Facial Expression Analysis Based on High Dimensional Binary Features

Samira Ebrahimi Kahou(✉), Pierre Froumenty, and Christopher Pal

École Polytechique de Montréal, Université de Montréal, Montréal, Canada
{samira.ebrahimi-kahou,pierre.froumenty,christopher.pal}@polymtl.ca

Abstract. High dimensional engineered features have yielded high performance results on a variety of visual recognition tasks and attracted significant recent attention. Here, we examine the problem of expression recognition in static facial images. We first present a technique to build high dimensional, $\sim$ 60k features composed of dense Census transformed vectors based on locations defined by facial keypoint predictions. The approach yields state of the art performance at 96.8% accuracy for detecting facial expressions on the well known Cohn-Kanade plus (CK+) evaluation and 93.2% for smile detection on the GENKI dataset. We also find that the subsequent application of a linear discriminative dimensionality reduction technique can make the approach more robust when keypoint locations are less precise. We go on to explore the recognition of expressions captured under more challenging pose and illumination conditions. Specifically, we test this representation on the GENKI smile detection dataset. Our high dimensional feature technique yields state of the art performance on both of these well known evaluations.

Keywords: Facial expression recognition · Smile detection · High-dimensional feature · Census transformation · Deep learning · GENKI · CK+

1 Introduction

Local binary patterns (LBPs) [1] are well known texture descriptors that are widely used in a number of applications. LBP features have been found to be particularly effective for face related applications [2]. As an example, high dimensional features based on LBPs have yielded highly competitive results on the well known Labeled Faces in the Wild face verification evaluation [3,4].

We are interested here in recognizing facial expressions in static imagery. Facial expression analysis can be a particularly challenging problem, especially when using imagery taken under "in the wild" conditions as illustrated by the recent Emotion Recognition in the Wild Challenge [5]. Here we examine both controlled environment facial expression analysis and an "in the wild" problem through evaluations of our proposed method using the Extended Cohn-Kanade (CK+) database [6,7] and the GENKI-4K smile detection evaluation. The CK+ database is a widely used standard evaluation dataset containing acted

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 135–147, 2015.
DOI: 10.1007/978-3-319-16181-5_10

expressions. The expressions to be recognized are based on Ekman's six basic universal categories of: happiness, sadness, surprise, fear, anger, and disgust [8]. The GENKI-4K [9] dataset contains comparatively low resolution images harvested from the web.

We provide a number of technical contributions in this paper. First, we provide a formulation of high dimensional features that is different from other standard formulations. Our descriptor is a high dimensional feature vector in which each dimension consists of the bits derived from Census transformation. Features are obtained based on image patches centered on facial keypoints. We use a slight variant of LBPs known as the Census transform [10]. To the best of our knowledge this representation yields the highest known performance on CK+ using the same evaluation criteria as in [7].

We go on to adapt our technique to be more robust to inaccurately localized facial keypoints using a multi-resolution technique and local Fisher discriminant analysis (LFDA) [11] - a recently proposed extension to the widely used Fisher discriminant analysis technique. The issue of keypoint localization accuracy is particularly important when turning to the problem of recognition in the wild, but even in controlled environments there are well known degradations in performance when per subject keypoint training data is not used to fit a facial keypoint model. Turning to the problem of smile recognition using in the wild GENKI imagery, it is much harder to detect a large number of keypoints due to the quality and variability of the imagery. For the GENKI evaluation in particular we are however able to detect five keypoints reliably. Adapting our method to this setting, here again our proposed method yields the highest known performance of which we are aware on this well known evaluation.

The remainder of this manuscript is structured as follows: We provide a brief review of some other relevant work in section 2, but also discuss other relevant work throughout this document. In section 3 we present our novel feature extraction technique based on high dimensional binary features, multi-scale patches and discriminative dimensionality reduction. In section 4 we benchmark our high dimensional feature vector technique using CK+, examining experimentally the issue of facial landmark prediction quality, its impact on prediction performance and our motivations for extending our basic formulation to include multi-scale analysis and discriminative dimensionality reduction. We then provide our experiments on GENKI-4K, where we also compare directly with a state of the art convolutional neural network technique that does not rely on keypoints. We provide conclusions and additional discussion in section 5.

