

Feature Selection in Source Camera Identification

Kai San Choi, Edmund Y. Lam, and Kenneth K. Y. Wong

Abstract—Source camera identification is the process of discerning which camera has been used to capture a particular image. In our previous work, we tackled the problem with a vector of thirty-six features to train and test the classifier. The features include the lens aberration parameters and statistical measurements from pixel intensities. In this paper, we focus on reducing the feature set by stepwise discriminant analysis. Simulation is carried out to evaluate the classifier's performance by using the full feature set, reduced feature sets and randomly selected feature sets. The results show that the reduced feature sets can decrease the processing time while also maintain or even improve the classification accuracy under some circumstances.

I. INTRODUCTION

With the advent of low-cost and high-resolution digital cameras, there is an increasing popularity of digital images. However, it is well-known that digital images can be manipulated easily by software and most of the time, the alternations leave no traces. The credibility of an image hinders the usefulness of digital images to be presented as news items or as evidence in court cases. As a result, in image forensics, one would like to ascertain the source and authenticity of a digital image. In this paper, we focus on distinguishing between images captured by a limited number of camera models.

An approach to solve the camera identification problem is to make use of the CCD noise patterns in digital cameras. Because of the manufacturing process, unique noise patterns (e.g. pixel non-uniformity, dust on lens, dark currents) may be introduced on the CCD sensor. Lukas et al. [1] proposed to use a Gaussian denoising filter to extract the pattern noise. The reference noise pattern of a camera is obtained by averaging a number of images. The source camera of an image is then determined by the correlation between its noise pattern and a candidate camera's reference noise pattern.

The camera identification problem can also be solved by extracting a set of features from images and training a classifier to distinguish images from different cameras. Kharrazi et al. [2] proposed to extract a number of features from pixel intensities in order to capture the differences in image processing methods among camera models. Demosaicing, gamma correction, color processing, white balance and compression are the standard processes in digital cameras [3]. However, the exact processing algorithms may vary from one manufacturer to another. Therefore, by capturing these traits, it is possible to classify the images from different camera models.

Authors are with the Department of Electrical and Electronic Engineering, University of Hong Kong, Pokfulam Road, Hong Kong {kaisan, elam, kywong}@eee.hku.hk.

In our previous work [4], we proposed to use lens radial distortion for the camera identification problem. As all lens elements inevitably produce some aberrations, they may leave some unique imprints on the images being produced. We incorporated our radial distortion parameters with the features proposed by Kharrazi et al. in the classification. Each image is represented by a vector of thirty-six features. We then use a support vector machine classifier (SVM) [5] to evaluate the success rate of the classification. A reasonably high accuracy is obtained in the classification.

Obviously, not all the features are equally important. Some of them may even be redundant and some of them are very susceptible to noise. They may decrease the classification accuracy and waste the computational resources. Therefore, in this paper, we aim at identifying a subset of features which better discriminate among cameras by the stepwise discriminant analysis [6]. We also investigate the impact of using the selected feature set on classification accuracy and computational performance.

This paper is organized as follows. In Section II, we first introduce our full set of candidate features. In Section III, we describe the details of the feature selection method. Experimental results for the three camera case are provided in Section IV. The future work and conclusion are presented in Section V.

II. CANDIDATE FEATURES

Two types of candidate features are constructed. The first type of features is lens radial distortion parameters. They are used to capture the geometric footprints left behind by the lens on the images. The second type of features is proposed by Kharrazi et al. They capture the photometric effect left by the color processing algorithms on the images.

A. Lens Radial Distortion

The lens radial distortion [7] causes straight lines in the object space to be rendered as curved lines on the film or camera sensor. It originates from the transverse magnification, M_T , which is the ratio of the image distance to the object distance. Since lens has a spherical surface, M_T is a function of the off-axis image distance, r , rather than a constant. In other words, a lens has various focal lengths and magnifications in different areas. Barrel distortion occurs when M_T decreases with r . Similarly, pincushion distortion occurs when M_T increases with r . Fig. 1 shows an example of barrel distortion and pincushion distortion.

Due to manufacturing cost, most lenses are limited to spherical surfaces with inherent radial distortion. The distortion can be corrected by manipulating the system variables

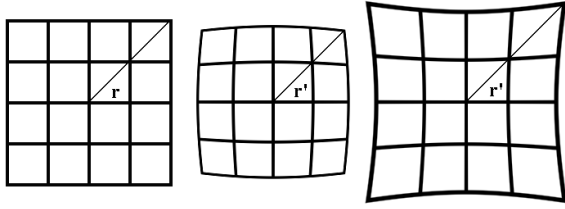


Fig. 1. Distortion of a rectangular grid. Left: Undistorted grid. Middle: Grid with barrel distortion. Right: Grid with pincushion distortion.

