Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009

THE APPLICATION OF INTELLIGENT SYSTEM TO DIGITAL IMAGE FORENSICS

CHENG-LIANG LAI, YI-SHIANG CHEN

Department of Informatics, Fo Guang University, R.O.C.
E-MAIL: CL.Lai@msa.hinet.net

Abstract:

Digital image capture devices, such as digital cameras or camera-equipped mobile phones, have become very common. Digital images also have the problem of being easy to edit and to tamper. As a result, digital forensics is now an important field in image processing. In this study, the features of images taken different cameras were used as the basis for determining the source of the digital images. The Genetic Algorithm was used to automatically select the most suitable and minimum number of features for the image content then the Support Vector Machine used for training and classification in order to identify the source cameras.

This study also used image editing software for the post-processing of images, including resizing , blurring and tamper in order to determine if the Genetic Algorithm selected features were still effective for identification after the images were tampered. The results showed that the features selected automatically using the Genetic Algorithm could not only use less features, but also achieved better identification rates for the source camera of the digital images, and save images to extract the time of features after Genetic Algorithm select optimal feature.

Keywords:

Image source identification; Genetic algorithm; Support vector machine; Feature extraction; Feature selection

1. Introduction

The dawn of the Digital Age has led to an explosion in the variety of devices capable of capturing digital images including digital cameras, mobile phones and video cameras. The hardware has also become increasingly advanced and consumer digital cameras with 10-Megapixel resolutions have become commonplace. With digital image capture devices becoming increasingly affordable, more and more people are now using basic equipment to produce or edit digital images. Back in the analog age, when cameras still used conventional film, after a photo was taken it had to be developed. This process was resistant to editing and forging so a prosecutor could rely on photographic evidence in court. In today's digital era, however, advances in technology means that digital images no longer have to be developed. A variety of cheap and powerful digital image editing and production tools are also available on the market. The increasing popularity of the Internet has also revolutionized how humans acquire and use information. Digital images and data can now be quickly spread and exchanged over the Internet and they are by nature easy to modify as well, so the reliability of digital images are now increasingly being challenged.

Many consumers have now replaced their conventional film camera with the digital camera, highlighting the lack among law enforcement agencies of an effective technique for identifying source cameras. While most cameras now record details such as camera model, ISO (International Organization for Standardization), date and time in the EXIF (Exchangeable Image File Format), this data can be easily modified or stripped through compression or using simple software tools. The EXIF is not a reliable way of identifying source cameras and cannot be used for forensic purposes [1].

Geradts et al. [11] proposed the use of defective pixels on the digital camera's CCD (Charge Coupled Device) sensor to determine the source camera of an image. This technique is unable to find defective pixels on high-end CCD sensors, and newer models of digital cameras now come with excellent compensation technology that renders this technique useless.

Bayram et al. [12] suggested using the CFA (Color Filter Array) interpolation algorithm to identify the source of a digital image. The method has difficulties with compressed images.

Another approach is to digitally watermark the digital image. While this is an effective method of identification, most digital cameras today do not have a real-time watermarking function. Even if a watermark was added, digital watermarks can still be maliciously tampered with so its signature is weakened or even removed altogether, making it impossible to effectively identify the source camera. Other more effective methods for identifying the

978-1-4244-3703-0/09/$25.00 ©2009 IEEE

source camera must therefore be developed.

For this study, the features of the image content are used to identify the source camera of an image. The GA (Genetic Algorithm) is also used to automatically search for the optimal features and the SVM (Support Vector Machine) in order to classify the source camera of a digital image.

2. The approach

In this study, we use the features within the image captured by the camera itself as the basis for identifying the source camera. In other words, the intent is to find features unique to that particular camera. Image processing techniques, GA and SVM were therefore applied to images taken by different cameras. First, the features of each photo were calculated then GA used to automatically find the optimize features. These features were then passed to the SVM for training and classification in order to identify the source camera of the image.

