Automatic Facial Expression Recognition using Bags of Motion Words

Liefei Xu
http://www.cs.stevens.edu/~lxu1/
Philippos Mordohai
http://www.cs.stevens.edu/~mordohai/

Department of Computer Science, Stevens Institute of Technology
Hoboken, NJ, USA

The human visual system is highly specialized in recognition tasks related to people and, in particular, faces. We are very adept at identifying discriminating features in people's faces and sensitive to the emotional states manifested by their facial expressions. Arguably, we are less sensitive in similar discriminative tasks for animals and objects. Computer vision research has followed a similar path by developing specialized techniques for face and expression recognition and more general methods for object recognition. Here, we investigate whether a general approach for expression recognition based on motion is feasible.

Several researchers [1, 2, 4, 6] have argued that motion provides strong cues for expression recognition. By relying exclusively on motion features these methods provide invariance to the subjects' ethnic background, facial hair, make up and accessories. Our approach differs from previous work by not being tailored to faces; we use general motion features, instead of highly detailed models of the human face, eyes and mouth. Since we do not employ face-specific models, we do not rely on FACS coding [3], but only use a single expression label per sequence for training.

Annotation requirements for training are very low: each video is labeled according to the displayed expression, but individual frames do not need to be labeled, fiducials do not need to be marked on the faces and the sequence does not need to begin or end with a neutral expression. An additional constraint we imposed on our method is that it should be applicable to videos captured by a single unknown camera. Face and expression recognition can benefit greatly from the availability of additional modalities such as 3D or infrared, but this restricts them to processing data captured by a particular sensor or at a specific location. Finally, we require our method to be fully automatic. The user should not have to identify neutral faces or the apex of the expression in the sequence, and more importantly the user should not have to localize the eyes or other features in the first frame in order to align a model with the input.

For each video sequence, we estimate optical flow fields between consecutive frames, which are then concatenated to generate optical flow fields with larger motions. On these motion fields, we compute a dense set of local descriptors of the motion vectors following the encoding of the SIFT descriptor [5], but without scale detection or rotation invariance. We, then, cluster the descriptors to generate a *motion vocabulary* in which the words are the cluster centroids [7].

During testing, we compute motion descriptors in the same way and assign each descriptor to one of the words. Thus, we obtain a histogram of word frequencies in each frame, which we use as the frame signature. We assign a label to each frame based on its signature. Finally, we classify sequences according to the labels assigned to the frames that comprise them. The processing steps are summarized in the following figure. We have compared several alternatives for the three main classification tasks: the assignment of descriptors to words, frame classification and sequence classification.

Here, we present recognition results on videos of the six universal expressions: happiness, sadness, fear, surprise, anger and disgust on a database of 600 video sequences. Our method could also be trained on data labeled in other ways, for instance according to AU activations, without modification.

[1] J. N. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2058, 1979.

[2] M.J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.

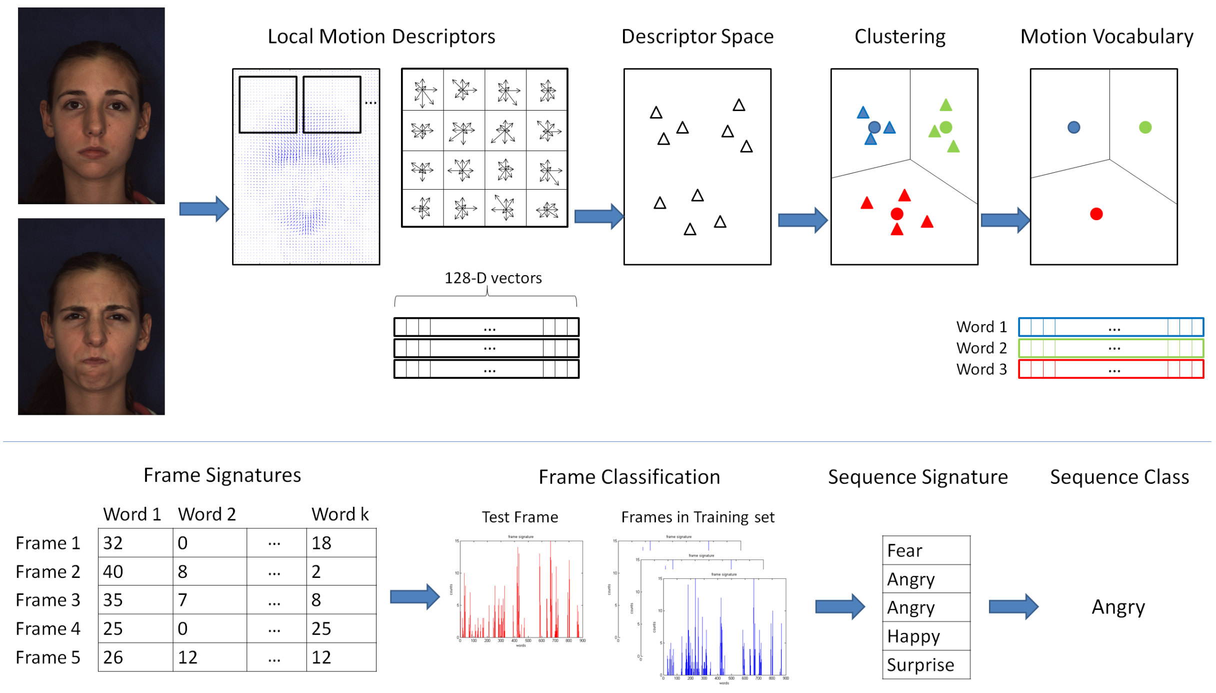
[3] P. Ekman and W.V. Friesen. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[4] I.A. Essa and A.P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

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

[6] K. Mase. Recognition of facial expression from optical flow. *IEICE*, E74(10):3474–3483, 1991.

[7] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Int. Conf. on Computer Vision*, pages 1470–1477, 2003.



	Word 1	Word 2		Word k
Frame 1	32	0	...	18
Frame 2	40	8	...	2
Frame 3	35	7	...	8
Frame 4	25	0	...	25
Frame 5	26	12	...	12