Histogram-Based Image Pre-processing for Machine Learning

Ayumi Sada
Tokyo Metropolitan University
Tokyo, Japan
sada-ayumi@ed.tmu.ac.jp

Yuma Kinoshita
Tokyo Metropolitan University
Tokyo, Japan
kinoshita-yuma@ed.tmu.ac.jp

Sayaka Shiota
Tokyo Metropolitan University
Tokyo, Japan
sayaka@tmu.ac.jp

Hitoshi Kiya
Tokyo Metropolitan University
Tokyo, Japan
kiya@tmu.ac.jp

***Abstract*—This paper proposes to use some image processing methods as a data normalization method for machine learning. Conventionally, z-score normalization is widely used for pre-processing of data. In the proposed approach, in addition to z-score normalization, a number of histogram-based image processing methods such as histogram equalization are applied to training data and test data as a pre-processing method for machine learning. We evaluate the effectiveness of the proposed approach by using a support vector machine algorithm and a random forest one. In experiments, the proposed scheme is applied to a face-based authentication algorithm with SVM/random forest classifiers to confirm the effectiveness. For SVM classifiers, both z-score normalization and image enhancement work well as a pre-processing method for improving the accuracy. In contrast, for random forest classifiers, a number of image enhancement methods work well, although z-score normalization is unuseful for improving the accuracy.**

***Index Terms*—Pre-processing, Contrast Enhancement, Support Vector Machines, Machine Learning, Random Forest**

I. Introduction

Machine learning and deep learning have been spreading in many fields such as face recognition, medical diagnosis, character recognition, and machine translation [1]–[5]. A lot of researchers have been seeking for efficient algorithms to obtain a high performance. In this paper, we focus on two machine learning algorithms: SVM [6], [7] and random forest [8], [9]. Support Vector Machine (SVM) is well known as a supervised machine learning method for binary classification. It has high capability by using margin maximization and nonlinear classification based on a kernel function [10]. Random forest is an ensemble learning method using decision trees. It is carried out by combining a plurality of decision trees and majority decision of prediction results of each decision tree. However, the performance of learning algorithms is known to be influenced by pre-processing of data such as data normalization [11], [12].

Data normalization which transforms data according to a certain rule, is widely used as pre-processing [13]. Data normalization enables us not only to improve the accuracy but also to speed up the learning. Conventionally, z-score normalization has been widely used as one of data normalization methods for most of machine learning applications. Nevertheless, studies on pre-processing for machine learning are very few, compared to studies on learning algorithms.

As for image processing, numerous image enhancement methods have been studied to improve the quality of images [14]–[19]. Histogram Equalization (HE) [20] is one of the most famous contrast enhancement method and various extended versions of HE have been proposed such as Adaptive Gamma Correction with Weighting Distribution (AGCWD) [21] and Contrast-ACcumulated Histogram Equalization (CACHE) [22]. AGCWD aims to prevent over-enhancement and under-enhancement caused by HE, by using an adaptive gamma correction and a modified probability distribution. CACHE adaptively controls the contrast gain according to the potential visual importance of intensities and pixels. Histogram Matching (HM) [23], [24] transforms an image histogram into any reference histogram. Because these image enhancement methods have been studied to improve the visuality of image, they have not been considered as pre-processing methods for machine learning.

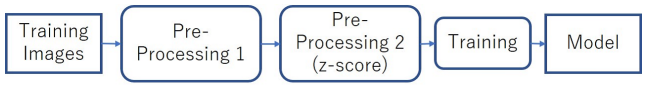
Because of such a situation, this paper proposes to use some image processing methods as a data normalization method for machine learning. In this paper, in addition to z-score normalization, histogram matching and various image enhancement algorithms including histogram equalization are applied to learning data and test one as one of pre-processing methods.

In this paper, we focus on two algorithms: SVM and random forest. The proposed framework, which consists of z-score normalization and contrast enhancement methods, was evaluated by using a face-based authentication algorithm with SVM/random forest, in terms of True Positive Rate and True Negative Rate. When we used SVM classifiers, both TPR and TNR were improved in all cases of using z-score normalization and contrast enhancement than using only z-score normalization. In contrast, for random forest classifiers, z-score normalization was not useful for improving the accuracy, and a number of image enhancement methods were useful even in such a case.