2.1. Extract feature

In our research, 33 image features were used as the basis for identifying the image source and GA used to automatically screen for the optimal features in order to achieve the best recognition rate. The features were divided into three types: color features, image quality features and wavelet domain. These were indicated respectively with a C, Q or W prefix. A detailed description of each prefix is provided below [2], [3], [4], [13]:

- *Color feature*

Color features refer to features related to the colors in the image that had not been subjected to signal processing. These include the mean, correlation coefficient, center of mass and energy ratio. These types of image color features have a total 9 features as shown below:

Average pixel value: C1, C2, C3
RGB pairs correlation: C4, C5, C6
Neighbor distribution center of mass: C7, C8, C9
RGB pairs energy ratio: C10, C11, C12

- *Image quality metrics*

Apart from color features, the quality of images varies between different cameras as well. We can generally determine the differences in the image quality of different cameras with the naked eye. For example, one camera might give sharper outlines but darker tones. Another camera might produce a brighter image and more vibrant colors but the outline is not as clear as the first camera. These visual differences can be used as identifying features using the Image Quality Metrics (IQM) proposed by Memom et al. [5]. Their types include pixel variance, correlation and spectrum, and there are 9 features in total:

Mean Square Error, MSE. Q1
Mean Absolute Error, MAE. Q2
Minkowski Difference, $\gamma = \infty$. Q3
Structural Content. Q4
Normalized cross correlation. Q5
Czekonowski correlation. Q6
Spectral magnitude error. Q7
Spectral phase error. Q8
Spectral phase-magnitude error. Q9
Block spectral magnitude error. Q10
Block spectral phase error. Q11
Block spectral phase-magnitude error. Q12

- *Wavelet domain statistics*

In image processing, an image may not only be represented in the spatial domain but also in the frequency domain as well. Commonly used methods for converting the spatial domain into the frequency domain include Fourier Transform, Discrete Cosine Transformation and Wavelet Transformation. When this type of features is processed using wavelet transformation, the individual frequency components of the image are extracted, producing a range of different high and low frequency bands. The three sub-bands with high frequency information were used in this study as they represented the more detailed elements within the image. 9 features were identified for this feature type: W1-W9.

2.2. Support Vector Machine

Conventional classification methods included neural networks, decision trees and the nearest neighbor method. In the past few years, Supply Vector Machine has entered common use and has demonstrated its effectiveness in many applications. SVM also offers better convergence than neural networks.

The Supply Vector Machine (SVM) is a learning method based on kernel functions that can be used for classification and non-linear regression. For classification applications, the main concept is to construct an optimal separating hyper-plane in a high dimension features. Through this plane, two groups of data can be effectively separated [6], [7]. In this study, the RBF (Radial Basis Function) kernel function is used to translate data to a high dimension features.

Parameter selection is an important issue for SVM as the parameters selected influence the accuracy of the prediction results. The Cross-Validation method was used in this study to select the optimal parameters. The combinations of C and γ parameters were subjected to the cross-validation method to determine the accuracy of each

parameter combination and determine the most suitable parameters for that particular data sample.

This study use LibSVM software [8] to analysis data, this software is currently used for classification on a frequent basis. For example, if it is given a stack of pre-classified data, after training with SVM a model is constructed. The model trained with SVM can then be given new data for classification prediction.

2.3. Genetic Algorithm

The Genetic Algorithm (GA) is one of the algorithms used for solving mathematical search problems and has applications for optimization engineering problems. By randomly producing multiple solutions and retaining the better solutions for further processing, after several generations the optimal solution can be determined. In a way, it simulates the survival of the fittest concept in evolution where only those most adapted to the natural environment survive to breed [6], [9].

Many conventional algorithms use a certain set of transfer rules to search for a certain point in space in order to determine the vector of the next point. The problem with such a point-to-point search method is the tendency to end up finding only the local optimum. The GA solution searches through multiple points in a random manner. It also uses selection/reproduction, crossover and mutation to find better generation. Compared to conventional algorithms, GA has a better chance of finding the global optimum. It also offers a reasonable algorithm for complex optimization problems with a large solution space.

