Emotion Distribution Recognition from Facial Expressions

Ying Zhou, Hui Xue, Xin Geng*
MOE Key Laboratory of Computer Network and Information Integration
School of Computer Science and Engineering
Southeast University, Nanjing, China
{zhouying1, hxue, xgeng}@seu.edu.cn

ABSTRACT

Most existing facial expression recognition methods assume the availability of a single emotion for each expression in the training set. However, in practical applications, an expression rarely expresses pure emotion, but often a mixture of different emotions. To address this problem, this paper deals with a more common case where multiple emotions are associated to each expression. The key idea is to learn the specific description degrees of all basic emotions for each expression and the mapping from the expression images to the emotion distributions by the proposed emotion distribution learning (EDL) method. Experimental results show that EDL can effectively deal with the emotion distribution recognition problem and perform remarkably better than the state-of-the-art multi-label learning methods.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

Emotion distribution learning; facial expression recognition; description degree

1. INTRODUCTION

The increasing applications of facial expression recognition,especially those in human computer interaction, have attracted a great amount of interests in the past decade. There are many single-emotion learning methods that have been used, for example, neural-network-based methods [8], support vector machine (SVM) [14] and hidden markov model (HMM) [15]. As a single-emotion problem, satisfactory recognition accuracy has been reached in the previous research [10].

However, according to Plutchik's wheel of emotions theory [11], there are a small number of basic emotions, and all the other emotions occur as combinations, mixtures or compounds of the basic emotions and can exist in varying degree of intensity or levels of arousal. So, the single-label learning methods used to recognize one basic emotion for each expression may not be suitable for the real-life facial expression recognition applications, where an expression rarely expresses only one basic emotion.

Fig. 1 gives three examples from the s-JAFFE database. The s-JAFFE database contains 213 facial expression images. Each image was rated by 60 persons. A five-level score is applied to each sample on the 6 basic emotions (happiness, sadness, surprise, anger, disgust and fear [3]), where 5 represents highest emotion intensity and 1 represents lowest emotion intensity. Then the average scores for each emotion on each expression image were obtained. From Fig. 1, we can see that each emotion in one expression image corresponds to a positive score, no matter it is high or low, which agrees with the Plutchik's theory that an expression can be viewed as a mixture of basic emotions [11].

Rudovic et al [13] proposed a multi-output Laplacian Dynamic Ordinal Regression method, which can output the probability of each emotion label and estimate intensity. However, it assumes only one correct emotion label for each expression and outputs the emotion with the highest probability as a result, which cannot match the mixture emotion cases in real life.

If each basic emotion is considered as a label, multi-label learning (MLL) [20] can be used to describe each expression image with several relevant emotions. However, MLL cannot learn the intensity of each emotion. MLL might first select a threshold, then the emotions with scores higher than the threshold are labeled as relevant emotions, while the others are labeled as irrelevant emotions. As in Fig. 1, the relevant emotions are set as +1 and the irrelevant emotions are set as -1. In such way, all the relevant emotions are deemed as equally important. As a result, the important information about the emotion intensity is lost.

To address the above problem, we propose an emotion distribution learning (EDL) algorithm in this paper. Different from the above algorithms, EDL considers that expressions are often the mixtures of basic emotions and allows different intensities in each emotion. Concretely, EDL uses a description degree d_x^y as a numerical indicator to measure the relationship of the emotion y to the expression x and indicate the relative emotion intensity. The sum of description degrees of all emotions for one expression is 1, meanwhile

*Corresponding author.

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806328.

each description degree is between 0 and 1. For a particular expression, the description degrees of all the emotions have a similar data form to probability distribution. As a result, we termed it as emotion distribution. Then, a learning process is invoked to learn the mapping from the expression image to the emotion distribution, which aims to minimize the difference between the true distribution and the predicted distribution. Furthermore, both the single-label learning and MLL can be considered as special cases of EDL in facial expression recognition.

The rest of the paper is organized as follows. Section 2 proposes the method of emotion distribution learning. In section 3, the experimental results are reported. Finally, several conclusions are drawn in Section 4.

