A system for forensic analysis of large image sets

Steven J. Simske, Margaret Sturgill, Paul Everest, George Guillory

HP Laboratories
HPL-2009-371

Keyword(s):
Image classification, image forensics, counterfeiting, security printing

Abstract:

Variable Data Printing (VDP) offers the ability to uniquely tag each item in a serialized list, which increases product security. However, it is often not cost-effective or practical to employ VDP, in which case each item is printed with the same set of images. The brand owner should be able to use such "static" printed areas to gauge if the printing was performed by an authentic or a counterfeit printer/label converter. In this paper, we describe a system that uses a small set of pre-classified images (either authentic or counterfeit images from the same source) for initial training, and thereafter adaptively classifies and aggregates images from multiple sources as they join the population to be classified. Authentic images and multiple sources of counterfeit images are identified, and secondary links between the non-compliant samples are provided. The system currently uses a set of 420 metrics which are filtered to a smaller set of metrics that can reliably describe our known set. This filtered set of metrics, or metric signature, is used for the search and clustering thereafter. We describe the use of this system to streamline and enhance investigations for a global brand protection program. For such an application, the system is tuned for high precision.

External Posting Date: December 17, 2009 [Fulltext] Approved for External Publication
Internal Posting Date: December 17, 2009 [Fulltext]

Presented at IEEE WIFS 2009, London, UK, Dec. 6-9, 2009.

A SYSTEM FOR FORENSIC ANALYSIS OF LARGE IMAGE SETS

Steven J. Simske[1] *Margaret Sturgill*[1], *Paul Everest*[2], *George Guillory*[3]

[1]Hewlett-Packard Laboratories, 3404 E. Harmony Rd., MS 36, Fort Collins CO 80528, USA
[2]Hewlett-Packard Co, 1000 NE Circle Blvd., MS 613B, Corvallis, OR 97330, USA
[3]Hewlett-Packard Co, 11445 Compaq Center Dr., MS M0301-107, Houston, TX 77070, USA

ABSTRACT

Variable Data Printing (VDP) offers the ability to uniquely tag each item in a serialized list, which increases product security. However, it is often not cost-effective or practical to employ VDP, in which case each item is printed with the same set of images. The brand owner should be able to use such "static" printed areas to gauge if the printing was performed by an authentic or a counterfeit printer/label converter. In this paper, we describe a system that uses a small set of pre-classified images (either authentic or counterfeit images from the same source) for initial training, and thereafter adaptively classifies and aggregates images from multiple sources as they join the population to be classified. Authentic images and multiple sources of counterfeit images are identified, and secondary links between the non-compliant samples are provided. The system currently uses a set of 420 metrics which are filtered to a smaller set of metrics that can reliably describe our known set. This filtered set of metrics, or metric signature, is used for the search and clustering thereafter. We describe the use of this system to streamline and enhance investigations for a global brand protection program. For such an application, the system is tuned for high precision.

Index Terms— ***Image classification, image forensics, counterfeiting, security printing***

1. INTRODUCTION

According to the recent World Economic Forum, counterfeiting currently comprises 8% of world trade [1]. This is a significant health, safety and security threat for the world's people. As with other broad security concerns, counterfeiting cannot be eliminated. A more practical and proactive plan is to attempt to determine the relative size of each counterfeiter, and target the largest counterfeiters for elimination first. Since counterfeit material shows up through different channels, we are working to create a method useful for identifying commonalities between these various samples. Samples exhibiting similar characteristics are logically suspected of originating from the same source. Selecting the proper sample sets for further investigation leads investigators to larger counterfeiting operations, potentially creating the largest possible disruption in the counterfeiting chain of operations for the amount spent in investigation and legal response. Optimizing return-on-investment (ROI) in brand protection is important to offset the disadvantages brands face, since counterfeiters do not spend money on brand development, marketing, quality control, etc.

To provide an effective ROI, we need to both classify suspect samples as authentic or counterfeit, and automatically detect clusters (aggregates of related images) in our suspect sample set. Although the system we have designed, built and deployed is not necessarily a final forensic tool, it is an effective front-end to traditional forensic analysis. The system discovers sets of samples that should be examined in greater detail, and determines the relative size of each potential counterfeit source: clusters formed are assumed to be part of the same counterfeit network, or at minimum having a supplier in common.