2.4. The optimum features

In this study, GA's reproduction, crossover and mutation processes were used to select 33 features and find the optimum features in order to improve the identification rate and classification performance. The detailed calculations for this study's GA were as shown below:

- Initial population

Start with Randomly generated population of 10 33-bit chromosomes. As each bit is binary encoded the corresponding features can be extracted, as shown in Fig. 1.

- Fitness function

The fitness function in this study was the identification rate of SVM classification. The fitness function is calculated by

$$\text{Fitness function} = \frac{P}{N} \times 100\% \qquad (1)$$

where N is the total number of images classified, and P the number of correctly classified images.

- Reproduction

The elitism method was used for reproduction in this study. The top two chromosomes from each generation in terms of fitness function value were retained for the next generation.

- Crossover

Double point crossover was used in this study. Two points were chosen at random from the chromosomes selected for crossover processing, and the bits switched between them to produce two new chromosomes, as shown in Fig. 2. In this study, the probability of crossover was set to 0.6.

- Mutation

Mutation was implemented by selecting a chromosome at random for mutation processing. A random point within the chromosome is chosen at random and the bit string at the mutation point changed, as shown in Fig. 3. In this study, probability of mutation was set to 0.1.

- Termination conditions

This study used the GA to select the optimum features in order to achieve the optimum classification identification rate. The identification rate and the number of selected features must therefore both be considered for the termination condition, so in this study, the termination conditions were set as an identification rate of at least 95% and the number of features being less than 20. Both conditions must be satisfied for the algorithm to terminate. Fig. 4 shows the flow chart for this study.

3. Experiment and results

This study used the GA to automatically select the optimum features and achieve high identification rates from the 33 features [10]. The experimental steps are listed below:

1. Four cameras were each used to take two sets of images. One set contained 150 photos and other set contained 90 images.
2. 60 images from the 150 image set were selected at random as the training data while remaining 90 images were used as test data for the classification experiment.
3. When GA search identifies the optimum features, the other 90 images were used to verify the identification rate of the optimum features.

Using this methodology, this study designed five cases based on similar image contents, dissimilar image contents

