RESEARCH ARTICLE OPEN ACCESS

Toward Instantaneous Facial Expression Recognition Using Privileged Information.

Michele Mukeshimana*, Xiaojuan Ban*, Yitong Li*
* School of Computer and Communication Engineering, University of Science and Technology Beijing, 100083 Beijing, China.

----------**********----------

Abstract:

The Facial Expression Recognition is one of determining factors in automatic recognition of humans' emotion. There are many works done in Facial Expression Recognition research, yet it has been difficult to build real-time and robust systems. This paper proposes the use of a classifier trained by Extreme Learning Machine to solve the speed issue, combined with the use of privileged information to improve the testing time and the reduction of the testing error. The study is done on features differently extracted from still facial datasets. The experimental results show that the proposed method improves the testing time making it feasible in real-time applications and more stable than the classical method, with time improvement reaching 25%.

Keywords — **Emotion recognition, Extreme Learning Machine, Learning Using Privileged Information, Facial Expression Recognition, Machine Learning, Real-Time Execution time.**

----------**********----------

I. INTRODUCTION

Facial Expression Recognition (FER) research has attracted many researchers for a long time because of its application in many domains. Human being are able to express emotion through many ways and among them, facial expression being most the most studied. The facial expression recognition starts by the face detection, feature extraction and classification [42]. Face detection is similar for all application the aim is to detect and localize the face on an image or video. The feature extraction depends on the application (expression recognition, identity recognition, motion recognition…) and the method used. The work in this paper focuses on Facial Expression Recognition classification.

The FER's main purpose is to detect and to identify the human emotion expressed by face. It is mainly applied in Human Computer Interaction (HCI) to introduce a natural way of communication in the interaction between man and machine. After, the work of Ekman and Friesen [1] proposing the Facial Action Coding System (FACS), many FER techniques have been introduced [2-8]. The main challenge is to obtain accurate and a real time technique to apply in a natural approach. This paper proposes a method that simultaneously enhances the recognition rate and reduces the execution time. The rest of the article is arranged as follows first, brief review of related work, second a description of the used method, and third, experiments and results as well as the discussion. At the end, conclusions and acknowledgements are expressed.

II. RELATED WORK

FER is one of the many of research domains in Affective Computing [9, 10, 11]. It mainly consists in three steps: face detection, feature extraction and face classification [43]. Concerning face detection, there are many proposed algorithms to locate a human face in a scene. In their works, Sing et al. [37] and Bouzerdoum and Chai [38] proposed a method for finding faces in images with controlled background such as based on skin colour, motion or mixture of both. More details can be found in surveys [12-15], [36], [39, 40]. The feature extraction consists of gathering resources that can help in the face description. The state-of art methods can be found in the work done by Al-Allaf [39]. The recognition (classification) is done by classifier based on support vector machine and neural networks [30, 41].

This paper focuses on Facial Expression Classification. Even though, many works are done for the previous steps (face detection and feature extraction) there is a big challenge in classification. Classification consists in identification and attribution of the expression of the observed face.

The applied method reduces the execution time and increases the recognition accuracy.

III. METHODS APPLIED

A. Learning Using Privileged Information

Learning Using Privileged Information (LUPI) is a new learning paradigm introduced by Vapnik et al. [16] as analogy to the human being learning system. LUPI paradigm introduces the use of the additional information in the training step of the classical learning method [17]. It is called Privileged Information because it is only available during the training stage. During the natural learning process, the teacher provides the student with elements for a deeper understanding and memorization of the experience. The same way, this additional information transfers information to the original data modality to enhance the training accuracy and reduce the testing error [18].

The classical learning paradigm of a supervised machine learning uses a set of N pairs (x_i, y_i) as training data with $x_i \in X$ and $y_i \in Y$ aiming to find the function that minimizes the probability of incorrect classifications [19]. The LUPI paradigm introduces another learning model; in the pairs of training set it adds a third component. The privileged information space has the same number of examples as the standard space but differ in attributes. Thus, in the

LUPI paradigm, a set of triplets (x_i, x_i^*, y_i) is given as training set, with $x_i \in X$, $x_i^* \in X^*$ and $y_i \in Y$. The objective is still the same, which is to find the function that guarantees the smallest probability of incorrect classification. The difference is that in the LUPI, we are given an additional information $x_i^* \in X^*$ which is different from $x_i \in X$. However, this information will only be available during the training stage that is why it is called privileged information. It sums up an idea of Vapnik [28] leading to integration in learning techniques, of element of exact science and art of data interpretation, exact science and humanities and exact science and emotions.

