Multi-view Facial Expression Recognition Analysis with Generic Sparse Coding Feature

Usman Tariq[1], Jianchao Yang[1,2], and Thomas S. Huang[1]

[1] Department of Electrical and Computer Engineering,
Coordinated Science Laboratory,
and Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[2] Adobe Systems Incorporated, San Jose, CA 95110, USA
{utariq2,jyang29,huang}@ifp.illinois.edu
http://www.illinois.edu, http://www.adobe.com

Abstract. Expression recognition from non-frontal faces is a challenging research area with growing interest. This paper works with a generic sparse coding feature, inspired from object recognition, for multi-view facial expression recognition. Our extensive experiments on face images with seven pan angles and five tilt angles, rendered from the BU-3DFE database, achieve state-of-the-art results. We achieve a recognition rate of 69.1% on all images with four expression intensity levels, and a recognition performance of 76.1% on images with the strongest expression intensity. We then also present detailed analysis of the variations in expression recognition performance for various pose changes.

1 Introduction

The increasing applications of facial expression recognition, especially those in Human Computer Interaction, have attracted a great amount of research work in this area in the past decade. However, much of the literature focuses on expression recognition from frontal or near-frontal face images [1, 2]. Expression recognition from non-frontal faces is much more challenging. It is also of more practical utility, since it is not trivial in real applications to always have a frontal face. Nonetheless, there are only a handful of works in the literature working with non-frontal faces. There has been experimental evidence in both face recognition and Psychology that non-frontal faces may achieve better recognition performance than frontal ones [2–4]. However, there has not been much effort on a detailed analysis of the effect of large pose variations (both pan and tilt angles) on the expression recognition performance. This paper, apart from achieving state-of-the-art results, also attempts to fill in these gaps.

1.1 Related Works

Most existing works focus on recognizing six basic expressions that are universal and recognizable across different cultures. These include anger (AN), fear (FE), disgust

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 578–588, 2012.

(DI), sad (SA), happy (HA) and surprise (SU) [2]. Some of the notable works in expression recognition focusing on frontal or near-frontal faces include [5–13]. For a comprehensive survey of the works in expression recognition please refer to [1] and [14]. In the following, we shall briefly review the papers that concentrate on non-frontal view facial expression recognition.

The works on non-frontal view expression recognition can be classified based upon the types of features employed. Some works use geometric features, e.g., Hu et al. [15] and Rudovic et al. [16, 17] use displacement or mapping of manually labeled key points to the neutral or frontal face views of the same subject. Whereas, some researchers extract various low-level features (e.g., SIFT) on pre-labeled landmark points and use them for further processing [2]. Some of such works include those by Hu et al. [18] and Zheng et al. [19].

Note that the aforementioned approaches require the facial key-points location information, which needs to be pre-labeled. However, in real applications, key-points need to be automatically detected, which is a big challenge itself in the case of non-frontal faces. To address this issue, there have been some attempts which do not require key-point locations; they rather extract dense features on detected faces[1]. The prominent examples in this category include works by Moore and Bowden [20, 21], Zheng et al. [22] and Tang et al. [23]. Moore and Bowden [20, 21] extract LBP features and its variants from non-overlapping patches. While, Zheng et al. [22] and Tang et al. [23] extract dense SIFT features on overlapping image patches. Zheng et al. [22] use regional covariance matrices for the image-level representation. Tang et al. [23], after dense feature extraction, represent the images with super vectors which are learnt based on ergodic hidden markov models (HMM).

It is worthwhile to mention that the BU3D-FE database [24] has become the de-facto standard for works in this area. Many works use five pan angle views rendered from the database (0°, 30°, 45°, 60°and 90°) [15, 18–21]. However, in real-world situations, we have variations in both pan and tilt angles. Thus, in more recent works [22, 23], people are working with a range of both pan and tilt angles.

Unlike many previous works, our work neither requires key-point localization nor needs a neutral face. We work with 35 views rendered from the BU-3DFE database (combination from 7 pan angles and 5 tilt angles). Unlike [22] and [23], we use all the four expression intensity levels. This work beats the state-of-the-art performance in the same experimental setting as [22] and [23]. Apart from performing better, it also does a significant analysis on the effect of pose and intensity variations on expression recognition results. To our best knowledge, such analysis has not been done before in the literature for such a wide range of pan and tilt angle views. This gives valuable insights to the the multi-view expression recognition problem.