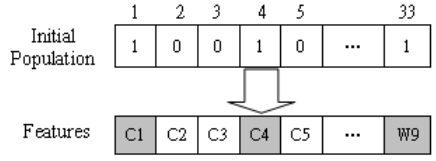


Figure 1. Initial population

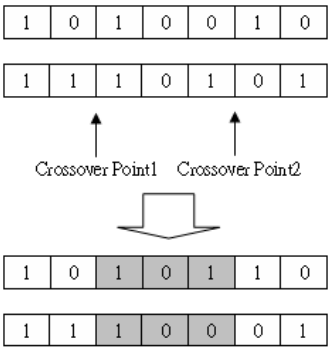


Figure 2. Double point crossover

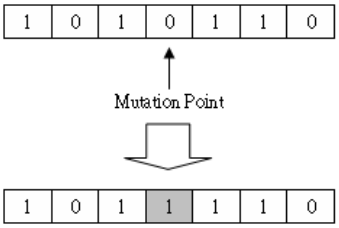


Figure3. Mutation

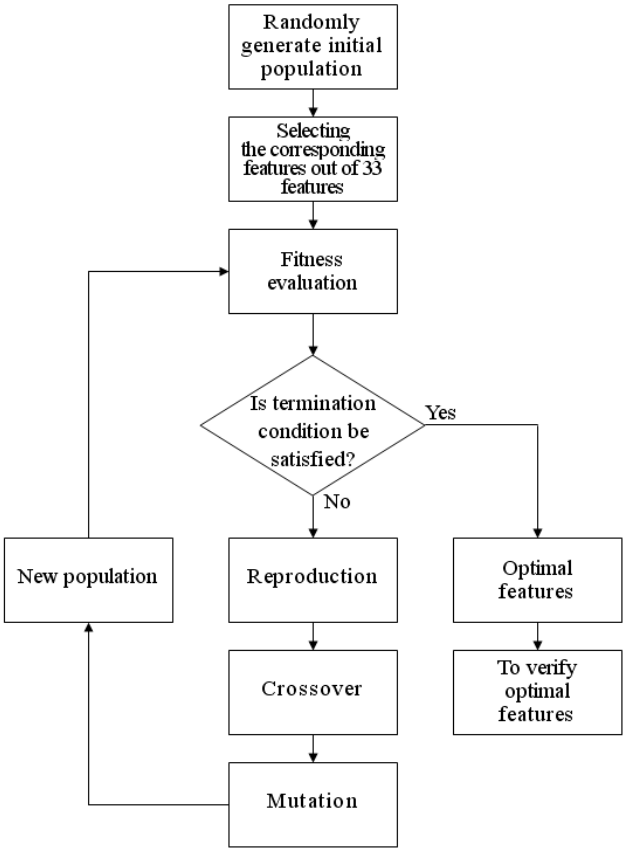


Figure 4. Research flow chart

and images were tampered with post-production software. This research, we also had discussion on time of operation in the case6.

- *Case1*

To verify that the GA's automatic search of optimum features contributed to the identification rate of the source camera for digital images, four different cameras (SONY T7, SONY T100, CANON 950IS and RICOH Caplio GX) were used to take photos of similar content as shown in Fig. 5 (a), (b), (c) & (d). The above methodology was then used for classification prediction. The experimental results were as shown in Table 1. The GA automatically located the optimum features from the 33 features and observations showed that there was no definite correlation between the type and the number of features selected. The GA did not tend to select a particular type of features because a particular feature offered a higher identification rate. The features selected automatically by the GA were able to achieve the same effect as the 33 features instead of being reduced because fewer features were used, as shown in Table 2.

- *Case2*

To determine whether the identification rate of the optimum features selected by the GA were affected by dissimilar image content, this study repeated the methodology from Case 1 and used the same cameras to take photos of dissimilar image content as shown in Fig. 6 (a), (b), (c) & (d). As shown in Table 3, the results showed that the optimum features found using the GA was able to achieve high identification rates as well and was not affected by differences in the image content. The features selected by GA delivered a certain level of identification ability for dissimilar image content and the average identification rates and the number of features used of the 4 different cameras were all superior to [10] as shown in Table 4. From the above results, it can be seen that differences in the images did not have an absolute influence on the identification rate of the features selected by the GA.

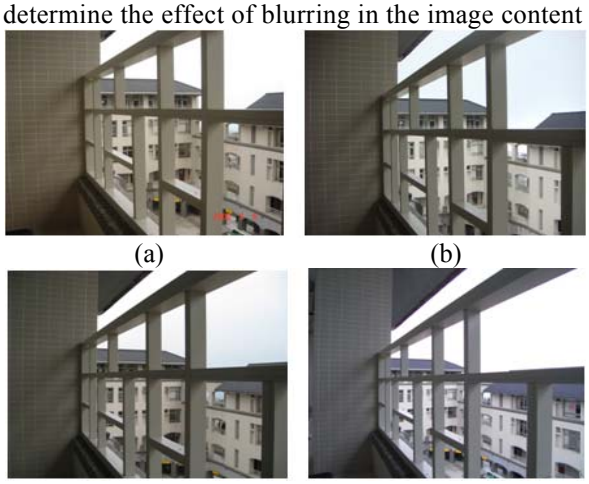
- *Case3*

Most people today use image-editing software to resize digital images for personal use. This study sought to determine whether resized images were still identifiable. Two different cameras (SONY T7 and CANON 950IS)

were used to take photos of similar content and the images resized using image-editing software from 1600x1200 to 1280x960, 800x600 and 480x320. Classification prediction was then conducted using the methodology described above. The results were as shown in Table 5-Table 8. A high level of identification rate was still achieved after the digital images were resized and the change in the size of the image did not affect the identification rate of the features selected by the GA. The GA-selected features can therefore be used to classify photos that were resized as well, and the images resized from 1600x1200 to 1280x960, 800x600 and 480x320, utilizes GA select number of features are 13, 17, 17 and 18 respectively.