In many learning scenarios, there are examples of privileged information like the poetic description of a digit recognition problem, rate of exchange information after moment t for a prediction of a currency exchange rate at moment t problem, etc. In the experiments, we use differently extracted and features, one set as privileged information to another. More details in the section 4.

B. *Extreme Learning Machine(ELM) using privileged Information*

The used classifier is trained on ELM algorithm. ELM is an emerging learning machine algorithm with three main characteristics, namely, extreme fast training speed, good generalization and a universal approximation capability.

1) Brief review of ELM: Extreme Learning Machine (ELM) is a learning algorithm firstly proposed by Huang et al. [20] for a single-hidden layer feedforward neural networks (SLFN) and was extended to generalized SLFN where the hidden layer doesn't need to be neuron alike [21,22]. It is described as follows:

Given a set of N training samples $(x_i, t_i), i = 1, .., N$ with $x_i \in X^d$ as the input vector for the ith sample and $t \in T$ its target value. The output of ELM with *L* hidden nodes is:

$$f(x) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x) = \beta.h(x) \quad (1)$$

With a_i and b_i are input weight and bias respectively between the input nodes and the hidden layer, β_i is the output weight between the ith hidden node and the output node and $G(a_i, b_i, x)$ is the output of the ith hidden node with respect to the input x, also called activation function in literature. $h(x)$ is the output vector of the hidden node with respect to the input x, which maps the data from d-dimension of the input space to the L-dimension of the hidden layer feature space H . For the binary classification applications, the decision function of ELM is:

$$f(x) = sign(\sum_{i=1}^{L} \beta_i G(a_i, b_i, x)) \quad (2)$$

Equation (1) can also be written in the matrix form as [20], [23]:

$$H\beta = T \quad (3)$$

Given N , L and C , number of the samples, hidden nodes and class respectively, H_{N*L} is the hidden-layer output matrix, β_{L*C} is the output weight and T_{N*C} is the target matrix. β can be directly calculated as:

$$\hat{\beta} = H^{\dagger} T \quad (4)$$

Where $H^{\dagger}$ is the Moore-Penrose generalized inverse of matrix H . There are several methods to compute the $H^{\dagger}$ like the orthogonal projection method, iterative method and singular value decomposition [24, 25]. In summary, the basic ELM algorithm consists mainly in 3 steps learning model:

The result decision function is: $f(x) = h(x)\hat{\beta} = h(x)H^{\dagger}T$.

ELM simultaneously tends to reach the smallest training error and the minimum norm of the output weight [25, 26]:

$$\min \sum_{i=1}^{N} \left\| \beta.h(x_i) - t_i \right\| \text{ and } \min \left\| \beta \right\| \quad (5)$$

The classification problem turns into to simultaneously solving the optimization problem in equation (7). Thus, the classification problem is as follows:

$$\min_{\beta,\xi} L_{PELM} = \frac{1}{2}\left\|\beta\right\|^2 + C\frac{1}{2}\sum_{i=1}^{N}\left\|\xi_i\right\|^2$$
$$s.t: h(x_i)\beta = t_i^T - \xi_i^T, i = 1, \cdots, N \quad (6)$$

Where $\xi_i = \left[\xi_{i,1}, \xi_{i,2}, \cdots, \xi_{i,c}\right]^T$ is the estimation of the training error vector of c output nodes with respect to the input sample x_i , and C is the regularization parameter, which represents the trade-off between the minimization of training errors and the maximization of the marginal distance.