2 Other Relevant Work

A number of modern, state of the art approaches to expression detection are based on handcrafted features, such as: Local binary patterns or LBP features [1], Histograms of oriented gradients or HOG features [12], or Lowe's Scale-invariant feature transform (SIFT) descriptors [13]. For example, the influential work of Shan et al. [14] studied histograms of LBP features for facial expression

recognition. They introduced Boosted-LBP by using AdaBoost [15] for feature selection. Their experiments showed that LBP features are powerful for low resolution images. Dahmane et al. [16] built face representation based on histograms of HOG features from dense grids. Their representation followed by nonlinear SVM outperforms an approach based on *uniform* LBP. Other work has used SIFT features for facial expression analysis [17], yielding competitive results on CK+.

Techniques based on convolutional neural networks have also yielded state of the art performance for the task of emotion recognition, including top performing results on competitive challenges [18–20]. The CK+ data and classification tasks were introduced in Lucey et al. [7]. They provided both the additional facial examples that were used to extend the original Cohn-Kanade (CK) dataset of [6], yielding the combined dataset known as CK+ as well a number of experimental analyses. They provided a variety of baseline experiments and a state of the art result at the time in which they combine a landmark based representation (SPTS) and appearance features both before and after shape normalization using landmarks, which they refer to as CAPP features. They combine two different classifiers for landmarks and appearance using a logistic regression on the outputs of the classifiers. This procedure yields their best result with an average accuracy of 83.33%.

Jeni et al. [21] used shape only information for expression recognition experiments with CK+; however, they removed the sequences with noisy landmarks. The work of Sikka et al. [17] compares the performance for a variety of techniques on the CK+ expression recognition task, including the well known uniform LBP histogram technique in [14] which they state yields 82.38% ±2.34 average accuracy. They state that their own bag of words architecture yields 95.85% ±1.4 average per subject accuracy using a leave one subject out evaluation protocol. Other work has also explored the problem of smile detection using the GENKI-4K data. Jain et al. [22] report 92.97% accuracy using multi-scale gaussian derivatives combined with an SVM, but they removed ambiguous cases and images with serious illumination problems (423 removed faces). Shan et al. [23] report 89.70% ±0.45 using an Adaboost based technique; however, they manually labeled eye positions which is not practical for many applications. Liu et al. [24] report 92.26% ±0.81 accuracy and also provide the splits used for their evaluation. We therefore use their splits in our evaluation below to permit our technique to be directly compared to their results.

3 Our Models

In this section, we present our technique which we show later is capable of obtaining state of the art results on both the CK+ and GENKI evaluations. We also present a deep neural network approach for expression recognition that we shall use for additional comparisons on the GENKI evaluation.

3.1 High Dimensional Engineered Features

Our high dimensional feature approach is conceptually simple. We extract a form of local binary pattern known as the Census transform for each pixel found within small image patches, each centered on a facial keypoint. Unlike previous work which typically creates histograms of LBPs, here we create our feature vector by concatenating the bits for each pixel of the image patch into a binary vector. We also concatenate bits obtained from patches extracted at multiple scales centered on the keypoints. As far as we are aware this is different from previous uses of LBP techniques which have relied on histogramming operations. This high dimensional binary feature vector is then projected into a smaller dimensional space via principal component analysis (PCA), followed by a recently proposed variation of multiclass Fisher Discriminant Analysis (FDA) known as local FDA or LFDA [11]. The resulting vector is then used within a Support Vector Machine (SVM). There are a number of choices to be made throughout this processing and classification pipeline and we search over key subsets of these choices using cross validation techniques. We discuss the different steps of our procedure in more detail below.