- *Case4*

Although most digital cameras now come with image stabilization, there is still a chance that camera wobble may affect the clarity of some photos. The case therefore sought to determine the effect of blurring in the image content on identification rate. For this case, Photoshop's motion blurring effect was applied to the photos of dissimilar contents taken from Case 2. All the pixels in each image were translated horizontally, the displacement interval are 50 pixels, 75 pixels and 100 pixels respectively depending on the level. The greater the interval, the more blurred the image became. The original image as shown in Fig. 7(a), the blur image of displacement interval are 50 pixels, 75 pixels and 100 pixels respectively as shown in Fig. 7(b)-(d). The results of the experiment were as shown in Table 9-Table 11. The blurred content proved to not affect the identification rate of the features selected by the GA and a high identification rate was maintained for blurred photos, the results showed that while high identification rates were possible with blurred images, the level of blurring did affect the identification rate, and the blur image of displacement interval are 50 pixels, 75 pixels and 100 pixels respectively, utilizes GA select number of features are 19, 18 and 19 respectively.



(a) (b) (c) (d)

Figure 5. Image samples (a)-(d) with similar content

Table 1. These features be selected by GA

	Features of select by GA				
Color Features	C2	C4	C6	C7	C10
Quality Features	Q1	Q3	Q6	Q7	Q11
Wavelet Domain	W2	W6	W7		

Table 2. Accuracy of these features be selected by GA

		Predicted				Accuracy (%)
		SONY -T7	SONY-T100	CANON -950IS	RICOH	
Actual	SONY-T7	90	0	0	0	100
	SONY-T100	0	90	0	0	100
	CANON -950IS	0	0	90	0	100
	RICOH	0	0	0	90	100

- *Case5*

Image editing software is now in common use and many people use its personalize graphics to touch-up their own digital images. To determine whether the GA selected optimum features could still classify digital images with graphics added in post-processing, this study used the same images

(a) (b) (c) (d)

Figure 6. Image samples (a)-(d) with dissimilar content

Table 3. Accuracy of these features be selected by GA

		Predicted				Accuracy (%)
		SONY -T7	SONY-T100	CANON -950IS	RICOH	
Actual	SONY-T7	89	1	0	0	98.8889
	SONY-T100	2	88	0	0	97.7778
	CANON -950IS	0	0	90	0	100
	RICOH	0	0	0	90	100

Table 4. Comparison of average accuracy and number of features

	Average accuracy (%)	The number of features
Tsai, et al. [10]	93	33
Our research	99	19

from Case 3 and then added graphics that took up 2%, 4%, 6% and 8% of the total pixel count. The methodology described above was then used for classification prediction.

The experimental results were as shown in Table 12-Table 15. After graphics were manually added through post-production to the digital images, the GA-selected optimum values continued to maintain a high identification rate. The results of the experiment also showed that there was no absolute correlation between the identification rate and the size of the added graphics. Adding larger graphics did not lower the identification rate, and the image added graphics that took up 2%, 4%, 6% and 8% of the total pixel count, utilizes GA select number of features are 13, 13, 15 and 15 respectively.