From Karush-Kuhn-Tucker (KKT) theorem [27], to train ELM is equivalent to solving the following dual optimization problem:

$$L_{DELM} = \frac{1}{2}\left\|\beta\right\|^2 + C\frac{1}{2}\sum_{i=1}^{N}\left\|\xi_i\right\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{i,j}(h(x_i)\beta_j - t_{i,j} + \xi_{i,j}) \quad (7)$$

Where β_j is the vector of the weights linking hidden layer to the jth output node and $\beta = \left[\beta_1, \cdots, \beta_c\right]$. From corresponding KKT optimality conditions, we compute and obtain:

$$\beta = H^T\left(\frac{1}{C} + HH^T\right)^{-1}T \quad (8)$$

Now, the output function of ELM classifier is:

$$f(x) = h(x)\beta = h(x)H^T\left(\frac{1}{C} + HH^T\right)^{-1}T \quad (9)$$

The generalization of ELM has been found insensitive to the dimension of the feature space and it can reach good performance as long as L is large enough [25].

2) ELM Using Privileged Information: As aforementioned, the privileged information is the additional information avail-able only during the training stage. Learning Using Privileged Information aims to increase the training accuracy and reducing the probability of the error of classification. It is similar to the case when

the teacher gives examples and interacts with the students during training time, but will not be there in the testing time where the student will be out of the teacher's supervision [19]. In using privileged information, it is possible to try transfer a set of useful feature for parameter in the privileged information space X^* into their image in non-privileged information space X

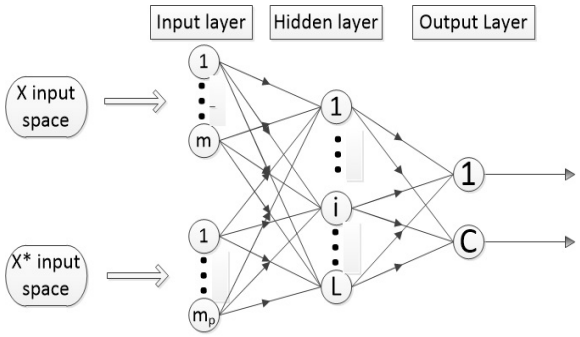

Fig. 1 ELM Using Privileged Information flowchart

Fig.1 represents the flow of the processes in the ELM using privileged information with a standard space of m-dimension, and the privileged information space of m_p-dimension. There are L hidden nodes in the hidden layer and the C-classes.

In ELM, as well as in other classical learning machine algorithm, the privileged information can be used as the correction function as proposed by Vapnik [28]. In this case, we are given N triplets (x_i, x_i^*, t_i) with $x_i \in X$, the space of non-privileged (or standard features), $x_i^* \in X^*$ the space of the privileged information, and $t_i \in T$ the target set.

The vector X is mapped into the feature space by $h(x) = [h_1(x),...,h_L(x)]$, the vector X^* into correcting space by $h^*(x^*) = [h_1^*(x^*),...,h_L^*(x^*)]$. The two functions $h(x)$ and $h^*(x^*)$ can be the same or different. They are all mapped into the same decision space, are both used to build the decision function. The correcting function, which estimates the slack value, is given by:

$$\xi(x) = \phi(x^*) = h^*(x^*)\beta^* \tag{10}$$

Then the classification problem thus becomes:

$$\min_{\beta,\beta^*} L_{PELM} = \frac{1}{2}\|\beta\|^2 + \frac{\lambda}{2}\|\beta^*\|^2 + C\frac{1}{2}\sum_{i=1}^{N}(h^*(x^*)\beta^*)^2 \tag{11}$$
$$s.t: h(x_i)\beta = t_i - h^*(x_i^*)\beta^*, i = 1,\cdots,N$$

With β_i^* the correcting weight connecting the ith hidden node to the output node in the correcting space, and λ is a parameter to adjust the relative weights of the two spaces. To minimize the equation (11) we construct the Lagrangian function as:

$$L(\beta,\beta^*,\alpha) = \frac{1}{2}\|\beta\|^2 + \frac{\lambda}{2}\|\beta^*\|^2 + C\frac{1}{2}\sum_{i=1}^{N}(h^*(x_i^*)\beta^*)^2 - \sum_{i=1}^{N}\alpha_i(h(x_i)\beta - t_i + h^*(x_i^*)\beta^*) \tag{12}$$