In the following, we first describe the BU-3DFE database used in this work in Section 2. Then we present the generic sparse coding feature from object recognition in Section 3. Multi-view expression recognition experiments are conducted in Section 4. And we present detailed discussions of the results in Section 5. Finally, Section 6 concludes our paper.

[1] Extraction of dense features essentially implies computing features on an entire image region from overlapping or non-overlapping image patches.

2 Database

The database used in this work is the publicly available BU3D-FE database [24]. It has 3D face scan and associated texture images of 100 subjects, each performing 6 expressions at four intensity levels. The facial expressions presented in this database include anger (AN), disgust (DI), fear (FE), happy (HA), sad (SA) and surprise (SU). Each subject also has a neutral face scan. Thus, there are a total of 2500 3D faces. The dataset is quite diverse and contains subjects of both gender with various races. Interested readers are referred to [24] for further details.

We used an openGL based tool from the database creators to render multiple views. We generated views with seven pan angles ($0°$, $\pm 15°$, $\pm 30°$, $\pm 45°$) and five tilt angles ($0°$, $\pm 15°$, $\pm 30°$). These views were generated for each subject with 6 expressions and 4 intensity levels, resulting in an image dataset with $5 \times 7 \times \times 6 \times 4 \times 100 = 84000$ images. Some sample images of a subject in various pan and tilt angles are shown in Figure 1.

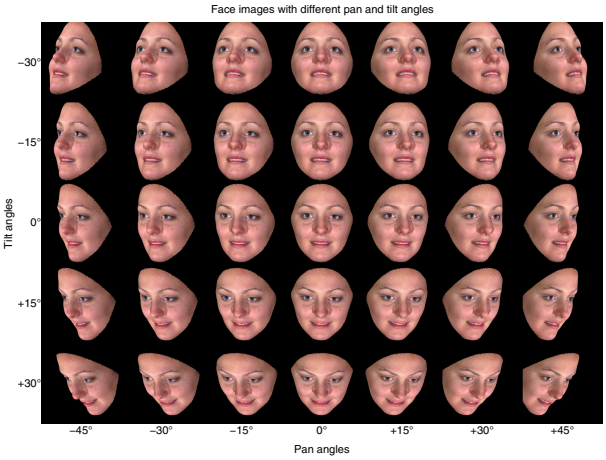


Fig. 1. Rendered facial images of a subject with various pan and tilt angles

3 The Generic Sparse Coding Feature

Recently, much progress has been made in learning mid-level feature representations for image classification [25–28]. These approaches typically follow a common pipeline that consists of three computational modules:

1. *Local feature extraction*: Local descriptors (e.g., raw patches, SIFT or HOG) are extracted from image patches densely sampled from the image to capture the local statistics.
2. *Descriptor encoding*: Each local descriptor is transformed into some code with desired properties (e.g., hard or soft vector quantization [28], LLC [27] or sparse coding [26]), such as compactness, sparseness, or statistical independence.

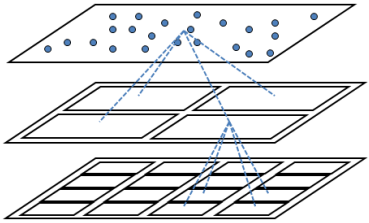


Fig. 2. Spatial pyramid structure for representing the image

3. *Spatial feature pooling*: The codes are then pooled (e.g., averaging [25] or taking the maximum [26, 27]) over different spatial locations across multiple spatial scales to obtain the image level feature representation.

With these mid-level feature representations, state-of-the-art recognition performances have been reported in object recognition and scene classification tasks on benchmark datasets, such as Caltech-101 [29], Caltech-256 [30] and Scene 15 [25]. In this work, we follow the line of mid-level feature learning and apply the image categorization technique for multiple view facial expression analysis. Specifically, we follow the ScSPM work [26] for building the facial image feature representation by max pooling the sparse codes of the local descriptors in a spatial pyramid. The following briefly describes the procedure for building feature representation based on sparse coding.