(indices, shapes, spacing, stops and etc.). However, the degree and order of compensation vary from one manufacturer to the others or even with different camera models by the same manufacturer. Therefore, lens from different camera models may have different degrees of radial distortion. Apart from the design, focal length also affects the degree of distortion [8]. Usually, lenses with short focal lengths have a larger degree of barrel distortion, while lenses with long focal lengths suffer more pincushion distortion. As a consequence, lenses from different cameras may leave unique imprints on the pictures being captured.

The lens radial distortion model can be written as an infinite series. In this paper, we only use the first and second order distortion parameters as an estimate of the degree of distortion in an image. The lens radial distortion can be written as:

$$r_u = r_d + k_1 r_d^3 + k_2 r_d^5 \quad (1)$$

where r_u and r_d are the undistorted radius and distorted radius respectively, and k_1 and k_2 are the first order and second order distortion parameters respectively. The radius is the radial distance $\sqrt{x^2 + y^2}$ of a point (x, y) from the center of distortion. The center of distortion is regarded as the center of an image in this paper. Devernay's straight line method [9] is used for computing the radial distortion parameters. An implementation of Devernay's algorithm in Matlab is publicly available [10]. We modified the program to estimate distortion parameters, k_1 and k_2 , from each image.

B. Features Proposed by Kharrazi et al.

Kharrazi et al. [2] proposed to extract thirty-four features from pixel intensities on an image. The features include average pixel value, RGB pairs correlation, center of mass of neighboring distribution, RGB pairs energy ratio, wavelet domain statistics and a set of Image Quality Metrics (IQM) [11][12]. It is believed that these features capture the differences in the color processing algorithms of different cameras.

III. FEATURE SELECTION METHOD

In this paper, we use an empirical method to select a subset of features. One simple statistical method to measure the discriminatory power of a feature is the analysis of variance (ANOVA) [13]. ANOVA is a statistical test for heterogeneity of means by analysis of group variances. It measures the F-ratio which is the ratio of the between-groups

variance over the within-groups variance. The larger the F-ratio is, the higher is the likelihood of that feature being more discriminative. However, the F-ratio only considers one individual feature at each time, and it provides no information about the interactions between features.

Stepwise discriminant analysis is a more advanced method based on ANOVA. It can overcome some the limitations of ANOVA, yet is still simple to use. In this paper, we use stepwise discriminant analysis to do the feature selection.

A. Stepwise Discriminant Analysis

In the stepwise discriminant analysis [6][14], features are chosen to enter or leave the model according to the significance level of an F-test (partial F statistic) from an analysis of covariance, where the features inside the model are covariates and the feature under consideration is the dependent feature. Stepwise selection begins with no features in the model. Initially, the F-ratio of each candidate features is calculated. The feature with the highest F-ratio is selected to the model. After the initial step, features are selected to move in or move out of the model by the significance level of the partial F statistic. If the feature in the model with the lowest partial F statistic has a lower value than the specified criterion (F-to-remove), that feature will be excluded from the model. On the other hand, if the feature outside the model with the highest partial F statistic has a higher value than the specified criterion (F-to-enter), that feature will be added to the model. The considerations of adding and removing features are taken alternatively. The selection process stops until all the features in the model exceed the F-to-remove and all the features outside the model are below the F-to-enter.

By using Monte Carlo simulations, Costanza and Afifi [14] recommends to select a minimum F-to-enter corresponding to a maximum α level, 4.0, and a minimum F-to-remove less than the F-to-enter, 3.9.

IV. EXPERIMENTAL RESULTS

The experiments are divided into two parts. The first part of the experiment is to identify a subset of features that is best discriminant to the cameras. The second part of the experiment is to verify the discriminatory power of the selected feature set.

In order to evaluate the result of feature selection on classifier accuracy and the computational time, three experiments were conducted. The first experiment included all the features (i.e. no feature selection) in the classification. The second experiment only included the selected features obtained from stepwise discriminant analysis. For the third experiment, we randomly selected the same number of features as the second experiment.