The KKT conditions of optimality are:

$$\frac{\partial L}{\partial \beta} = 0 \rightarrow \beta = \sum_{i=1}^{N}\alpha_i(h(x_i))^T = H^T\alpha \tag{13}$$

$$\frac{\partial L}{\partial \beta^*} = 0 \rightarrow \lambda\beta^* + C(H^*)^T H^*\beta^* - (H^*)^T\alpha = 0 \tag{14}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow h(x_i)\beta - t_i + h^*(x_i^*)\beta^* = 0 \rightarrow H\beta + H^*\beta^* = T \tag{15}$$

From equation (13), (14) and (15), the output weight β is:

$$\beta = \left(H + H^*\left(\lambda I + C\left(H^*\right)^T H^*\right)^{-1}\left(H^*\right)^T\left(H^T\right)^{\dagger}\right)^{\dagger} T \tag{16}$$

Then the output function of the ELM classifier becomes:

$$f(x) = h(x)\beta = h(x)\left(H + H^*\left(\lambda I + C\left(H^*\right)^T H^*\right)^{-1}\left(H^*\right)^T\left(H^T\right)^{\dagger}\right) \tag{17}$$

This solution is the same as the one proposed by Zhang et al. [29] and the corresponding algorithm was called ELM+. The work is to apply it in Face Expression Recognition having features differently extracted. Its application has the effect of reducing testing time and improving the classification performance.

IV. Experiments

A. *Experiments description*

In experiments, features used are extracted from the Japanese Female Face Expression dataset (JAFFE) [45] and the MMI datasets [46]. The feature extraction was done using three methods. The first method is the Local Binary Patterns (LBP) [31, 32], the second is Local Direction Pattern (LDN) [33, 34] and the third is the Edge Orientation Histograms (EOH) [35]. The three methods have different outlooks and more details can be found in the work in [30] and references therein.

From the three methods, the features extracted are distinct but complementary because they describe the same image from different point of view and they are all local-based descriptor. They are well eligible for LUPI paradigm [28]. In the work, three types of features are successively grouped each as standard feature and another as privileged information instead of concatenating them in single feature vector.

TABLE I
THE DESCRIPTIONS OF THE USED DATASETS

Dataset	Feature number	Samples number
JAFFE LBP	13276	213
JAFFE LDN	14401	213
JAFFE EOH	8101	213

MMI LBP	13276	504
MMI LDN	14401	504
MMI EOH	8101	504

The JAFFE LBP and MMI LBP are dataset of the features extracted by LBP method. JAFFE LDN and MMI LDN are dataset of the features extracted by LDN method. JAFFE EOH and MMI EOH are dataset of the features extracted by EOH method.

We have utilized a five-fold cross validation to evaluate the algorithms. All the simulations and evaluations are done with MATLAB R2014b running on an Intel® Core™ i5-4590 CPU @ 3.30GHz with 4.00GB RAM. Some parts of the codes are from the codes available on the Extreme Learning Machine (ELM) website[44] Graphs and tables are done with Microsoft Office Excel 2016.

B. Results and Discussions

1) Performance evaluation: In first experiment, different possible combinations of three groups of features were made and the results are in the Table II.

TABLE II

THE PERFORMANCE IN TRAINING ACCURACY AND TESTING ACCURACY

Standard features→ Privileged Information ↓	JAFFE/MMI LBP		JAFFE/MMI LDN		JAFFE/MMI EOH	
	Training Accuracy	*Tesing Accuracy*	*Training Acuracy*	*Tesing Accuracy*	*Training Acuracy*	*Tesing Accuracy*
JAFFE LBP			99.7	97.78	100	81.67
JAFFE LDN	99.85	81.67			100	86.11
JAFFE EOH	97.02	72.78	95.68	94.4		
MMI LBP			99.88	99.04	99.94	82.93
MMI LDN	100	99.04			100	97.6
MMI EOH	97.75	73.56	98	79.56		

From the above results, the contribution of all feature sets is distinct because they are different one from the other in the description they aim. The accuracies are expressed in percentages. The best recognition in training stage is observed when the EOH features is used as standard information and other features as privileged information. Thus, the training accuracy is more improved when the more accurate (more informative) set is used (Table III) in training stage as standard features and others as privileged information.

The best improvement in testing accuracy is more observed when the LDN is used as privileged information. The LDN features have the highest number of features among others (Table I) and their contribution is more important than others. From the results obtained when using single set in Table III. This satisfies the condition for the theory of knowledge transfer[19]. The knowledge transfer is better applied when the training size is the largest.