2. EMOTION DISTRIBUTION LEARNING

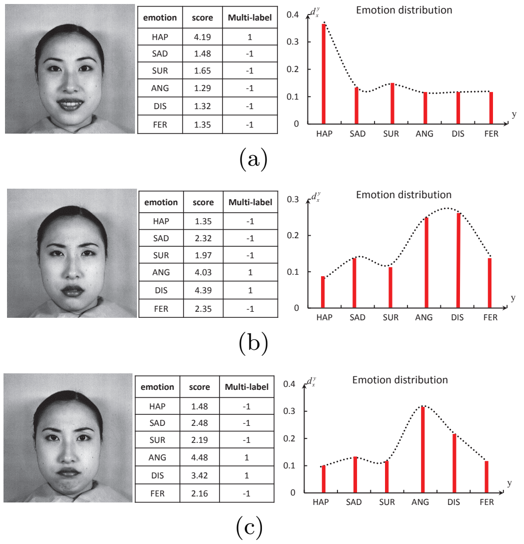
2.1 Emotion Distribution

The goal of EDL is to learn a mapping from an expression image space $\mathcal{X} = \mathbb{R}^m$ to the emotion distribution over a finite set of labels $\mathcal{Y} = \{y_1, y_2, ...y_c\}$. Each label represents one of the basic emotions. As discussed in Section 1, facial expression is often composed of one or more emotions, and each emotion has its own intensity. We use the description degree d_x^y to indicate the intensity of emotion y for the facial expression x, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Normalize the emotion intensity to make $d_x^y \in [0, 1]$, and $\sum_y d_x^y = 1$ to constitute the emotion distribution.

Fig. 1 shows some typical examples from the s-JAFFE database together with their multi-labels and emotion distributions. The threshold is chosen as 2.5 in MLL scenarios to illustrate the difference between MLL and EDL more clearly. The emotion distribution is represented by a curve. There are six values at the horizontal axis labeled by the six basic emotions. The values at the vertical axis represent the description degrees of each emotion. For Fig 1(a), the description degree of happiness is the highest, and all the other five emotions' description degrees are significantly lower. Happiness can be considered as the only relevant emotion. Single-emotion learning, MLL and EDL algorithms can all deal with such case well. In Fig 1(b), anger and disgust have similar emotion intensity and their description degrees are significantly higher than other emotions, which means that both anger and disgust might be the relevant emotions. In such case, single-emotion learning algorithms can no longer work. But both MLL and EDL algorithms can match such case well. In Fig 1(c), anger's description degree is the highest, while disgust's description degree is a little lower than that of anger but significantly higher than the other four emotions. In this case, both anger and disgust might be the relevant emotions, but they have different emotion intensity. For considering the relevant emotions as equally important, MLL might lose the important information about the intensity difference between the relevant emotions. Consequently, measuring emotion intensity is very important, as with emotion intensity one can know how much each emotion is and how many emotions are related to a particular expression.

2.2 Learning from Emotion Distribution

Given a training set $G = \{(x_1, E_1), (x_2, E_2), ..., (x_n, E_n)\}$, where $x_i \in \mathcal{X}$ is an expression instance and $E_i = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, ..., d_{x_i}^{y_c}\}$ is the emotion distribution associated with x_i. The goal of emotion distribution learning is to learn a conditional probability mass function $p(y|x)$ from G, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Assume that $p(y|x)$ is a parametric model $p(y|x; \theta)$, where θ is the model parameter vector.



emotion	score	Multi-label
HAP	4.19	1
SAD	1.48	-1
SUR	1.65	-1
ANG	1.29	-1
DIS	1.32	-1
FER	1.35	-1

(a)

emotion	score	Multi-label
HAP	1.35	-1
SAD	2.32	-1
SUR	1.97	-1
ANG	4.03	1
DIS	4.39	1
FER	2.35	-1

(b)

emotion	score	Multi-label
HAP	1.48	-1
SAD	2.48	-1
SUR	2.19	-1
ANG	4.48	1
DIS	3.42	1
FER	2.16	-1

(c)

Figure 1: Typical expressions from the s-JAFFE database together with their scores, multi-labels, emotion distributions.

Many different criteria can be used to measure the distance between two distributions, such as Squared $\mathcal{X}^2$, Euclidean, Jeffery's divergence, Kullback-Leibler (K-L) divergence and so on. Here we use Jeffery's divergence defined by

$$D_J(Q_a||Q_b) \;=\; \sum_j (Q_a^j - Q_b^j) log \frac{Q_a^j}{Q_b^j}, \qquad (1)$$

where Q_a^j and Q_b^j are the j-th element of the two distributions Q_a and Q_b, respectively. Jeffery's divergence is balanced, which makes $D_J(Q_a||Q_b)$ equal to $D_J(Q_b||Q_a)$.