A. Cameras and Test Images

In our experiments, three different cameras were used. They are recent models from three different manufacturers. The Camera A and Camera B were used to produce 1600×1200 images, while Camera C was used to take 2560×1920 images. The images were taken with no flash, auto-focus, no

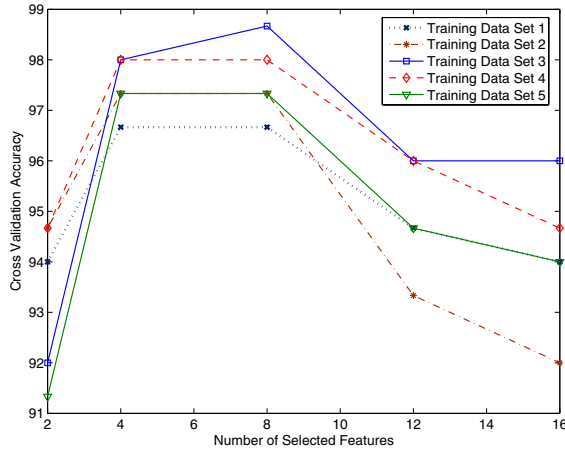


Fig. 4. Cross validation accuracy as a function of the number of features used. There is an increase in the accuracy by adding the number of features to 8. However, the accuracy starts to drop when the number of selected features is beyond 8.

TABLE II
FEATURES SELECTED IN THE 10 TRAINING DATA SETS AND THEIR NUMBER OF OCCURRENCES

Number	Feature	No. of occurrences
1	Lens radial distortion parameter k_1	10
2	Lens radial distortion parameter k_2	10
3	Spectral phase error	10
4	Czenkonowski correlation	9
5	Spectral magnitude error	8
6	Mean square error	8
7	Mean absolute error	4
8	Mean of vertical subband on green plane	4
9	Mean of diagonal subband on blue plane	3
10	Center of mass of neighboring distribution on red plane	3

rics (IQM), which provide qualitative data about the image quality by measuring the variation between the filtered and original image. It seems that they are independent of the changes in image content. On the other hand, the eighth to tenth features may be somewhat dependent on the image content because they are measurements of a particular color plane. As a result, they appear in lower ranks.

C. Classification Results for Each Feature Set

The first experiment used all thirty-six features in the classification. It acts as a control experiment to compare the differences in classification accuracy both with and without feature selection. In the second experiment, we used selected feature sets to perform the classification. In the third experiment, we randomly selected the same number of features as the second experiment to train and test the classifier. This experiment is also a control experiment and shows the difference in performance between stepwise feature selection

TABLE III
THE CONFUSION MATRIX FOR CAMERA IDENTIFICATION BY FULL FEATURE SET

		Predicted (%)		
		Camera A	Camera B	Camera C
Actual (%)	Camera A	91	9	0
	Camera B	11.6	87.8	0.6
	Camera C	0.6	0.4	99
Average Accuracy (%)	92.6			

TABLE IV
THE CONFUSION MATRIX FOR CAMERA IDENTIFICATION BY SELECTED FEATURE SETS

		Predicted (%)		
		Camera A	Camera B	Camera C
Actual (%)	Camera A	96.6	3.4	0
	Camera B	5.4	93.8	0.8
	Camera C	0.2	0.2	99.6
Average Accuracy (%)	96.67			

and random selection.

A SVM classifier was used in order to see the effectiveness of the feature selection. We use the SVM classifier available in the LibSvm package [15]. Each time, we used a training data set to train the classifier and used a corresponding testing data set to evaluate the classification results. The average classification accuracy of the 10 training and testing data sets in the first and second experiment are shown in Tab. III and Tab. IV respectively.

The average classification accuracy by the full feature set and the selected feature sets is 92.6% and 96.67% respectively. The selected feature sets had a slightly better performance because the feature selection process removed some noisy and redundant features. These features can degrade the accuracy of the SVM classifier, causing it to use less-than-optimal features in the classification [16][17]. As expected, the random feature sets had a relatively low accuracy (79.69%). This shows that the stepwise discriminant analysis is important in choosing a feature subset.