TABLE III

SINGLE SET CLASSIFICATION ACCURACY

Dataset	Classification accuracy	Dataset	Classification accuracy
JAFFE LDN	86.11	MMI LDN	85.34
JAFFE LBP	86.67	MMI LBP	86.54
JAFFE EOH	90.00	MMI EOH	91.11

The advantages of the LUPI over the classical paradigm in feature fusion was evaluated. The results are represented in Fig.2.

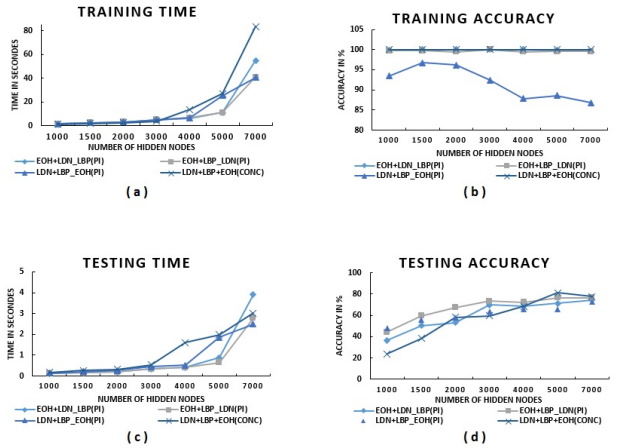

Fig. 2 Comparison of LUPI paradigm and classical paradigm on JAFFE dataset: (a) training time, (b) testing time, (c) training accuracy, (d) testing accuracy

The graphs represent the time of training and testing, and their respective accuracies according to the number of the hidden nodes. The EOH+LDN_LBP (PI) (EOH+LBP_LDN (PI), LDN+LBP_EOH (PI)) means that the EOH and LDN (EOH and LBP, LDN and LBP) features are taken as standard attributes (non-privileged) and the LBP (LDN, EOH) features operate as Privileged Information. The classical paradigm uses the concatenated vector. The results prove that the LUPI is better than the classical method in information fusion. The execution time for the classical information fusion is higher than in LUPI, Fig. 2 (a) and (c). The recognition rate is also improved in the LUPI paradigm, Fig.2 (b) and (d). Therefore, the learning using privileged information in features fusion improves both execution time and recognition rates.

The comparison to other methods is done by comparing the proposed method to basic ELM. The results are represented in the following table:

RTABLE IV

COMPARISON TO CLASSICAL METHODS

Datasets	Basic ELM	LUPI method
LBP-EOH	83.89	72.77
LDN-EOH	91.34	94.44
EOH-LBP	83.89	81.67
EOH-LDN	75.56	86.11
LDN-LBP	93.25	97.78
LBP-LDN	78.89	81.67

In Table IV, there are results of comparing the proposed technique to basic ELM method. The use of proposed method mostly outperforms basic ELM.

In addition to the performance improvement, it tends to generalize earlier than the Basic ELM as shown on the Fig. 3.

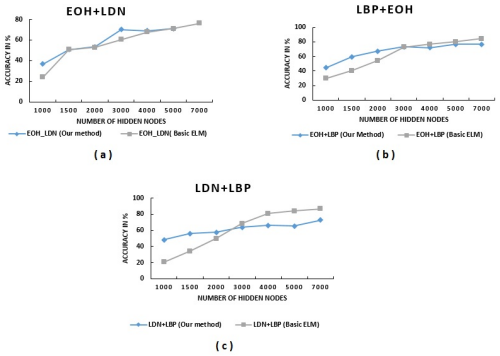

Fig. 3 Accuracy evaluation on JAFFE dataset: (a) EOH feature and LDN features (b) LBP features and EOH (c) LDN features and LBP features

Fig. 3 illustrates the comparison in information fusion. The proposed method converges earlier than the Basic ELM and tends to stabilize, while the Basic ELM remains sensitive to the increase of the number of hidden nodes.

Another advantage is the reduction of the testing time (Fig.4), making the technique candidate for the real-time applications.