First, we densely extract local descriptors from the image, and represent the image as sets of local descriptors in a three level spatial pyramid $\mathbf{X} = \left[X_{11}^0, X_{11}^1, X_{12}^1, ..., Y_{44}^2\right]$, where X_{ij}^s is a matrix containing local descriptors from the (i, j)-th spatial block in the s-th spatial scale. As shown in Figure 2, on the s-th image spatial scale, there are 2^s evenly divided spatial blocks in total. Given the dictionary D offline trained [31] from randomly sampled descriptors, we encode the local descriptors into sparse codes by

$$\hat{Z}_{ij}^s = \arg\min_Z \|X_{ij}^s - DZ\|_2^2 + \lambda \|Z\|_1, \tag{1}$$

where the ℓ^1-norm enforces sparsity of the representation and λ controls the sparsity penalty. After encoding all local descriptors into sparse codes, we can similarly represent these sparse codes in the spatial pyramid $\mathbf{S} = \left[\hat{Z}_{11}^0, \hat{Z}_{11}^1, \hat{Z}_{12}^1, ..., \hat{Z}_{44}^2\right]$. The final feature representation is obtained by max pooling over the sparse codes in each spatial block across different spatial scales, i.e.,

$$\beta = \left[\beta_{ij}^s\right], \quad \beta_{ij}^s = \max(|\hat{Z}_{ij}^s|), \tag{2}$$

where β is a concatenation of β_{ij}^s over all (i, j, s) and the "max" operation is performed over each row of $\hat{Z}_{ij}^s$. As shown in [26], max pooling in conjunction with sparse coding works well with linear classifiers, achieving surprisingly good results on image classification tasks. The framework is also backed up by biophysical evidences in the visual cortex (V1) [32].

4 Multi-view Expression Recognition

In this section we detail our work on multi-view expression recognition. We conduct extensive experiments on the 84,000 face images extracted from the BU-3DFE database in 35 views with 4 intensity levels (as outlined in Section 2). The 100 subjects in the database, are randomly divided into five partitions. We do 5-fold cross validation on the 84,000 images and then average the results. In each fold, images from four subject partitions (80% subjects) are used as training and images from the remaining partition (20% subjects) are used as testing. Thus we ensure that the training and testing datasets do not simultaneously contain images from the same subject. We first extract dense SIFT features from images on a regular grid with step size of 3 pixels in both horizontal and vertical directions. Then a randomly sampled subset of these SIFT features is used to train the dictionary, $\mathbf{D} \in \mathbb{R}^{128\times1024}$. This dictionary is then used to sparsely encode the SIFT features extracted from each image, from which max pooling is applied to obtain the image-level representation.

We choose to adopt a 'universal' approach for classification, as in [22] and [23]. For such an approach, in essence, the classifier is trained on the entire training set with all the poses. Thus the 'universal' approach does not require a pose detection step in testing. This not only saves computation but also avoids possible pose estimation errors. We used linear SVM [33] as the classifier. Its computational complexity is $O(n)$ in training and constant in testing. Thus, it can scale up well with large scale datasets.

The overall recognition accuracy for 5-fold cross-validation, averaged across all the subjects, expressions, intensity levels and poses, comes out to be **69.1%**. The respective class confusion matrix is shown in Table 1. The effect of varying expression intensities on expression recognition, averaged for all the poses is plotted in Figure 3. The effects of variations in pan and tilt angles on expression recognition performance are shown in Figure 4. Similarly the effects of variations in pan and tilt angles for various expression intensity levels are shown in Figure 5. Figure 6, on the other hand, shows the effect of the simultaneous variations of pan and tilt angles on the average recognition performance.

Note that no other previous work in the literature experimented with all the expression intensity levels and all the subjects for the aforementioned pan and tilt angles.

Table 1. Classification confusion matrix for over-all recognition performance averaged over all poses and expression intensity levels

Overall classification		Predicted					
		AN	DI	FE	HA	SA	SU
Ground Truth	AN	**64.2**	8.4	4.1	2.2	18.1	3.1
	DI	10.9	**70.1**	5.8	3.9	5.2	4.3
	FE	7.5	9.5	**51.1**	13.7	9.5	8.7
	HA	2.1	4.3	9.4	**81.2**	1.7	1.4
	SA	19.6	5.2	7.2	2.3	**63.4**	2.3
	SU	1.8	3.0	4.7	3.0	2.6	**85.0**