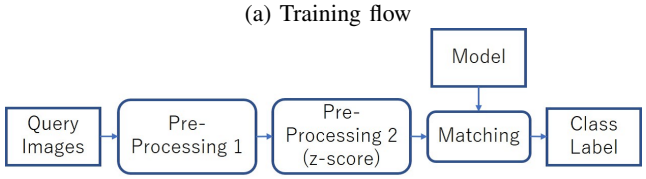
II. Scenario

Figure1 shows the outline of the proposed approach. Machine learning consists of two stages: training stage and test (query) stage, where pre-processing is performed in both stages. Conventionally, z-score normalization has been widely used as a pre-processing method, and its effectiveness has

978-1-5386-6309-7/18/$31.00 ©2018 IEEE



(a) Training flow



(b) Test flow

Fig. 1: Machine learning example with pre-processing

been evaluated in most of leaning algorithms[1]. In z-score normalization, input data $\mathbf{X}$ is transformed into output data that have average $\mu_z = 0$ and standard deviation $\sigma_z = 1$, computed by

$$z_i = \frac{x_i - \mu}{\sigma}, \tag{1}$$

where x_i and z_i are an element of input data $\mathbf{X} = \{x_i\}$ and an element of output data $\mathbf{Z} = \{z_i\}$, and μ and σ are the average vale and the standard deviation of input data respectively.

In this paper, we propose to add a new pre-processing method as pre-processing 1 in Fig.1. The aim of z-score normalization is to normalize data, while some image processing methods such as histogram equalization allow us not only to normalize data, but also to produce higher quality data than the original one as shown in Fig.2. Compared to the use of low quality data, it is expected that the use of high quality data improves the performance of machine learning algorithms.

III. Proposed Method

As described in Scenario, in this paper, some image processing methods are applied to machine learning algorithms as a pre-processing method. Here, image processing methods used as a pre-processing method are summarized.

A. Histogram Equalization

Histogram equalization (HE) is a method in image processing of contrast adjustment using the image histogram. Through this adjustment, the intensity can be better distributed on the histogram. For a grayscale image, the transformation T of the luminance value using HE is given by

$$T(k) = K \cdot cdf(k), \tag{2}$$

where k is a luminance intensity, K is the maximum luminance intensity of the input image (typically 255), and $cdf(\cdot)$ is the cumulative distribution function (CDF) of luminance. The method can lead to better views of structure in input images, and moreover, can work to justify the variation among input images, even when each input image has a different distribution on the histograms. In this paper, HE is carried out by using a function in MATLAB.

B. AGCWD

HE often produces unrealistic effects in photographs, due to too strong emphasis. In order to solve this issue, AGCWD (Adaptive Gamma Correction with Weighting Distribution), which is one of contrast enhancement methods, has been proposed. In AGCWD, the transformation T is given by

$$T(k) = K \cdot (k/K)^{\mu}, \tag{3}$$

where μ is obtained by

$$\mu = 1 - cdf_{\omega}(k). \tag{4}$$

Here, a weighted CDF $cdf_{\omega}(k)$ is obtained by the following formula

$$cdf_{\omega}(k) = \sum_{k=0}^{K} \frac{pdf(k)}{\sum_{k'=0}^{K} pdf_{\omega}(k')}, \tag{5}$$

where, pdf is the probability distribution function (PDF) of luminance and pdf_{ω} is calculated as

$$pdf_{\omega}(k) = pdf_{max} \cdot \left(\frac{pdf(k) - pdf_{min}}{pdf_{max} - pdf_{min}} \right)^{\eta}. \tag{6}$$

In this case, pdf_{max}, pdf_{min} indicate the maximum value and the minimum value of pdf respectively. According to the literature [21], the parameter $\eta = 0.8$ was set. We try to use this algorithm for machine leaning.