The processing time required for training and testing of each feature set is shown in Tab. V. The time records were normalized to the processing time required for testing the selected feature sets. The training time of the selected feature sets was about 15% less than that of the full feature set because the processing time of a classifier depends on the number features [16]. It is interesting that the training time of randomly selected feature sets were longer than that of the selected feature sets, though they had the same number of features. During the training of the SVM classifier, we require to search the parameters (C and γ) for the radial

TABLE V
TOTAL PROCESSING TIME REQUIRED IN THE 10 TRAINING AND
TESTING DATA SETS

Feature Set	Normalized Training Time	Normalized Testing Time
All Features	194.9	1.2
Selected Features	165.4	1.0
Randomly Selected Features	191.4	1.1
Note: The time is normalized to the processing time required for testing the selected features.		

basis function kernels. If the features contradict with each other, it may take more iterations to settle the parameter selection. A simple experiment can be done to verify this. Therefore, a good reduced feature set can reduce the training time by decreasing the amount of data input to the classifier and the number of iterations in the parameter selection.

V. CONCLUSION

In this paper, we investigate the problem of feature selection in source camera identification. We propose to use an empirical method to select a subset of features in order to improve the classifier's performance. Stepwise discriminant analysis is used to identify a subset of useful features. Simulations were done on comparing the full feature set, reduced feature sets and randomly selected feature sets. The results show that the reduced feature sets can improve the classifier's accuracy and reduce the required processing time.

Although our reduced feature sets show encouraging results on the classification performance, they are not optimal feature sets. In stepwise discriminant analysis, the selection process does not consider the relationships between features that have not been selected during the selection process. Therefore, some important features may be excluded in the process. In the future, a more sophisticated feature selection method based on the features' statistics as well as features' theoretical meanings can be used to further improve the result.

REFERENCES

- [1] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan, "Determining digital image origin using sensor imperfections," in *Image and Video Communications and Processing*, 2005, vol. 5685 of *Proc. SPIE*, pp. 16–20.
- [2] Mehdi Kharrazi, Husrev T. Sencar, and Nasir Memon, "Blind source camera identification," in *International Conference on Image Processing*, 2004, pp. 709–712.
- [3] Jim Adams, Ken Parulski, and Kevin Spaulding, "Color processing in digital cameras," *IEEE Micro*, vol. 18, no. 6, pp. 20–30, 1998.
- [4] Kai San Choi, Edmund Y. Lam, and Kenneth K.Y. Wong, "Source camera identification using footprints from lens aberration," in *Digital Photography II*, 2006, vol. 6069 of *Proc. SPIE*, pp. 155–162.
- [5] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley, New York, 2001.
- [6] A. A. Afifi and S. P. Azen, *Statistical Analysis: A computer oriented approach*, Academic Press, INC., 1972.
- [7] Eugene Hecht, *Optics*, Addison Wesley, San Francisco, California, 2002.

- [8] Ben Tordoff and David W. Murray, "Violating rotating camera geometry: the effect of radial distortion on self-calibration," in *Proc. 15th International Conference on Pattern Recognition*, 2000, vol. 1, pp. 423–427.
- [9] Frédéric Devernay and Olivier Faugeras, "Automatic calibration and removal of distortion from scenes of structured environments," in *Investigative and Trial Image Processing*, 1995, vol. 2567 of *Proc. SPIE*, pp. 62–67.
- [10] P. D. Kovesi, *Matlab and Octave Functions for Computer Vision and Image Processing*, Software available at <http://www.csse.uwa.edu.au/~pk/research/matlabfns/>.
- [11] Ismail Avcibas, Nasir Memon, and Blent Sankur, "Steganalysis using image quality metrics," *IEEE Transactions on Image Processing*, vol. 12, no. 2, pp. 221–229, 2003.
- [12] Ismail Avcibas, Blent Sankur, and Khalid Sayood, "Statistical evaluation of image quality metrics," *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 206–223, 2002.
- [13] Marcello Pagano and Kimberlee Gauvreau, *Principles of Biostatistics*, Duxbury Thomson Learning, 2000.
- [14] *SAS/STAT User's Guide, Version 8*, SAS Institute, 1999.
- [15] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a Library for Support Vector Machines*, Software available at <http://www.csie.ntu.edu.tw/~ccjlin/libsvm>, 2001.
- [16] Daphne Koller and Mehran Sahami, "Toward optimal feature selection," 1996, Proceedings of the Thirteenth International Conference on Machine Learning, pp. 284–292.
- [17] Huan Liu, Edward R. Dougherty, Jennifer G. Dy, Kari Torkkola, Eugene Tuv, Hanchuan Peng, Chris Ding, Fuhui Long, Michael Berens, Lance Parsons, Zheng Zhao, Lei Yu, and George Forman, "Evolving feature selection," *IEEE Intelligent Systems*, vol. 20, no. 6, pp. 64–76, 2005.