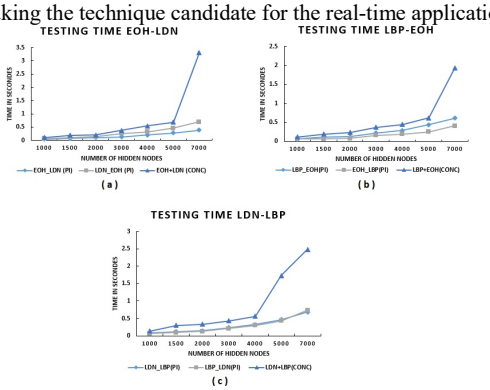

Fig. 4 Testing time improvement on JAFFE dataset: (a) EOH feature and LDN features (b) LBP features and EOH (c) LDN features and LBP features

Fig.4 represents the comparison with the use of the concatenation for the testing time. LDN_EOH(PI) (EOH_LDN(PI)) signifies that the LDN (EOH) attributes are taken as standard attributes and the EOH (LDN) attributes operate as the Privileged Information, EOH_LDN(CONC) signify that the EOH and LDN features are concatenated. In case of concatenation, the classification is based on classical learning. The proposed method requires less time than classical one. Moreover, it is better applicable for large-scale problem than the classical method because the testing time increases dramatically in case of the concatenation when the number of the hidden nodes increases.

2) *Evaluation of the user-defined parameters*: The user defines three main parameters, namely, the number of the hidden nodes L, factors λ and C (equation (16)). The relating experiments aim to determine their impact on time of execution and accuracy.

The number of hidden nodes is arbitrarily chosen in the following sequence: 1000, 1500, 2000, 3000, 4000, 5000 and 7000. The system attained the best performance in training and test when hidden number is around 2^{12}. The evaluation results are represented in the Fig.5

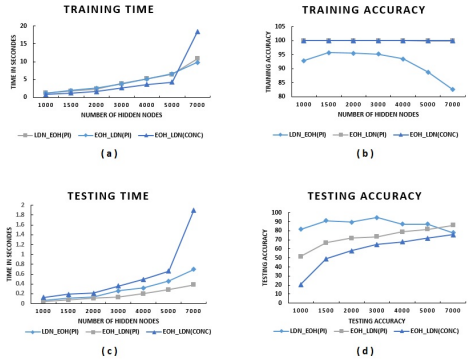

Fig. 5 Effect of the hidden nodes on JAFFE dataset (EOH_LDN features) (a) Training time (b) Training Accuracy (c)Testing Time (d) Testing Accuracy

In Fig. 5 the number of hidden nodes influences the execution time and accuracy. The training time is higher in LUPI paradigm [17], especially when we use the LDN features as standard features because its features' number is the highest. The proposed method achieves desirable classification precision quicker than the classic learning paradigm. So, it is better to use the LUPI paradigm than simple concatenation of multiple features.

To evaluate the impact of λ and C on the classification precision, their values was arbitrarily chosen respectively from 10^{-7} to 10^{+6}. The results are represented in Fig. 6.

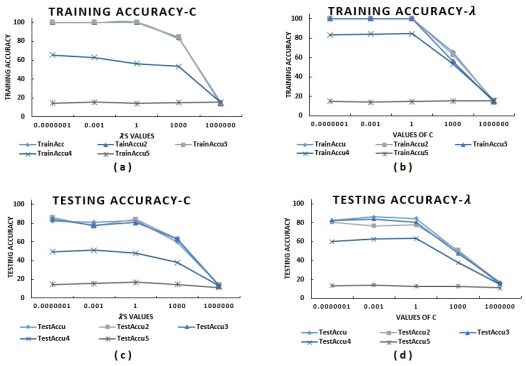

Fig. 6 The Impact of the λ and C on the classification precision on JAFFE EOH_LDN features. (a) Training space when C is fixed and λ varies (b) Training space when λ and C varies (c) Testing space when C is fixed and λ varies (d) Testing space when λ and C varies

Fig.6 represent the effect of the C(λ) value on the accuracies in the training and testing spaces respectively, according to the variation of the λ (C) value. TrainAccu (TestAccu) is the training accuracy and the testing accuracy when the stable value C (λ) is 10^{-7}; and the following value is 10^{-3}, 10^{0}, 10^{+3}, 10^{+6}. The best results are obtained when the value of C and λ are smaller or equal to one. It is better to choose C and λ belonging to] 0, 1].