C. CACHE

CACHE (Contrast-ACcumulated Histogram Equalization) method, which is an extended method of HE, can adjust the contrast of images according to the potential visual importance of input images. In this method, the potential visual importance of each pixel is computed by using its dark-pass filtered gradients. The dark-pass filtered gradients $\psi(h)$ is computed by

$$\psi_l(h) = -\sum_{h' \in \mathcal{N}(h)} \min\left(\frac{I_l(h) - I_l(h')}{K}, 0 \right), \tag{7}$$

where $h = (x, y)$ denotes the coodinates of a pixel, $\mathcal{N}(h)$ is a set of neighboring coodinates of h, I_l is obtained by downsampling the input image $I(h)$ by a factor 2^l. By using Eq.(7), the potential visual importance Φ is computed as

$$\Phi(h) = \left(\prod_{l=1}^{L} \max(U(\psi_l(h)), \varepsilon) \right)^{1/L}, \tag{8}$$

where $U(\cdot)$ denotes an upsampling operator, a modified PDF is obtained by using Φ, as follows:

$$pdf_c(k) = \frac{\sum_x \sum_y \Phi(h) \delta(I(h), k)}{\sum_x \sum_y \Phi(h)}, \tag{9}$$

where $\delta(I(h), k)$ is the Kronecker delta. The value of $pdf_c(k)$ is changed only when the luminance $I(h)$ equals to k. Therefore, the final transformation T is given by

$$T(k) = \sum_{k'=0}^{k} K \cdot pdf_c(k'). \tag{10}$$

D. Histogram Matching

Histogram Matching (HM) is the transformation of an image so that its histogram matches a reference histogram. Thus, choosing a reference image, other images are transformed so that their histograms match the histogram of the reference image. Note that we may directly choose a reference histogram such as Gaussian distribution, although a reference image can be chosen. A lot of machine learning models assume that input data follow Gaussian distribution.

IV. Experiment

In this experiment, we focused on a SVM algorithm and random forest as one of machine learning algorithms. To confirm the effectiveness, the above image processing methods were applied to a face recognition algorithm with SVM / random forest classifiers, where SVM was carried out as a dual problem. We evaluated the proposed method in terms of True Positive Rate (TPR) and True Negative Rate (TNR) in this paper.

A. Data Set

We used Extended Yale Face Database B [8] that consists of 2432 frontal facial images with 192×168-pixels of 38 people (see Fig.2). 64 images for each person were divided into 48 training ones and 16 test ones for each person.

B. Assessment

In the experiment, TPR and TNR were used to evaluate the performance. TPR and TNR are calculated by

$$TPR = \frac{TP}{TP + FN} \times 100[\%], \tag{11}$$

$$TNR = \frac{TN}{TN + FP} \times 100[\%], \tag{12}$$

where TP, TN, FN and FN represent the number of true positive, true negative, false positive, and false negative matches, respectively.

C. Experiment 1 (SVM)

We show the result of face recognition using SVM as a classifier.

1) experiment condition: In the experiment, a one-versus-the-rest classifier was trained for each class, that is, 38 SVM classifiers were trained in total. In addition, we utilized a linear kernel function for SVMs. As a feature vector, we used a vector obtained by concatenating the columns of an image. Here, all images were resized from original 192×168 pixels to 32×32 pixels.

2) Result and Discussion: Table 1 shows results under various pre-processing conditions, where the condition of Pre-processing 1 (skip) and Pre-processing 2 (z-score normalization) corresponds to the conventional one, and HM (Gaussian) indicates that HM was carried out with Gaussian distribution as a reference histogram. From Table 1, it is first confirmed that two processings: pre-processing 1 and pre-processing 2 are absolutely effective for improving the performance of the SVM classifiers. Moreover, the use of pre-processing 1 led to a higher performance, even when z-score normalization was carried out as pre-processing 2. The best scores were obtained when two pre-processing methods performed.