The Census Transform. The Census transform [10] is computed as follows. If $\mathbf{p} = \{u, v\}$ is the index of a pixel and $I(\mathbf{p})$ is its intensity, define $\xi(\mathbf{p}, \mathbf{p}') = 1, \text{if } I(\mathbf{p}') < I(\mathbf{p})$; otherwise $\xi(\mathbf{p}, \mathbf{p}') = 0$. The Census transform simply concatenates the bits obtained from comparisons using a fixed ordering of pixels within spatial neighborhood around the pixel. The result is a bit string with ones representing the pixels that are less than the value of the central pixel. Using $\bigotimes$ to denote concatenation, the census transform for the pixel at location $\mathbf{p} = \{u, v\}$ is simply

$$I^c(\mathbf{p}) = \bigotimes_{j=-n}^{n} \bigotimes_{i=-m}^{m} \xi(I(u, v), I(u + i, v + j)), \tag{1}$$

typically computed using a window of size $(2m+1) \times (2n+1)$. In other words, for a given image patch the CT simply compares each pixel with the center pixel. If its value is greater than the center pixel's value it assigns 0 and 1 otherwise. Common window sizes are 3 and 5. In our experiment, we used 3 as the window size which allows the information to be stored in an 8-bit binary number if desired. The ability to store such descriptors using a binary encoding means that even if our descriptor is of extremely high dimension the information can be stored in a highly compact format. Various other operations using these types of binary descriptors can also be implemented very efficiently.

Keypoint Guided Feature Extraction. As outlined above, we construct our descriptors by cropping small patches out of the larger facial image, applying the Census transform to each pixel for each patch and concatenating the resulting bits into a high dimensional vector. In our experiment below, each scale yields 19,992 features for CK+ and 4,312 for GENKI, due to the different number of keypoints produced by different methods. Patches are extracted centered on

each landmark, excluding the face contour. The patches have two parameters that are optimized by cross validation: patch width, defined in proportion to face size and the patch size. The optimal values for our initial CK+ experiment for example were 2/5ths of the face size and 9 pixels in width respectively. Each cropped patch is also resized before computing the Census transform allowing us to control both the dimensionality and the size or spatial extent of the patch separately. We will also present experiments where we extend this approach by extracting patches at each keypoint at three different scales. Depending on the experiment this produces about 60k features.

To obtain keypoints there are a variety of automated placement techniques which can be applied depending on the circumstances. For example, the CK+ dataset comes with landmark positions that were estimated by fitting an Active Appearance Model (AAM) [25]. AAMs can yield state of the art performance when labeled keypoints have been provided to train models for each subject of interest. For our first set of experiments we use the landmarks provided with the CK+ data. However, AAMs yield poor performance when per subject training data is unavailable. In many real world situations it is impractical to label keypoints for each subject. For this reason there has been a great deal of recent activity focused towards improving alternative approaches that are not identity dependent. For our second CK+ experiments we use the structured max margin technique of [26]. For GENKI experiments we use the convolutional neural network cascade technique in [27].

Dimensionality Reduction. As we shall see in our experimental work, our high dimensional Census feature technique can yield encouraging results on the CK+ evaluation. However, Working with high dimensional vectors can be impractical for many applications. We therefore employ a two phase dimensionality reduction procedure based on an initial projection using PCA followed by LFDA [11]. LFDA obtains a discriminative linear projection matrix through minimizing an objective function of the same form as FDA. The underlying problem is therefore also equivalent to solving a generalized eigenvalue problem. More precisely, a projection matrix $\mathbf{M}$ is obtained from

$$\arg\max_{\mathbf{M}} \mathrm{Tr}\Big\{(\mathbf{M}^T\mathbf{S}_W\mathbf{M})^{-1}\mathbf{M}^T\mathbf{S}_B\mathbf{M}\Big\}, \tag{2}$$

where there are $i = 1, \ldots, n$ feature vectors $\mathbf{x}_i$ with class labels C_i, given by $c = 1, \ldots, n_c$ class indices, and

$$\mathbf{S}_W = \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \tag{3}$$

which defines a *local* within-class scatter matrix using

$$\mathbf{W}_{i,j} = \begin{cases} \mathbf{A}_{i,j} & C_i = C_j = c \\ 0 & C_i \neq C_j, \end{cases} \tag{4}$$