The above formula calculates the sum of all the distances between the description degrees of the same emotion, i.e., the superscripts of Q_a and Q_b are the same (i.e., j). One possible problem of the definition in Eq. (1) is that the relationship among different emotions is not considered. In fact, some basic emotions often appear together, e.g., disgust and fear, and some often conflict to each other, e.g., happiness and sadness. Thus the weighted Jeffery's divergence is proposed here as

$$D_{wJ}(Q_a||Q_b) \;=\; \sum_{j,k} \lambda_{jk} (Q_a^j - Q_b^k) log \frac{Q_a^j}{Q_b^k}, \qquad (2)$$

where the weight λ_{jk} models the relationship between the j-th emotion and the k-th emotion in the distribution, which can be calculated by

$$\lambda_{jk} = \begin{cases} \dfrac{1}{\Lambda_j}(\rho_{jk})^{\eta} & |\rho_{jk}| \geq \varepsilon \\ 0 & otherwise \end{cases} \qquad (3)$$

where $\rho_{jk} = \frac{\sum_i (d_{x_i}^{y_j} - \overline{d_x^{y_j}})(d_{x_i}^{y_k} - \overline{d_x^{y_k}})}{\sqrt{\sum_i (d_{x_i}^{y_j} - \overline{d_x^{y_j}})^2}\sqrt{\sum_i (d_{x_i}^{y_k} - \overline{d_x^{y_k}})^2}}$ means correlation coefficient between the j-th emotion and the k-th emotion. $\Lambda_j = \sum_k (\rho_{jk})^{\eta}$ is a normalization factor that makes sure $\sum_k \lambda_{jk} = 1$. η is a positive odd number, which controls the degree the correlation coefficient works. ε is a threshold. If a couple of emotions have an absolute value of correlation coefficient smaller than ε, they are considered to have no relationship.

Then the optimal model parameter vector θ^* is determined by

$$
\begin{aligned}
\theta^* &= \arg\min_\theta \sum_i D_{wJ}(E_i||\hat{E}_i) \\
&\quad -\xi_1 \frac{1}{n} \sum_k \|\theta_k - \overline{\theta}\|_2^2 + \frac{1}{2}\xi_2 \sum_{k,r} \theta_{kr}^2 \\
&= \arg\min_\theta \sum_{i,j,k} \lambda_{ij}(d_{x_i}^{y_j} - p(y_k|x_i,\theta))(lnd_{x_i}^{y_j} - lnp(y_k|x_i,\theta)) \\
&\quad -\xi_1 \frac{1}{n} \sum_k \|\theta_k - \overline{\theta}\|_2^2 + \frac{1}{2}\xi_2 \sum_{k,r} \theta_{kr}^2,
\end{aligned}
\tag{4}
$$

where E_i is the ground truth emotion distribution of the i-th example and the $\hat{E_i}$ is the predicted one by $p(y|x_i;\theta)$. The second term is a regularizer to prevent too smooth output to emphasize the important emotions, and the third term is another regularizer to prevent unstable output. ξ_1 and ξ_2 are the balance factors.

As to the form of $p(y|x;\theta)$, similar to the work of Geng et al [5], we assume it to be a maximum entropy model [2], i.e.,

$$
p(y_k|x_i;\theta) = \frac{1}{\mathcal{Z}_i}\exp(\sum_r \theta_{kr} x_i^r), \tag{5}
$$

where $\mathcal{Z}_i = \sum_k \exp(\sum_r \theta_{kr} x_i^r)$ is the normalization factor, x_i^r is the r-th feature of x_i, and θ_{kr} is an element in θ. Substituting Eq.(5) into Eq.(3) yields the target function of θ.

$$
\begin{aligned}
T(\theta) &= \sum_i \mathcal{Z}_i + \sum_{i,j,k} \lambda_{jk}[\frac{1}{\mathcal{Z}_i} exp(\sum_r \theta_{kr} x_i^r) \\
&\quad (\sum_r \theta_{kr} x_i^r - ln\mathcal{Z}_i - lnd_{x_i}^{y_j}) - d_{x_i}^{y_j} \sum_r \theta_{kr} x_i^r] \\
&\quad -\xi_1 \frac{1}{n} \sum_k \|\theta_k - \overline{\theta}\|_2^2 + \frac{1}{2}\xi_2 \sum_{k,r} \theta_{kr}^2.
\end{aligned}
\tag{6}
$$