Table 2. Performance comparison with previous works on the strongest expression intensity

Zheng et al [22]	Tang et al. [23]	Ours
68.2%	75.3%	**76.1**%

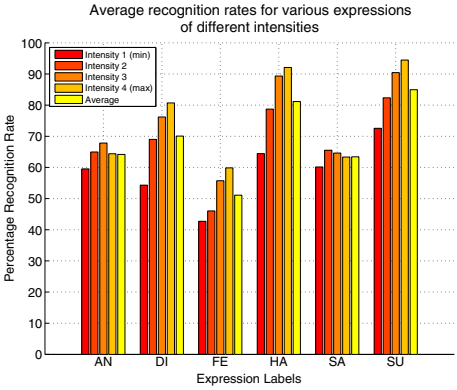
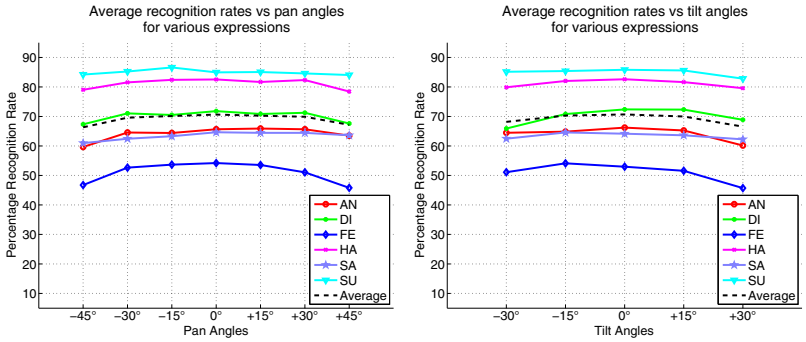


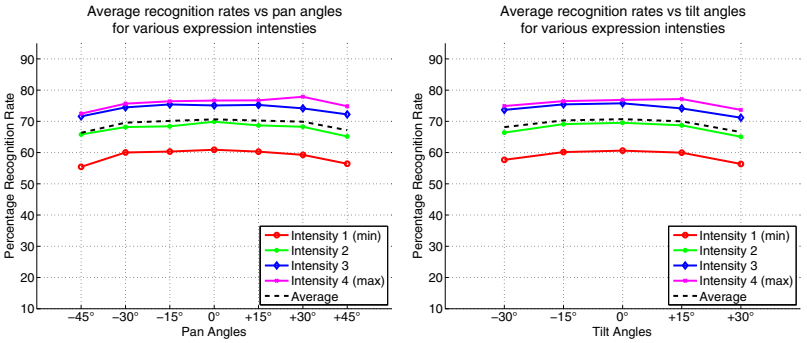
Fig. 3. Recognition performance for various expressions with different intensities



(a) (b)

Fig. 4. Effects of changes in pan (a) and tilt (b) angles on the recognition performance of various expressions

Zheng et al. [22] and Tang et al. [23] follow the same experimental setting of 5-fold cross validation with the same set of pan and tilt angle views, but only focus on the strongest expression intensity level. Hence their image dataset consists of 21,000 images. To fairly compare the performance of this work to those of [22] and [23], we repeat the experiments on the strongest expression intensity level. The comparison of our results with those of [22] and [23] is given in Table 2.



(a) (b)

Fig. 5. Effects of changes in pan (a) and tilt (b) angles on the recognition performance of various expression intensity levels

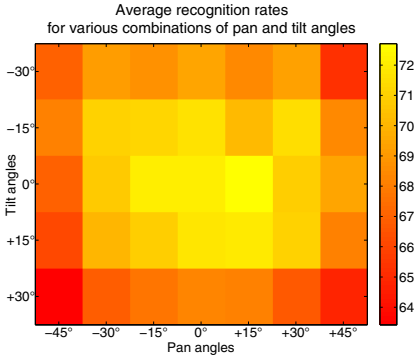


Fig. 6. Effects of changes in both pan and tilt angles on the overall expression recognition performance

5 Discussions

Our work with multi-view expression recognition shows promising results compared to the other state-of-the-art works. Unlike [22] and [23], we experiment with all the four expression intensity levels, which is harder compared to working with just the strongest expression intensity level. This can also be observed from Figures 3 and 5 that the most subtle expressions are the most difficult ones to recognize.