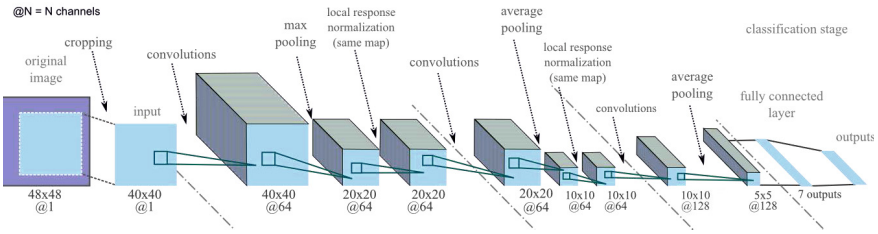

Fig. 1. The architecture of the convolutional neural network used in our experiments

and a *local* between-class scatter matrix defined by

$$\mathbf{S}_B = \frac{1}{2} \sum_{i,j=1} \mathbf{B}_{i,j} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \tag{5}$$

where

$$\mathbf{B}_{i,j} = \begin{cases} \mathbf{A}_{i,j} \left(\frac{1}{n} - \frac{1}{n_c}\right) & C_i = C_j = c \\ \frac{1}{n} & C_i \neq C_j, \end{cases} \tag{6}$$

and for both types of local scatter matrix one uses an affinity matrix $\mathbf{A}$ defined, for example by

$$\mathbf{A}_{i,j} = \exp(\|\mathbf{x}_i - \mathbf{x}_j\|^2). \tag{7}$$

3.2 A Deep Convolutional Neural Network Approach

We shall also compare with a deep convolutional neural network approach to expression recognition based on the framework presented in [28] which was used to win the recent ImageNet challenge. The particular architecture we used here for expression recognition is shown in Fig. 1. A similar deep neural network architecture and training approach for expression recognition in the wild was used in [18] to win the recent Emotion Recognition in the Wild Challenge [29] where the goal was to predict expressions in short clips from movies. In [18] the deep network was only trained on the Toronto Face Database TFD [30] - a large set of different standard expression datasets including Cohn-Kanade and a dataset mined from Google image search results [31] containing 35,887 images tagged with the corresponding emotion categories. In contrast for our GENKI experiments here we do not use additional training data.

Since this implementation and architectural variants of it have won a number of competitive challenges we believe the approach is representative of a state of the art deep neural network approach for expression recognition with wild imagery. We therefore use it here to provide a point of comparison for our GENKI experiments.

4 Experiments and Results

Here we provide two sets of experiments. First, we present experiments using the standard CK+ evaluation and our high dimensional feature technique. We examine in particular the sensitivity of our approach to keypoint localization quality, the results of which partly motivated the development of the multi-resolution extensions to our basic approach - making it more robust to inaccurate keypoints. We then present results for the smile detection problem using the GENKI-4K dataset, comparing with the deep convolutional neural network approach presented above.

For our last CK+ experiment with noisy keypoints and for our GENKI experiment we apply our full approach in which multi-scale patches are extracted and feature descriptors are reduced in dimensionality using LFDA. The dimensionality reduction is applied on a per patch basis. For PCA we search in the region of dimension reductions that capture 95% of the variance. For LFDA we search in the region of reductions that reduce the final output to 5-20% of the original dimensionality. It is interesting to note that the multi-scale descriptor has about 60k dimensions for our CK+ experiment and is reduced to about 6k dimensions.

4.1 Experiments on CK+

The CK+ database [6,7] is a widely used benchmark for evaluating emotion recognition techniques. It is perhaps more precise to characterize the emotion recognition task using CK+ as facial expression recognition since the majority of sequences were acted. The evaluation includes image sequences with 6 basic expressions. Each sequence starts with a neutral face and ends with an image showing the most exaggerated variation of a given expression. CK+ has large variation in gender, ages and ethnicity. The database consists of 593 image sequences of 123 different subjects and covers both spontaneous and acted expressions. Only one expression "Happy" is spontaneous and it's because some actors smiled during video recordings. CK+ dataset includes labels for expressions, landmarks and labels for the Facial Action Coding System (FACS). We focus here on the expression recognition task.

We use the CK+ data in our work to benchmark and evaluate our approach on a standard dataset before tackling data that is of principal interest to our work in which expressions are exhibited by subjects in natural and spontaneous situations. We begin by placing our high dimensional feature technique in context with the state of the art by showing the complete result of Lucey et al.'s top performing SPTS+CAPP technique discussed in more detail in our literature review [7]. To evaluate our technique performance when high precision keypoints are not available we then show the impact of using realistic keypoint predictions from the keypoint predictor in [26].