The minimization of the function $T(\theta)$ can be effectively solved by the limited-memory quasi-Newton method L-BFGS [7]. The computation of L-BFGS is mainly related to the first-order gradient of $T(\theta)$, which can be achieved by

$$
\begin{aligned}
\frac{\partial T(\theta)}{\partial \theta_{kr}} &= \sum_{i,j,k} \lambda_{jk}[p_{ik} x_i^k (1 - p_{ik})(\sum_r \theta_{kr} x_i^r - ln\mathcal{Z}_i \\
&\quad -lnd_{x_i}^{y_j} + 1)] - \sum_i x_i^r (1 - p_{ik}) - \xi_1 \frac{1}{n}[(\theta_{kr} - \overline{\theta}_r) \\
&\quad -\frac{1}{c}\sum_k (\theta_{kr} - \overline{\theta}_r)] + \xi_2 \sum_{k,r} \theta_{kr},
\end{aligned}
\tag{7}
$$

where $p_{ik} = \frac{1}{\mathcal{Z}_i}\exp(\sum_r \theta_{kr} x_i^r)$.

In order to compare with the multi-label learning methods, labels in the predicted distribution should be divided into two sets, i.e, the relevant and irrelevant sets. For this purpose, an extra virtual label y_0 is added into the label set, i.e., the extended label set $\mathcal{Y}' = \mathcal{Y} \cup \{y_0\}$=$\{y_0, y_1, y_2 \cdots y_c\}$. Using the new extended label set to do the training process, the optimal parameter vector θ^* is learned. As y_0 is the label that distinguishes the relevant and irrelevant emotions directly, it is initialized the same as the threshold used in MLL. Given a test image x', its emotion distribution is predicted by $p(y|x';\theta^*)$. The description degree of y_0 splits the predicted distribution into two sets. The emotions with the description degree higher than y_0's are regarded as the relevant emotions, and the rest emotions are regarded as irrelevant emotions. So that EDL can realize the function of MLL without setting the threshold manually.

3. EXPERIMENTS

To demonstrate the effectiveness of the proposed EDL algorithms, we have performed extensive experiments on two

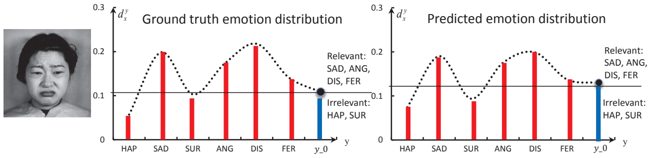


Figure 2: Typical example of the emotion distribution predicted by EDL.

widely used facial expression databases: s-JAFFE[9] and s-BU_3DFE [16], both of them are extended to the emotion distribution case proposal in this paper.

The s-JAFFE database contains 213 grayscale images posed by 10 Japanese female models. Each image is of the size 256 × 256 pixels and each model has 2-4 samples for each expression. The images are cropped manually so that the eyes are at the same positions, and then the cropped images are resized to 110 × 140 pixels. The features are extracted by the method of Local Binary Patterns (LBP) [1]. We set the diagram to 2 and the number of neighbours to 16. LBP histograms are then used as feature vectors. The dimensionality of each feature vector is eventually reduced to 243.

The s-JAFFE database is scored by 60 persons on the 6 basic emotions (i.e., happiness, sadness, surprise, fear, anger and disgust) with a 5 level scale (5 represents highest emotion intensity, while 1 represents lowest emotion intensity). The average score of each emotion is used to represent the specific emotion intensity.

The second database named s-BU_3DFE is much larger than s-JAFFE. There are 2500 examples in this database. 23 students are asked to score the s-BU_3DFE database by the same method of scoring s-JAFFE. The specific scores on each basic emotion are obtained and transferred into emotion distributions. We preprocessed the facial images and extracted the feature from s-BU_3DFE using the same method as s-JAFFE.

EDL is compared with four existing LDL methods and 7 widely used MLL methods. For each compared method, several parameter configurations are tested and the best performance is reported. The virtual label and the threshold value used in MLL are all set to 2.3. Besides, the η, ε, ξ_1 and ξ_2 are set as 5, 0.25, 0.0001, 0.001 respectively. For the LDL methods, k in AA-KNN is set to 6. Linear kernel is used in PT-SVM. The number of hidden-layer neurons for AA-BP is set to 60. For the MLL methods, the value of k is set to 6 in ML-KNN, ratio is 0.02 and μ is 2 in ML-RBF. Linear kernel is used in LIFT. Rank-SVM uses the RBF kernel with the width σ equals to 1. Ten-fold cross validation is conducted in each algorithm.