TABLE I: Experiment Result (SVM)

Pre-processing 1	Pre-processing 2	TPR(%)	TNR(%)
skip	skip	93.26	99.82
HE	skip	98.36	99.96
AGCWD	skip	94.08	99.84
CACHE	skip	37.66	98.32
HM (Gaussian)	skip	98.68	99.96
skip	z-score normalization	96.22	99.90
HE	z-score normalization	98.68	99.96
AGCWD	z-score normalization	98.52	99.96
CACHE	z-score normalization	98.19	99.95
HM (Gaussian)	z-score normalization	98.52	99.96

TABLE II: Experiment Result (random forest)

Pre-processing 1	Pre-processing 2	TPR(%)	TNR(%)
skip	skip	96.71	99.91
HE	skip	96.38	99.90
AGCWD	skip	97.37	99.93
CACHE	skip	97.70	99.94
HM (Gaussian)	skip	95.07	99.87
skip	z-score normalization	96.71	99.91
HE	z-score normalization	96.22	99.90
AGCWD	z-score normalization	97.37	99.93
CACHE	z-score normalization	97.70	99.94
HM (Gaussian)	z-score normalization	94.90	99.86

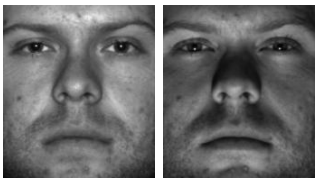
D. Experiment 2 (Random forest)

Random forest is an ensemble learning method for classification, regression and other tasks. A multitude of decision trees is used as weak learners in random forest.

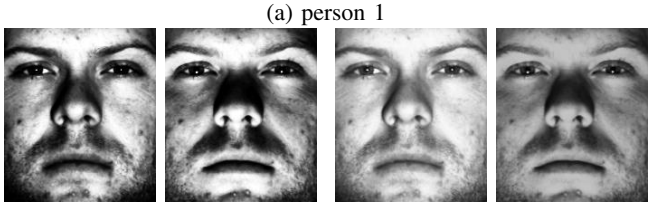
Random forest is an ensemble of decision trees. The decision tree is invariance when the magnitude relation of each feature is not changed. z-score normalization does not change the magnitude relation of each feature quantity, so random forest would not be influenced by z-score normalization. In this section, we show the result of face recognition using random forest as a classifier.

1) experiment condition: The output of the decision tree was a class, and the identification used its majority vote. The number of trees was 100, and the number of division was 10. We used 'bootstrap aggregation' on the ensemble of decision trees used for classification.

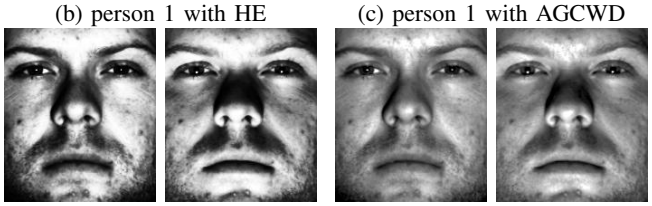
2) Result and Discussion: From Table 2, it is confirmed that the TPR and TNR values were not effected by z-score normalization. This is due to the fact that z-score normalization does not change the magnitude relation of each feature quantity as mentioned before. Moreover, the use of AGCWD and CACHE in pre-processing 1 led to a higher performance than the other case regardless of z-score normalization. Therefore, some contrast enhancement work well in pre-processing. Both of the best TPR and TNR scores were obtained when CACHE was applied in pre-processing 1.