Our data to date supports this assumption. HP has deployed our approach internally to aid in its anti-counterfeiting efforts. While we cannot provide specific data on HP's counterfeit rates or intervention statistics, we describe how our access to "known" counterfeits is used to optimize our system settings. We describe the methods used in our "image forensics" in Section 2. Results are provided in Section 3. A brief discussion comprises Section 4.

2. METHODS USED

Variable Data Printing (VDP) is an excellent means to incorporate high-density security material directly on a physical object of value (package, label, document, ticket, etc.). In some cases, VDP may not be available or affordable, obviating the use of printing for authentication. However, due to the "variable" nature of static printing (color, ink, substrate, and finishing variability), even static printed areas can be analyzed to gauge authenticity of the print. By extracting such non-variable images and comparing them to known-authentic sets of the same images, we can determine the authenticity of an unassigned sample. For example, on HP's Inkjet cartridge packages, a suitable image is that of the girl that appears on a variety of SKUs (Figure 1).

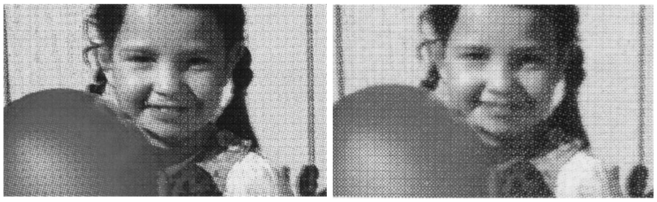

Figure 1. Scan of an image from an authentic sample (left) and a counterfeit sample (right).

Once an image of interest (IOI) is selected, a population of related samples is analyzed to initiate the system. These may be a selection of authentic samples of the same SKU, or a selection of samples seized from a single counterfeiter. Once those samples are processed, they will comprise the initial, or base, population.

2.1 Describing Samples

The initial version of the Image-Based Forensic System (IBFS) uses a feature set of 420 image metrics, from which an image class-specific "signature" of metrics is selected. These metrics, once defined, provide a description of the population of interest. The same set of metrics can be also used, comparatively, to find additional clustering information of the suspect samples.

The initial 420 metric set consists of 42 metrics that are computed for each of ten representations of the image. These "maps" transform the original RGB (red, green, blue) images as follows: (1) R channel, (2) G channel, (3) B channel, (4) Cyan, C = (G+B-R+255)/3, channel, (5) Magenta, M = (R+B-G+256)/3, channel, (6) Yellow, Y = (R+G-B+255)/3, channel, (7) Hue, (8) Saturation = max(R,G,B) * (1 – min(R,G,B)/sum(R,G,B)), (9) Intensity = (R+G+B)/3, and (10) Pixel Variance ("edge" space), the latter defined as the as mean difference (in intensity) between a pixel and its four diagonally closest neighboring pixels. For each of these 10 maps, there are 42 metrics, assigned to one of three broad sets of metrics:

(1.) Histogram metrics: Mean, Entropy, StdDev, Variance, Kurtosis, Pearson Skew, Moment Skew, 5% Point (value of index in histogram below which 5% of the histogram lies), 95% Point (value of index in histogram below which 95% of the histogram lies), and 5% to 95% Span.

Projection Profile Metrics for both (2.) the horizontal and (3.) the vertical profiles: Entropy, StdDev, Entropy, Delta StdDev, Mean, Mean Longest Run, Kurtosis, Skew, Moment Skew, Delta Kurtosis, Delta Pearson Skew, Delta Moment Skew, Lines Per Inch, Graininess, Pct In Peak, Delta Mean. For the "Delta" metrics, the differences between consecutive profiles in the projection data are used as the primary statistics. These metrics are computed for each IOI.

Once metrics are computed for each of the samples, the expected population description for each of the metrics is computed. The set of 420 metrics is a large set to quickly comb through with traditional methods such as SVD, PCA, etc. It is distilled to a more manageable set rapidly by taking into account the fact that each of the metrics is not fully independent. Our goal is to select a set of metrics that are most independent.