V. CONCLUSION

The work aimed to improve the Facial Expression Recognition classification accuracy, using the new learning paradigm of Learning Using Privileged Information (LUPI). The classifier is trained using the Extreme Learning Machine algorithm using Privileged Information (ELM+).

The outcomes prove that the proposed technique is feasible and applicable in real life situation. The use of attributes from different perspectives and the improvements obtained make the proposed method extensible to more complex problems integrating real-time Face Expression Recognition.

Considering the testing time improvement, the method is applicable to the large-scale problems. Future works can extend the study on more datasets with multimodal features and even video recordings

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under Grant No.61272357, No.61300074, 61572075.

REFERENCES

[1] P. Ekman and W. Friesen, *Facial Action Coding System*. Consulting Psychologists Press, 1978

[2] F. Kawakami, H. Yamada, S. Morishima and H. Harashima, "Construction and Psychological Evaluation of 3-D Emotion Space," *Biomedical Fuzzy and Human. Sciences*, vol.1 (1), pp.33-42, 1995

[3] M. Rosenblum, Y. Yaccob and L. S. Davis, "Human Expression recognition from Motion using a radial basis function network architecture," *IEEE Trans. On Neural Networks*, vol.7 (5), pp. 1121-1138, 1996

[4] G. J. Edwards, C. J. Taylor and T. F. Cootes, "Interpreting face images using active appearance models," *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. p. 300, 1998*

[5] D. Keltner, P. Ekman, "Facial expression of emotion." In: *Lewis M, Haviland-Jones J M (eds), Handbook of emotions*. Guilford Press, New York, NY, pp 236–249 ,2000

[6] L. Ma, K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Trans. System, Man, and Cybernetics (Part B)*, vol.34 (4), pp.1588-1595, 2003.

[7] R. Goecke, A. Asthana and A. Dhall, "A SSIM-based approach for finding similar facial expressions. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pp.815-820 , March 2011

[8] L. Chen, H. Boughrara, M. Chtourou, and C. Ben Amar, "Face recognition under varying facial expression based on Perceived Facial Images and local feature matching," *2012 International Conference on Information Technology and e-Service (ICITeS)*, pp.1-6, March 2012.

[9] R.W. Picard, "Affective Computing," MIT Media Laboratory Perceptual Computing Section Technical Report No 321Press MIT Press1997.

[10] P. Ekman, "*Facial expressions of emotion: New findings, new questions*," Psychological Science, 3, pp.34–38, 1997.

[11] P. Ekman, *Basic emotions*, In T. Dalgleish & M. Power (Eds.), Handbook of cognition and emotion, pp. 45–60, Sussex, UK.

[12] E. Murphy-Chutorian, and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE TPAMI*, 2009.

[13] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE TPAMI*, 2002.

[14] W. Zhao, R. Chellappa, P. Phillips and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, 2003.

[15] R. Xiao, Q. Zhao, D. Zhang and P. Shi, "Facial expression recognition on multiple manifolds," *Pattern Recognition*, vol.44, pp.107–116, 2011.

[16] V. Vapnik, A. Vashist and N. Pavlovich, "Learning using hidden information: Master class learning," *Proceedings of NATO Workshop on Mining Massive Data Sets for Security*, pp. 3-14, 2008.

[17] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, 22(5-6), pp.544-557, 2009.

[18] V. Sharmanska, N. Quadrianto, C. H. Lampert, "Learning to Rank Using Privileged Information".

[19] V. Vapnik and R. Izmailov, "Learning Using Privileged Information: Similarity Control and knowledge Transfer", *Journal of Machine Learning Research* 16, pp.2023-2049, 2015.

[20] Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: "Extreme learning machine: A new learning scheme of feedforward neural networks," *Proc. of International Joint Conference on Neural Networks IJCNN*, vol. 2, pp. 985–990, 2004.

[21] G.-B. Huang, L. Chen and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," IEEE *Trans. Neural Netw.* Volume 17 (4), pp.879-892, 2006.

[22] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, volume 70, pp. 489–501, 2006.