Fig. 2 gives one example from s-JAFFE database by EDL. The ground truth emotion distribution is obtained by normalizing the scores and the virtual label y_0. As can be seen, the curve of the predicted emotion distribution is very similar as the ground truth distribution, which demonstrates that EDL can learn the varying intensities of all the basic emotions well. Furthermore, the trained description degree of the virtual label y_0 can act as the threshold automatically rather than be set heuristically in MLL, which splits the predicted emotion distribution into relevant and irrelevant sets.

Table 1 reports the experimental results of EDL and several LDL algorithms. The best performance on each measure is highlighted by boldface. The two-tailed t-tests with 5% significance level are performed to see whether the dif-

Table 1: Experimental results of Label Distribution Learning Methods

database	Algorithm	Evaluation Criterion					
		Euclidean(↓)	Sϕrensen(↓)	Squared $\mathcal{X}^2$(↓)	K-L(↓)	Intersection(↑)	Fidelity(↑)
s-JAFFE	EDL	**0.0957±0.0068**	**0.1002±0.0059**	**0.0339±0.0043**	**0.0346±0.0045**	**0.8998±0.0059**	**0.9914±0.0011**
	AA-KNN [4]	0.1306±0.0117●	0.1273±0.0110●	0.0534±0.0086●	0.0556±0.0099●	0.8727±0.0110●	0.9863±0.0023●
	PT-Bayes [4]	0.1682±0.0219●	0.1644±0.0168●	0.0835±0.0215●	0.0916±0.0269●	0.8356±0.0168●	0.9784±0.0059●
	PT-SVM [4]	0.1696±0.0117●	0.1689±0.0099●	0.0812±0.0094●	0.0854±0.0110●	0.8311±0.0099●	0.9792±0.0025●
	AA-BP [4]	0.1908±0.0208●	0.1880±0.0195●	0.1139±0.0210●	0.1100±0.0273●	0.8120±0.0195●	0.9685±0.0058●
s-BU_3DFE	EDL	**0.1055±0.0023**	**0.1061±0.0025**	**0.0402±0.0017**	**0.0420±0.0020**	**0.8939±0.0043**	**0.9898±0.0046**
	AA-KNN [4]	0.1549±0.0036●	0.1464±0.0042●	0.0697±0.0036●	0.0743±0.0031●	0.8536±0.0008●	0.9821±0.0036●
	AA-BP [4]	0.1648±0.0076●	0.1595±0.0065●	0.0760±0.0061●	0.0808±0.0063●	0.8405±0.0017●	0.9804±0.0061●
	PT-Bayes [4]	0.1659±0.0044●	0.1606±0.0049●	0.0766±0.0039●	0.0830±0.0037●	0.8394±0.0010●	0.9803±0.0039●
	PT-SVM [4]	0.1701±0.0032●	0.1638±0.0047●	0.0799±0.0044●	0.0877±0.0029●	0.8362±0.0007●	0.9794±0.0044●