The final set of metrics is selected based on the individual metric independence; that is, $1.0-\mu(\rho_i^2)$, where $\mu(\rho_i^2)$ is the mean correlation of the metric with the set of other metrics. The metrics are initially grouped together in logical (related) sets to extract the most independent metrics. For example, we correlate histogram Kurtosis, Moment Skew and Pearson Skew and select the metrics according to Equation 1. The same method is used then to prune on the metric set level, and then one more time on all of the remaining metrics. This approach provides a fast filtering mechanism and a variety of metrics for the signature. These metrics can be selected using different filtering approaches (e.g. favoring at least one metric from each map), with similar performance. Regardless, for purposes of statistical testing, the final set of metrics obeys Equation 1:

$$1.0 <= \sum i=1\ldots S\ \mu(\rho_i^2) <= 2.0 \text{ and} \approx 1.0 \qquad \textbf{Equation 1}$$

That is, we add metrics based by their ranked mean independence until the sum just exceeds 1.0. The final set of "surviving" N metrics, along with their population statistics, provides the description of a referent sample population.

2.2 Classifying samples

This set of "surviving" N metrics is then compared statistically to the different aggregated (known) populations of images. Each statistical comparison (t-test) is performed with a pre-set p-value of α, and a sample differs from a known population with statistical significance for M (where 0 <= M <= N) of these metrics. We now wish to know if the sample is statistically significantly different from the known population. For this, there are two settings: (1) α, as described above, and (2) the experiment-wise confidence, or ε. We then note that the probability that a given number of metrics, M, which varies from 0 to the number of metrics, N, will be statistically significant for a new sample is given by:

$$\alpha^M *(1- \alpha)^{N-M} \qquad \textbf{Equation 2}$$

which is then multiplied by the number of combinations of M metrics that are statistically significantly different, which is ${}_N C_M$, or N!/[M!*(N-M)!]. Thus, for each value of M, the probability that a random sample will have M metrics statistically significantly different at probability α is:

$$\alpha^M *(1- \alpha)^{N-M} * N!/[M!*(N-M)!] \qquad \textbf{Equation 3}$$

Finally, the minimum M required for which the experiment-wide error rate (ε), or the accepted probability of making a Type I error in declaring a new sample different, is:

$$\Sigma_{M\ldots N}\{\alpha^{M} *(1- \alpha)^{N-M} * N!/[M!*(N-M)!]\} < \varepsilon$$ **Equation 4**

Equations 2-4 describe the means by which different sample groups are compared, and how new (unclassified) samples are compared to existing sets of samples. Additional aspects of the service are outlined in Section 2.4.

2.3 Aggregating Samples

To aggregate samples, an additional metric MWMD, based on the mean weighted distance of metrics defining the referent set, is calculated. A sample can belong to an aggregate only if its MWMD belongs to the population of the aggregate's MWMDs.

$$MWMD = \sum \left(\frac{\mu_s - \mu_p}{\sigma_p} \right) \times W_{metric}$$ **Equation 5**

This metric, MWMD, aggregates samples with an existing constrained aggregate (aggregates that are specifically defined as belonging together; e.g., known authentic or known counterfeit images). If the MWMD belongs to the population, then the sample also belongs to the aggregate.

When adding samples to an unconstrained aggregate (an aggregate created automatically by the software without user interference), the system still uses MWMD to group the samples. However, some assumptions must be made. Since we need a metric population to calculate MWMD values, a known constrained aggregate is used as a referent. By calculating the MWMD using the referent aggregate's metric population, groups of samples that cluster in the MWMD space can be considered for aggregation. Figure 2 shows the MWMD distribution for several sample populations in a single aggregate space.

Since the MWMD collapses the large set of metrics to a single primary metric, it is not sufficient to cluster by MWMD only. While two samples can have the same MWMD, they can be far away from each other in metric space (see Figure 3). To get around this problem, a second aggregation step is performed: the new aggregate must be self-consistent. Self-consistency is determined by calculating the best set of metrics describing the new aggregate and then verifying that the MWMD population of the new aggregate has no outliers. The speculative aggregate is then recursively pruned (i.e., outliers are removed and the signature is re-calculated) until there are no outliers or there are too few samples in the aggregate. Samples that were pruned out are returned to the unassigned samples pool. Thus, the system is dynamic (changing with each new sample added) and learning (increases its confidence as aggregates change in size and correlation).