We intend to address a series of questions in this work. For instance, we consider whether the recognition performance is affected in the same manner across different expressions with change in their intensity level. Please refer to Figure 3 for this purpose. It displays the recognition rates for various expressions of different intensities, averaged across all the pan and tilt angle views. It can be observed that the recognition performance of the disgust (DI), fear (FE), happy (HA) and surprise (SU) expressions increases with the increase of expression intensity levels. However, this trend is not

strictly followed for the anger (AN) and sad (SA) expressions. Note that the variations in the recognition performance from the least intense (level 1) to the most intense (level 4) anger and sad expressions is much smaller compared to the other expressions. This may stem from the reason that it may be harder for the subjects to display such expressions in varying intensity levels.

Another point to analyze is how does the variation in pan or tilt angles affects the recognition performance. Please consider Figures 4 and 5 for this purpose. Note that here, the results are averaged across the all the intensity levels, across corresponding tilt angles in Figure 4(a) and across the respective pan angles in Figure 4(b); similarly, the recognition rates are averaged across all the expressions, across corresponding tilt angles in Figure 5(a) and across the respective pan angles in Figure 5(b). One may note that, the average expression recognition performance has its maximum value on 0° pan or tilt angle. There is a slight performance drop up till $\pm 30^\circ$ pan angle (Figures 4(a) and 5(a)) and beyond that the performance drop is more significant. Similarly the average performance drop beyond $\pm 15^\circ$ tilt angle (Figures 4(b) and 5(b)) is more significant. Thus, the frontal and near-frontal views give better average recognition performance.

Then we can ask how do individual expressions respond to change in pan or tilt angles. In Figure 4, one can observe that there are three 'clusters' of curves. The first cluster has only the fear expression performance. It is significantly worse compared to the other expressions for all the variations in pose (in pan or tilt angles). This may be due to greater variation in expressing fear amongst the subjects. The other group of curves giving similar performance are for disgust, anger and sad expressions. And the group of curves giving the best performance is for the happy and surprise expressions. One can note from these figures that the negative expressions (fear, anger, disgust, sad) perform significantly worse than the positive ones, for all the variations in pan and tilt angles. One can also make some interesting observations from the variation in tilt angles in Figure 4(b). For instance, for the fear and anger expressions, -30° tilt angle view give better performance compared to $+30^\circ$ tilt angle view. For the disgust expression, however, the positive tilt angle views give better performance compared to the negative tilt angle views. For the other expressions, the trend is approximately symmetric.

We can also analyze the effect of change in pan and tilt angles on individual intensity levels. One can notice from Figure 5 that the curves for the four intensity levels are more or less parallel, meaning thereby that the individual intensities are affected more or less similarly with the change in pan or tilt angles. The two strongest expression intensity levels perform significantly better than the other two, for all the pan or tilt angle variations. However, in real life situations, the expressions are subtle, in general, thus posing a harder research problem for recognition. The third strongest expression intensity still performs significantly better than the most subtle expression intensity level. The expression intensities, in general, achieve a maximum recognition rate at 0° pan or tilt angle. There is also a significant performance dip at pose angles beyond $\pm 30^\circ$ pan or $\pm 15^\circ$ tilt, for all the expression intensities.

Similarly, to address the variation of both pan and tilt angles on average recognition performance please refer to Figure 6. Please note that each 'box' in this figure, gives the average recognition performance of 2400 images in the corresponding pan and tilt angle view combination. Also note that there is a significant performance decrease be-

yond $\pm 30^\circ$ pan and $\pm 15^\circ$ tilt angle view. Other than that, the performance seems more or less comparable (in the middle). The view with 15° pan and 0° tilt gives the best performance. However, it is very close to the frontal view performance.

6 Concluding Remarks

Our work sets a new state-of-the-art for multi-view facial expression recognition on the BU3D-FE database. We also provide a detailed analysis of variations in expression recognition performance with changes in a range of pan angles, tilt angles and both. Such an in-depth analysis is the first of its kind with such a wide range of pan and tilt angle variations. This can aid in designing various expression recognition systems. Also, unlike many other works, the approach used in this work neither requires any key point detection nor does it need a neutral face, and thus is more suitable for practical purposes.