High Dimensional Binary Feature Vectors. For our first experiment here we created a high dimensional binary vector from densely sampled keypoint locations as discussed in section 3. We give the resulting vector to a linear support vector machine using the implementation in [32]. We perform leave one subject

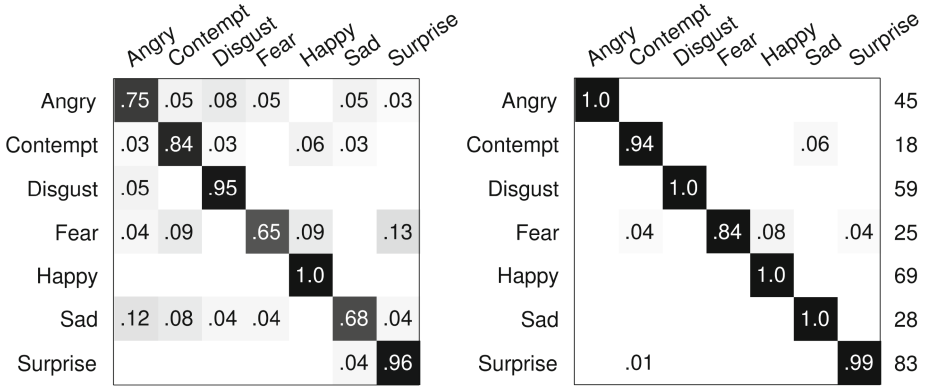

Fig. 2. (**left**) A confusion matrix for expression detection from the SPTS + CAPP result of Lucey et al. [7]. The average per class recognition rate was 83.3%. The matrix is row normalized as in [7]. (**right**) The confusion matrix for expression detection on CK+ using our high dimensional binary features. The average per class accuracy is 96.8%. The overall average accuracy is 98.2%. We give the number of examples per class in the column on the right.

out experiments and optimize hyperparameters using an inner cross validation procedure within the training set. Results are shown in Fig. 2 (right). We are aware of no other published result with higher performance. The best result of which we are aware on CK+ also gives an accuracy of 96% [21]; however, they exclude five subjects from their evaluation. Table 1 provides comparison of our results to other methods.

The Impact of Noisy Keypoints. As we have discussed, in many practical situations it is not possible to obtain highly accurate keypoints such as those possible when using an AAM trained on labeled examples of each subject. For this reason we perform the same experiment above but using the keypoint detector of [26]. As seen in Fig. 3 (left), there is a drop in performance (i.e. 90.0% vs 96.8%), but it is not as dramatic as one might expect due in part to the improved quality for subject independent keypoint predictions afforded by [26].

The Impact of Multiscale Patches. We then evaluated the hypothesis that the use of multiscale patches centered on each keypoint could make the approach more robust to keypoint localization errors. The result of this experiment is shown in Fig. 3 (right). While we cannot recover the original performance, we do see a slight boost in performance over the original single resolution technique.

4.2 Smile Detection Experiments

The GENKI-4K dataset [9,33] consists of 4,000 facial images labelled with pose and smile content. The images are relatively low resolution and in jpeg for-

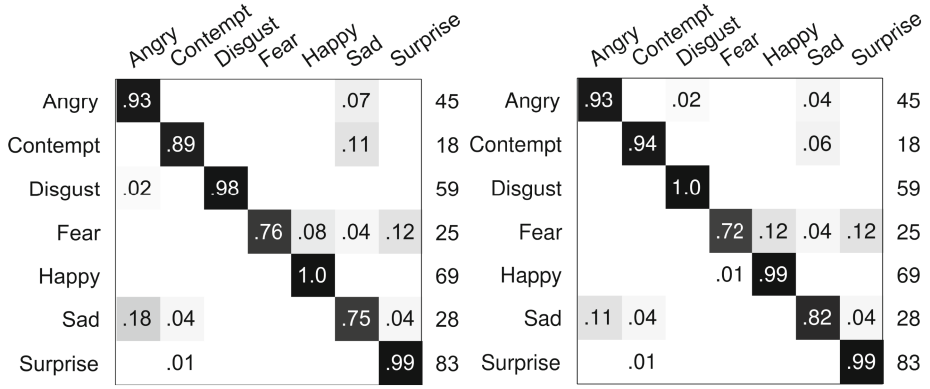

Fig. 3. (**left**) Confusion matrix for expression detection on CK+ using our high dimensional binary features, but based on less accurate keypoints. The average per class accuracy is 90.0%. The overall average accuracy is 93.4%. (**right**) The average per class accuracy when using our multi-scale strategy increases to 91.3% as does the average accuracy, which increases to 94.5%.