Table 2: Experimental results of Multi-label Learning Methods

database	Algorithm	Evaluation Criterion				
		Average Precision(↑)	Coverage(↓)	Hamming Loss(↓)	One Error(↓)	Ranking Loss(↓)
s-JAFFE	EDL	**0.9037±0.0300**	**2.6913±0.3680**	**0.2540±0.0352**	**0.1175±0.0687**	**0.1374±0.0316**
	ML-RBF [17]	0.8651±0.0738●	3.2675±0.4157●	0.2484±0.0810●	0.1810±0.1088●	0.2005±0.1085●
	ML-KNN [20]	0.8455±0.0605●	3.4310±0.3822●	0.2790±0.0617●	0.1976±0.0616●	0.2184±0.0878●
	LIFT [18]	0.8050±0.1042●	3.5397±0.3604●	0.3254±0.0777●	0.2690±0.1679●	0.2515±0.1268●
	Rank-SVM [20]	0.7516±0.0982●	3.9198±0.2844●	0.3356±0.1127●	0.3008±0.1390●	0.3276±0.1189●
	MLLOC [6]	0.7502±0.0998●	3.9127±0.2146●	0.5734±0.0639●	0.3214±0.1496●	0.3399±0.1275●
	BP-MLL [19]	0.7435±0.1009●	4.0786±0.2630●	0.3591±0.1133●	0.3444±0.1477●	0.3593±0.1210●
	ECC [12]	0.7380±0.0952●	3.9500±0.2305●	0.3591±0.1133●	0.3484±0.1130●	0.3424±0.1264●
s-BU_3DFE	EDL	**0.7861±0.0216**	**0.5236±0.0758**	**0.1167±0.0069**	**0.3667±0.0349**	**0.1761±0.0178**
	ML-RBF [17]	0.7157±0.0370●	0.6756±0.1139●	0.1271±0.0075●	0.4360±0.0559●	0.2165±0.0300●
	ML-KNN [20]	0.6224±0.0282●	0.9712±0.1103●	0.1387±0.0043●	0.5699±0.0362●	0.2947±0.0243●
	LIFT [18]	0.6062±0.0323●	1.0204±0.1225●	0.1443±0.0045●	0.5913±0.0413●	0.2515±0.0296●
	Rank-SVM [20]	0.6016±0.0321●	0.9876±0.1231●	0.2118±0.0164●	0.6101±0.0380●	0.3069±0.0338●
	MLLOC [6]	0.4970±0.0288●	1.4832±0.2538●	0.1453±0.0042●	0.7140±0.0247●	0.4352±0.0557●
	BP-MLL [19]	0.5367±0.0210●	1.1692±0.1043●	0.1817±0.0098●	0.7024±0.0280●	0.3480±0.0263●
	ECC [12]	0.4419±0.0185●	1.8396±0.0677●	0.1426±0.0046●	0.7484±0.0317●	0.5438±0.0105●

ferences between EDL and the baseline algorithms are statistically significant. The results of t-tests are given right after the performance of each method, where • indicates significance difference. As can be seen, EDL performs best on all criteria. The comparison results of EDL with seven multi-label classifiers are tabulated in Table 2. Similarly, EDL performs best on all evaluation measures. This implies that EDL can not only match more complex cases, but also perform better than MLL methods in the scenario of MLL, owing to the consideration in varying intensity of the basic emotions.

4. CONCLUSIONS

This paper presents the problem of emotion distribution recognition, and proposes to solve it by the algorithm of EDL. Different from the previous facial expression recognition methods, EDL can output the intensity of all the emotions, which well matches the reality that one facial expression is often the mixture of the basic emotions with different intensity. The experimental results show that EDL performs significantly better than some state-of-the-art multi-label algorithms and label distribution learning algorithms.

5. ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (61273300, 61232007, 61375057), Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), and Natural Science Foundation of Jiangsu Province of China (BK20131298).

6. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[3] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124, 1971.

[4] X. Geng and R. Ji. Label distribution learning. In *Proceedings of the 13th IEEE International Conference on Data Mining Workshops*, pages 377–383, 2013.

[5] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

[6] S.-J. Huang and Z.-H. Zhou. Multi-label learning by exploiting label correlations locally. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 949–955, Toronto, Canada, 2012.

[7] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989.

[8] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, Columbus, OH, 2014.

[9] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The japanese female facial expression (JAFFE) database. http://www.kasrl.org/jaffe.html, 1998.

[10] V. J. Mistry and M. M. Goyani. A literature survey on facial expression recognition using global features. *International Journal of Engineering and Advanced Technology*, 2:653–657, 2013.

[11] R. Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1, 1980.

[12] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

[13] O. Rudovic, V. Pavlovic, and M. Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641, 2012.

[14] C. Song, W. Liu, and Y. Wang. Facial expression recognition based on hessian regularized support vector machine. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 264–267. ACM, 2013.

[15] T.-H. Wang and J.-J. J. Lien. Facial expression recognition system based on rigid and non-rigid motion separation and 3d pose estimation. *Pattern Recognition*, 42(5):962–977, 2009.

[16] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *7th IEEE international conference on automatic face and gesture recognition*, pages 211–216, 2006.

[17] M.-L. Zhang. Ml-rbf: Rbf neural networks for multi-label learning. *Neural Processing Letters*, 29(2):61–74, 2009.

[18] M.-L. Zhang. Lift: Multi-label learning with label-specific features. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1609–1614, Barcelona, Spain, 2011.

[19] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[20] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.