(a) person 1

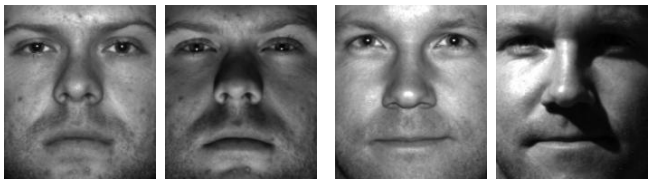


(b) person 1 with HE (c) person 1 with AGCWD

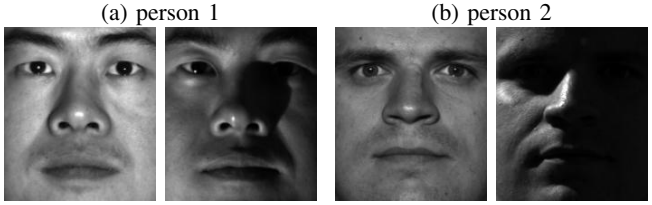


(d) person 1 with CACHE (e) person 1 with HM

Fig. 2: Examples of pre-processing



(a) person 1 (b) person 2



(c) person 3 (d) person 4

Fig. 3: Examples of Extended Yale Face Database B

V. Conclusion

This paper has proposed to use some image processing methods as a pre-processing method for machine learning algorithms. In the experiments, a number of image enhancement methods were applied to a face recognition algorithm with SVM classifiers and random forest ones to demonstrate the effectiveness of the proposed approach. For SVM classifiers, it was better to apply both z-score normalization and contrast enhancement as pre-processing. On the other hand, for random forest classifiers, AGCWD and CACHE worked well, but z-score normalization did not improve the accuracy.

Conventionally, z-score normalization is one of the most frequently used pre-processing for machine learning. However, the usefulness depends on machine learning algorithms. Even in such a case, some image enhancement methods such as AGCWD and CACHE can improve the accuracy. We plan to apply the proposed approach to various machine learning algorithms including deep learning as a future work.

References

[1] S. Gong, S. J. McKenna, and A. Psarrou, "From images to face recognition," *Image Processing, Imperial College Press*, 1999.

[2] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[3] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] S. M. Weiss and C. A. Kulikowski, *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc., 1991.

[6] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[7] S. R. Gunn *et al.*, "Support vector machines for classification and regression," *ISIS technical report*, vol. 14, no. 1, pp. 5–16, 1998.

[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[10] T. Maekawa, Y. Kinoshita, S. Shiota, and H. Kiya, "Privacy-preserving svm processing by using random unitary transformation," vol. 41, no. 28, pp. 13–18, 2017.

[11] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised leaning," *International Journal of Computer Science*, vol. 1, no. 2, pp. 111–117, 2006.

[12] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[14] Y.-T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE transactions on Consumer Electronics*, vol. 43, no. 1, pp. 1–8, 1997.

[15] S.-D. Chen and A. R. Ramli, "Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation," *IEEE Transactions on consumer Electronics*, vol. 49, no. 4, pp. 1301–1309, 2003.

[16] M. D. Abràmoff, P. J. Magalhães, and S. J. Ram, "Image processing with imagej," *Biophotonics international*, vol. 11, no. 7, pp. 36–42, 2004.

[17] C. CHIENCHENG, K. Yuma, S. Sayaka, and K. Hitoshi, "An image contrast enhancement scheme with noise aware shadow-up function," in *Proc. IEEE International Conference on Consumer Electronics*. IEEE, 2018, pp. 185–186.

[18] Y. Kinoshita, T. Yoshida, S. Shiota, and H. Kiya, "Pseudo multi-exposure fusion using a single image," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*. IEEE, 2017, pp. 263–269.

[19] Y. Kinoshita, S. Shiota, and H. Kiya, "A pseudo multi-exposure fusion method using single image," *IEICE Trans. Fundamentals*, vol. 101, no. 11, 2018.

[20] S.-L. Lee and C.-C. Tseng, "Color image enhancement using histogram equalization method without changing hue and saturation," in *Consumer Electronics-Taiwan (ICCE-TW), 2017 IEEE International Conference on*, 2017, pp. 305–306.

[21] Y.-S. Chiu, F.-C. Cheng, and S.-C. Huang, "Efficient contrast enhancement using adaptive gamma correction and cumulative intensity distribution," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, 2011, pp. 2946–2950.

[22] X. Wu, X. Liu, K. Hiramatsu, and K. Kashino, "Contrast-accumulated histogram equalization for image enhancement," *IEEE*, pp. 3190–3194, 2017.

[23] A. N. Avanaki, "Exact global histogram specification optimized for structural similarity," *Optical review*, vol. 16, no. 6, pp. 613–621, 2009.

[24] F. Balado, "Optimum exact histogram specification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.