- *Case6*

This research utilizes the GA select optimal features, reduce number of feature in order to save images to extract the time of features. In this case, we were discussion on GA spends time on select optimal feature and images to extract the time of features. In our research, the experiment is finished in the apparatus of the following hardware and software: Pentium 4 3.2GHz, 2GB Memory, Microsoft Windows XP Professional and MATLAB 7. In the case1, GA spends 4 hours on select optimal feature, calculating optimal features of images take 22 minute, and calculating 33 features of images take 49 minute. In the case2, GA spends 4 hours on select optimal feature, and calculating optimal features of images take 32 minute. From the above, GA spend long time on select optimal feature, but save image to extract the time of features after select optimal feature, and achieved better identification rate.

Table 5. Accuracy of original images size 1600 × 1200

		Predicted		Accuracy (%)
		SONY-T7	CANON-950IS	
Actual	SONY-T7	90	0	100
	CANON-950IS	0	90	100

Table 6. Accuracy of the image after resize to 1280 × 960

		Predicted		Accuracy (%)
		SONY-T7	CANON-950IS	
Actual	SONY-T7	90	0	100
	CANON-950IS	1	89	98.8

Table 7. Accuracy of the image after resize to 800 × 600

		Predicted		Accuracy (%)
		SONY-T7	CANON-950IS	
Actual	SONY-T7	90	0	100
	CANON-950IS	5	85	94.4

Table 8. Accuracy of the image after resize to 480 × 320

		Predicted		Accuracy (%)
		SONY-T7	CANON-950IS	
Actual	SONY-T7	85	5	94.4
	CANON-950IS	0	90	100

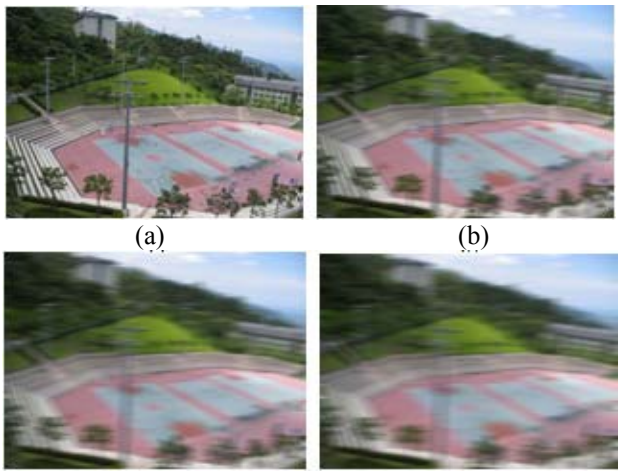


Figure 7. (a) original image, (b) The blur image of displacement interval 50 pixels, (c) The blur image of displacement interval 75 pixels, (d) The blur image of displacement interval 100 pixels.

4. Conclusions

In this study, image processing techniques and the SVM method is used to find the features of the relationship between different digital cameras and the images they capture. The features are then selected, crossed and mutated using the GA as a screening process to identify the optimal features. The source camera for an image can then be

Table 9. Accuracy where blurring was horizontal displacement by 50 pixels

		Predicted				Accuracy (%)
		SONY-T7	SONY-T100	CANON-950IS	RICOH	
Actual	SONY-T7	89	0	1	0	98.8889
	SONY-T100	0	90	0	0	100
	CANON-950IS	0	0	90	0	100
	RICOH	0	0	1	89	98.8889

Table 10. Accuracy where blurring was horizontal displacement by 75 pixels

		Predicted				Accuracy (%)
		SONY-T7	SONY-T100	CANON-950IS	RICOH	
Actual	SONY-T7	89	0	0	1	98.8889
	SONY-T100	0	90	0	0	100
	CANON-950IS	0	0	90	0	100
	RICOH	2	0	0	88	97.7778

Table 11. Accuracy where blurring was horizontal displacement by 100 pixels

		Predicted				Accuracy (%)
		SONY-T7	SONY-T100	CANON-950IS	RICOH	
Actual	SONY-T7	89	0	1	0	98.8889
	SONY-T100	1	89	0	0	98.8889
	CANON-950IS	0	0	89	0	98.8889
	RICOH	5	0	1	84	93.3333

Table 12. Accuracy where a graphic 2% of the size of the total pixel counts was added to the image

