyramid matching, a technique that performs histogram pooling and increases the tolerance to registration errors. This matching scheme encodes componential information at various scales (see Figure 2f), and the layer that does not divided the image to subregions conveys holistic information.

High-level Data-driven Representations

All representations discussed so far describe local texture (see Figure 2a-f). Implicitly or explicitly, their features encode the distribution of edges. Recent approaches aim instead at obtain data-driven higher-level representations to encode features that are semantically interpretable from an affect recognition perspective. Two methods that generate such representations are NMF and sparse coding. Alternatively, various feature learning approaches can also be used [9].

Hierarchical Representations

Low-level representations are robust against illumination variations and registration errors. On the other hand high-level representations can deal with issues such as identity bias and generate features that are semantically interpretable. Hierarchical representation encode information in a low to high level manner. Hierarchical representations can alternatively be designed straightforwardly by cascading well-established low and high level representations such as Gabor filters and sparse representations [3].

Part-based Representation

Part-based representation process faces in terms of independently registered parts and thereby encode componential information. They discard configural information explicitly as they ignore the spatial relations among the registered parts (see Figure 2j and 2). Ignoring the spatial relationships reduces the sensitivity to head-pose variation. Part-based representations proved successful in spontaneous affect recognition tasks where head-pose variation naturally occurs.

References

[1] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004. 1

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. 1

[3] S. F. Cotter. Sparse representation for accurate classification of corrupted and occluded facial expressions. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 838–841, 2010. 2

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 2

[5] 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

[6] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993. 2

[7] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797 – 1803, 2009. 1

[8] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews. Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197 – 205, 2012. 1

[9] B. R. Paredes, A. Argyriou, N. Berthouze, and M. Pontil. Exploiting unrelated tasks in multi-task learning. In *International Conference on Artificial Intelligence and Statistics*, pages 951–959, 2012. 2

[10] 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

[11] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision*, pages 250–259, 2012. 2

[12] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007. 1

[13] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki. Robust FFT-based scale-invariant image registration with image gradients. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1899–1906, 2010. 1

[14] L. Wiskott, N. Krüger, N. Kuiger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997. 2

[15] Y. Zhu, F. D. la Torre, J. F. Cohn, and Y. J. Zhang. Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *IEEE Transactions on Affective Computing*, 2(2):79–91, 2011. 1