[23] J-T. Xu, H-M. Zhou and G-M. Huang, "Extreme Learning Machine Based Fast Object Recognition," *Proc. 15th Int. Conf. Inform. Fusion (FUSION)*, pp. 1490-1496, 2012.

[24] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*. New York: Wiley, 1971.

[25] G-B. Huang, H-M.Zhou, X-J. Ding and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," I*EEE Trans. Syst. Man. Cybern.* Part B42 (2), pp.513–29 (2012)

[26] Z, Bai, G-B Huang, D-W. Wang, H.Wang and M.B. Westover "Sparse extreme learning machine for classification," *IEEE Trans. Cybern.* Volume 44 (10). pp.1858-1870, 2015.

[27] R. Fletcher, *Practical Methods of Optimization*, Volume 2 Constrained optimization. New York: Wiley ,1981.

[28] V. Vapnik, Empirical Inference Science. Afterword of 2006, Springer, 2006.

[29] W-B. Zhang, H-B. Ji, G-S. Liao, Y-Q. Zhang, "A Novel Extreme Learning Machine using privileged information" *Neurocomputing*, Volume 168, pp. 823-828, 2015.

[30] X.-J. Ban, Y.-T. Li, G. Yang and Y. Wang, "Multiple Features Fusion for Facial Expression Recognition Based on ELM," *Proceedings of CCIS 2016,* 2016.

[31] M.B. López, A.Nieto, J. Boutellier, J. Hannuksela and O.Silvén, "Evaluation of Real-time LBP computing in multiple architectures," *Journal of Real-Time Image Processing*, 2014.

[32] T. Ojala, M. Pietikäinen and D. Harwood, "A comparative study of tecture measures with classification based on featured distributions," *Pattern Recognition*, 29 (1), pp.51-59 ,1996.

[33] A. R. Rivera, J. R. Castillo, O. Chae, "Local Directional Number Pattern for Face Analysis: Face and Expression Recognition," *IEEE Transactions on Image Processing*, vol.22 (5), pp.1740-1752, 2013.

[34] T. Jabid, M.H. Kabir and O. Chae, "Local directional pattern (LDP) for face recognition," *Proc. of IEEE Int. Conf. Consum. Electron.*, pp.329-330, 2010.

[35] S. N. Kausar and S. S. Gawande, "Local Directional Number Pattern for Face Analysis: Face and Expression," *International Journal of Research in Science & Engineering*, volume 1 (2), 2012.

[36] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Transactions on Pattern analysis and Machine Intelligence* Volume 27(10), pp.1615-1630, 2005.

[37] S. K. Sing, D.S. Chauhan, M. Vatsa, and R. Singh, "A Robust Skin Colour Based Face Detection Algorithm," *Tamkang Journal of Science and Engineering* Vol.6 (4), pp. 227-234 ,2003.

[38] S. L. P. Bouzerdoum, A. D. Chai, "A Novel Skin Colour Model in YCbCr Colour Space and Its Application to Human Face Detection," *Proceedings of 2002 International Conference on Image Processing* Vol.1, pp.289-292, 2002.

[39] N. Ismail, M. I. MD. Sabri, "Review of Existing Algorithms for Face Detection and Recognition," *Recent Advances in Computational Intelligence, Man-Machine Systems and Cybernetics*, pp. 30-39, 2009.

[40] O. N. A. Al-Allaf, "Review of Face Detection Systems Based Artificial Neural Networks Algorithms," *The International Journal of Multimedia & its Applications (IJMA)* Vol.6 (1), pp.1-16, 2014.

[41] H. Kabir, T. Jabid, O. Chae, "Local Directional Pattern Variance (LDPv): A Robust Feature Descriptor for Facial Expression Recognition," *The International Arab Journal of Information Technology,* Vol.9 (4), pp. 382-339, 2012.

[42] P. Viola, M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision* 57 (2), pp.137-154 ,2004.

[43] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12): 1424-1445, 2000.

[44] Extreme Learning Machine, www.ntu.edu.sg/home/egbhuang/elm_codes.html
[45] Japanese Female Facial Expression(JAFFE) Database, www.kasrl.org/jaffe_download.html
[46] MMI Facial Expression Database, www.mmifacedb.eu