Table 1. CK+ Experiments: Comparison and summary

Method	%
Lucy et al. (2010) [average accuracy] using a landmark based representation and appearance features [7]	83.33
Sikka et al. (2012) [average accuracy] LBP histogram architecture [14,17]	82.38
Sikka et al. (2012) [average per subject accuracy] bag of words [17]	95.85
Our technique [average accuracy], accurate keypoints	96.8
Our technique [average class accuracy], accurate keypoints	98.2
Our technique [average accuracy], noisy keypoints	94.5

mat. This dataset has large variations in pose, illumination and ethnicity. We extracted faces from the original images using a combination of the opencv's Haar cascade face detection [34] and the convolutional neural network cascade of [27]. Where these detectors failed to detect any face, we just kept the original.

The resolution of imagery in this dataset was such that we were only able to detect a set of 5 keypoints reliably for our high dimensional feature technique. In order to cover the whole face we computed 6 more points located between eyes, mouth corners and the nose. We provide a comparison with the convolutional neural network (Convnet) architecture discussed in section 3.2, which does not rely on keypoints. For both our high dimensional feature technique and our ConvNet experiments we split the dataset into 4 equal folds using the precise splits defined in [24].

For each experiment with the convolutional neural network, we used random cropping with a 4-pixel border for 48×48 images. Also images were flipped

horizontally with a probability of 0.5 at each epoch. The model with no pre-processing yielded 91.5% 1-fold accuracy. We explored preprocessing with isotropic smoothing [35,36], yielding 91.5%, and histogram equalization on the grayscale imagery, which yielded 91.7%. From these experiments we found that these preprocessing options did not alter performance in a substantial way. We therefore ran a full four fold experiment using grayscale faces with no pre-processing at 96×96 pixel resolution, which yielded 92.97% ±0.71 accuracy.

Using our complete high dimensional feature technique consisting of both the initial feature construction and including the use of multi-resolution patches and the local fisher discriminant analysis step, followed by the application of an SVM with radial basis function kernel for the final classification, we were able to achieve 93.2% ±0.92 average accuracy. We place our results here in context with prior work in Table 2.

Table 2. GENKI-4K Experiments (Accuracies)

Method	%
Shan et al. (2012), using an Adaboost based technique; however, they manually labeled eye positions [23]	89.70
Jain et al. (2013), using multi-scale Gaussian derivatives combined with an SVM; however, they removed ambiguous cases & images with serious illumination problems (423 faces removed) [22]	92.97
Liu et al. (2013), using HOG features and SSL [24]	92.29
Liu et al. (2013), with only labeled data	91.85
Our ConvNet at 48 × 48 pixel resolution (no preprocessing)	91.5
Our ConvNet at 96 × 96 pixel resolution (±0.71)	93.0
Our high dimensional LBP technique (±0.92)	93.2

5 Final Conclusions and Discussion

It is important to emphasize that traditionally LBP based techniques have used histogramming operations to create underlying feature representations. In contrast, in our work we do not compute histograms and use bits directly. For example previous work [17] has given an accuracy of 82.38% on CK+ for a traditional LBP approach using histograms computed on grid locations defined by a face bounding box using a boosted SVM classification approach. Since we use LFDA to learn a discriminative reduced dimensionality space, our work thus also blurs the lines between traditional notions of engineered feature representations and learned representations. Since we use LBP-like descriptors defined by keypoint locations, in a sense we also blur the lines between keypoint vs. non-keypoint based representations. We hope that our results here will help motivate further work exploring other alternative approaches using LBP descriptors as underlying input representations.

Acknowledgments. We thank Yoshua Bengio, Pascal Vincent, Ian Goodfellow, David Warde-Farley, Mehdi Mirza and David Krueger for helpful discussions. We also thank NSERC and Ubisoft for their support.