Acknowledgments. This work was partly supported by Intel under the Avascholar project of the Illinois-Intel Parallelism Center (I2PC), by the Project from committee on science and technology of Chongqing (Grant No. cstc2011ggC0042) and by a research grant from Cisco.

References

1. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 39–58 (2009)
2. Zheng, W., Tang, H., Huang, T.S.: Emotion recognition from non-frontal facial images. In: Konar, A., Chakraborty, A. (eds.) Advances in Emotion Recognition. Wiley (in press, 2012)
3. Bruce, V., Valentine, T., Baddeley, A.: The basis of the 3/4 view advantage in face recognition. Applied Cognitive Psychology 1, 109–120 (1987)
4. Liu, X., Chen, T., Rittscher, J.: Optimal pose for face recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1439–1446 (2006)
5. Pantic, M., Bartlett, M.S.: Machine analysis of facial expressions. In: Delac, K., Grgic, M. (eds.) Face Recognition, pp. 377–416. I-Tech Education and Publishing (2007)
6. Lucey, S., Ashraf, A.B., Cohen, J.F.: Investigating spontaneous facial action recognition through aam representations of the face. In: Delac, K., Grgic, M. (eds.) Face Recognition, pp. 275–286. I-Tech Education and Publishing (2007)
7. Chang, Y., Hu, C., Feris, R., Turk, M.: Manifold based analysis of facial expression. Image and Vision Computing 24, 605–614 (2006)
8. Valstar, M.F., Gunes, H., Pantic, M.: How to distinguish posed from spontaneous smiles using geometric features. In: Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI 2007, pp. 38–45 (2007)
9. Anderson, K., McOwan, P.W.: A real-time automated system for the recognition of human facial expressions. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 36, 96–105 (2006)

10. Valstar, M., Pantic, M., Patras, I.: Motion history for facial action detection in video. In: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 635–640 (2004)
11. Dhall, A., Asthana, A., Goecke, R., Gedeon, T.: Emotion recognition using phog and lpq features. In: 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011, pp. 878–883 (2011)
12. Zeng, Z., Tu, J., Pianfetti Jr., B.M., Huang, T.S.: Audio-visual affective expression recognition through multistream fused hmm. IEEE Transactions on Multimedia 10, 570–577 (2008)
13. Tian, Y., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 97–115 (2001)
14. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognition 36, 259–275 (2003)
15. Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., Huang, T.S.: A study of non-frontal-view facial expressions recognition. In: Proceedings - International Conference on Pattern Recognition (2008)
16. Rudovic, O., Patras, I., Pantic, M.: Regression-based multi-view facial expression recognition. In: Proceedings - International Conference on Pattern Recognition, pp. 4121–4124 (2010)
17. Rudovic, O., Patras, I., Pantic, M.: Coupled Gaussian Process Regression for Pose-Invariant Facial Expression Recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 350–363. Springer, Heidelberg (2010)
18. Hu, Y., Zeng, Z., Yin, L., Wei, X., Zhou, X., Huang, T.S.: Multi-view facial expression recognition. In: 2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008 (2008)
19. Zheng, W., Tang, H., Lin, Z., Huang, T.S.: A novel approach to expression recognition from non-frontal face images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1901–1908 (2009)
20. Moore, S., Bowden, R.: The effects of pose on facial expression recognition. In: British Machine Vision Conference (2009)
21. Moore, S., Bowden, R.: Local binary patterns for multi-view facial expression recognition. Computer Vision and Image Understanding 115, 541–558 (2011)
22. Zheng, W., Tang, H., Lin, Z., Huang, T.S.: Emotion Recognition from Arbitrary View Facial Images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 490–503. Springer, Heidelberg (2010)
23. Tang, H., Hasegawa-Johnson, M., Huang, T.: Non-frontal view facial expression recognition based on ergodic hidden markov model supervectors. In: 2010 IEEE International Conference on Multimedia and Expo., ICME 2010, pp. 1202–1207 (2010)
24. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, vol. 2006, pp. 211–216 (2006)
25. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178 (2006)
26. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, pp. 1794–1801 (2009)
27. Wang, J., Yang, J., Yu, K., Gong, Y., Huang, T.: Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
28. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)

29. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106, 59–70 (2007)
30. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007)
31. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: NIPS, pp. 801–808 (2007)
32. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 994–1000 (2005)
33. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)