		Predicted		Accuracy (%)
		SONY-T7	CANON-950IS	
Actual	SONY-T7	88	2	97.7778
	CANON-950IS	0	90	100

Table 13. Accuracy where a graphic 4% of the size of the total pixel counts was added to the image

		Predicted		Accuracy (%)
		SONY-T7	CANON-950IS	
Actual	SONY-T7	90	0	100
	CANON-950IS	0	90	100

Table 14. Accuracy where a graphic 6% of the size of the total pixel counts was added to the image

		Predicted		Accuracy (%)
		SONY-T7	CANON-950IS	
Actual	SONY-T7	85	5	94.4444
	CANON-950IS	0	90	100

Table 15. Accuracy where a graphic 8% of the size of the total pixel counts was added to the image

		Predicted		Accuracy (%)
		SONY-T7	CANON-950IS	
Actual	SONY-T7	86	4	95.5556
	CANON-950IS	0	90	100

identified by training and classifying these features using the SVM.

From the above, it can be shown that the type of image features had no absolute bearing on classification identification rate. Features of a different type were not discarded because using a particular type of features produced a particularly low identification rate. An optimizing algorithm must therefore be used to make the best choices. In this study, the GA produced excellent results. Using the GA, good results could still be achieved with post-processed images, and save time of operation after utilizes GA to select optimal feature.

References

[1] K. S. Choi, E. Y. Lam, and K. K. Y. Wong, "Source camera identification using footprints from lens aberration", Proceedings of the SPIE, Vol.6069, pp. 60690J-1~ 60690J-8, 2006.

[2] Tran Van Lanh , K. S. Chong , Sabu Emmanuel and M. S. Kankanhalli, "A survey on digital camera image forensic methods", IEEE ICME2007, PP. 16-19, July 2-5, 2007.

[3] M. Kharrazi, H. T. Sencar, and N. Memon, "Blind Source Camera Identification", Proc. ICIP' 04, Singapore, October 24-27,2004.

[4] M. J. Tsai and G. H. Wu, "Using image features to identify camera sources", IEEE ICASSP2006, vol II, pp. 297-300, May 15-May 19, 2006, Toulouse, France.

[5] I. Avcibas, N. Memon, and B. sankur, "Steganalysis using image quality metrics", IEEE transactions on Image Processing, January pp. 221-229, 2003.

[6] C. L. Huang and C. J. Wang, "A GA-based feature selection and parameters optimization", Expert Systems With Applications Vol: 31, Issue: 2, pp. 231-240 August, 2006.

[7] Xiang-Yang Wang, Hong-Ying Yang and Chang Ying Cui, "An SVM-based robust digital image watermarking against desynchronization attacks", Signal Processing, Vol: 88, Issue: 9, pp.2193-2205 September 2008.

[8] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines", 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[9] S. S. J. Owais, P. Kromer, & V. Snasel, "Query optimization by genetic algorithms," Proceedings of Dateso, pp. 125–137, ISBN 80-01-03204-3, 2005

[10] M. J. Tsai, C. L. Lai and J. Liu "Camera/mobile phone

source identification for digital forensics.", IEEE ICASSP pp. 221-224 2007, Hawaii, USA.

[11] Z. J. Geradts, J. Bijhold, M. Kieft, K. Kurosawa, K. Kuroki, and N. Saitoh, "Methods for Identification of Images Acquired with Digital Cameras", Proc. of SPIE, Enabling Technologies for Law Enforcement and Security, vol. 4232, February 2001.

[12] S. Bayram, H. T. Sencar, N. Memon, and I. Avcibas, "Source Camera Identification Based on CFA Interpolation", ICIP 2005.

[13] C. L. Lai, Y. H. Chen and C. W. Chiu, "Source camera identification based on support vector machine with genetic algorithm", The 21th IPPR Conference on CVGIP 2008, August 24-26, 2008, Yilan, Taiwan