References

1. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (1996)
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006)
3. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 3025–3032. IEEE Computer Society, Washington, DC (2013)
4. Lu, C., Tang, X.: Surpassing human-level face verification performance on LFW with gaussianface. In: Technical report arXiv:1404.3840 (2014)
5. Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., Bartlett, M.: Multiple kernel learning for emotion recognition in the wild. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI 2013, pp. 517–524. ACM, New York (2013)
6. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
7. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
8. Eisert, P., Girod, B.: Analyzing facial expressions for virtual conferencing. IEEE Computer Graphics and Applications, 70–78 (1998)
9. GENKI-4K: The MPLab GENKI Database, GENKI-4K Subset. http://mplab.ucsd.edu
10. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 801, pp. 151–158. Springer, Heidelberg (1994)
11. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. The Journal of Machine Learning Research **8**, 1027–1061 (2007)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893 (June 2005)
13. Lowe, D.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
14. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image Vision Comput. **27**(6), 803–816 (2009)

15. Freund, Y., Schapire, R.E.: A desicion-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, Paul M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
16. Dahmane, M., Meunier, J.: Emotion recognition using dynamic grid-based HoG features. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pp. 884–888 (March 2011)
17. Sikka, Karan, Wu, Tingfan, Susskind, Josh, Bartlett, Marian: Exploring bag of words architectures in the facial expression domain. In: Fusiello, Andrea, Murino, Vittorio, Cucchiara, Rita (eds.) ECCV 2012 Ws/Demos, Part II. LNCS, vol. 7584, pp. 250–259. Springer, Heidelberg (2012)
18. Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, c., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., Mirza, M., Jean, S., Carrier, P.L., Dauphin, Y., Boulanger-Lewandowski, N., Aggarwal, A., Zumer, J., Lamblin, P., Raymond, J.P., Desjardins, G., Pascanu, R., Warde-Farley, D., Torabi, A., Sharma, A., Bengio, E., Côté, M., Konda, K.R., Wu, Z.: Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI 2013, pp. 543–550. ACM, New York (2013)
19. Tang, Y.: Deep learning using support vector machines. CoRR abs/1306.0239 (2013)
20. Rifai, S., Bengio, Y., Courville, A., Vincent, P., Mirza, M.: Disentangling factors of variation for facial expression recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 808–822. Springer, Heidelberg (2012)
21. Jeni, L., Takacs, D., Lorincz, A.: High quality facial expression recognition in video streams using shape related information only. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2168–2174 (November 2011)
22. Jain, V., Crowley, J.: Smile detection using multi-scale gaussian derivatives. In: 12th WSEAS International Conference on Signal Processing, Robotics and Automation, Cambridge, United Kingdom (February 2013)
23. Shan, C.: Smile detection by boosting pixel differences. Trans. Img. Proc. **21**(1), 431–436 (2012)
24. Liu, M., Li, S., Shan, S., Chen, X.: Enhancing expression recognition in the wild with unlabeled reference data. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part II. LNCS, vol. 7725, pp. 577–588. Springer, Heidelberg (2013)
25. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60**(2), 135–164 (2004)
26. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886 (June 2012)
27. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 3476–3483. IEEE Computer Society, Washington, DC (2013)
28. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1106–1114 (2012)
29. Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion recognition in the wild challenge 2013. In: ACM ICMI (2013)

30. Susskind, J., Anderson, A., Hinton, G.: The toronto face database. Technical report, UTML TR 2010–001, University of Toronto (2010)
31. Carrier, P.L., Courville, A., Goodfellow, I.J., Mirza, M., Bengio, Y.: FER-2013 Face Database. Technical report, 1365, Université de Montréal (2013)
32. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011). http://www.csie.ntu.edu.tw/cjlin/libsvm
33. Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., Movellan, J.: Toward practical smile detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(11), 2106–2111 (2009)
34. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I-511–I-518 (2001)
35. Štruc, V., Pavešić, N.: Gabor-based kernel partial-least-squares discrimination features for face recognition. Informatica **20**(1), 115–138 (2009)
36. Štruc, V., Pavešić, N.: Photometric normalization techniques for illumination invariance, pp. 279–300. IGI-Global (2011)
37. Dollár, P.: Piotr's Image and Video Matlab Toolbox (PMT). http://vision.ucsd.edu/pdollar/toolbox/doc/index.html