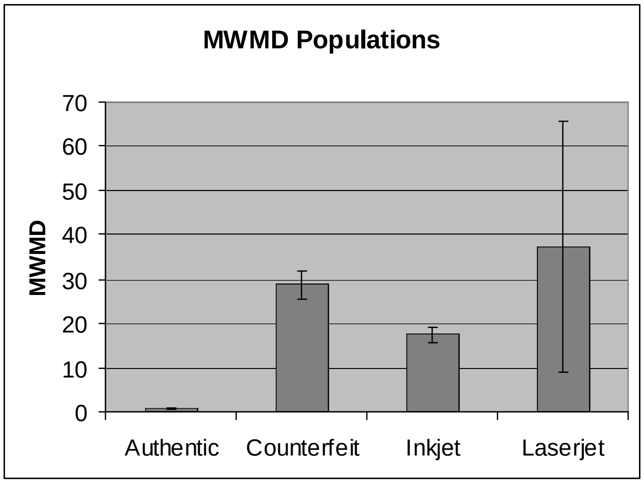

Figure 2. MWMD mean and standard deviation for several sample populations in the metric space of the *Authentic* aggregate. Since the metric space used is calculated based on the *Authentic* aggregate, the corresponding *Authentic* MWMD values are very small (0.75 ± 0.1). Note the lack of overlap in comparing the *Authentic* and *Counterfeit* samples.

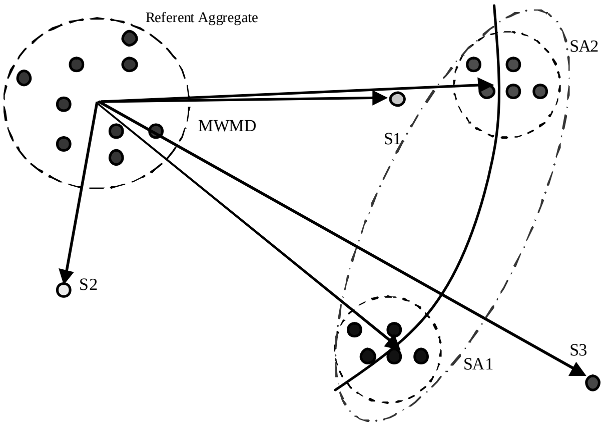

Figure 3. Samples with the same MWMD should not necessarily be aggregated together. Sample sets SA1 and SA2 have the same MWMD (single aggregate in MWMD space) but are far away from each other in the full metric space.

2.4 System Description

The system consists of two parts. A sample database and an aggregation system (see Figure 4 for a simplified diagram). The database contains all samples collected, their pedigree and any other metadata associated with them. The data base keeps track of known populations in the sample collection. These aggregates are not necessarily comprised of authentic or counterfeit samples, but may be aggregated based on properties of interest to the end-user.

To determine links between samples one or more referent aggregate(s) is(are) selected from the database. Then the samples of interest are determined and the initial aggregation is performed. The new aggregates and are then output and presented to the operator. If warranted, the new aggregates can be stored in the database for future use.

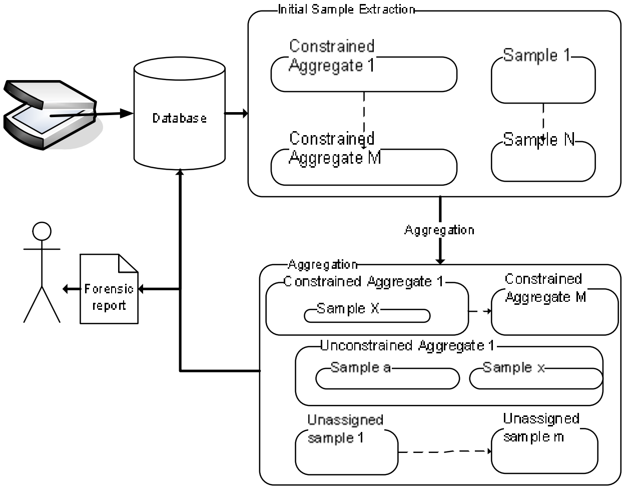

Figure 4. Simplified system diagram for the Image-Based Forensic Service.

3. RESULTS

A common scenario for the use of IBFS is the classification of samples into pre-determined, or “known”, groups. In this example, we start with 30 authentic images of the girl with a ball and 29 actual counterfeits. We then create two constrained aggregates with 15 authentic and 15 counterfeit samples, respectively. We then added the remaining 15 authentic and 14 counterfeits into the system. The classification result is shown in Table 1.

Set	Authentic	Counterfeit
Training set (constrained aggregate)	15	15
Number of samples correctly classified	15	8
Number of samples incorrectly classified	0	0
Number of samples unclassified.	0	6

Table 1. Classification results for simple two group classification.

The data in Table 1 have a recall rate of r=.79, a precision of p=1.0, and a corresponding accuracy, $(2pr/(p+r))$, of A=.88. The above example leaves several counterfeit samples in the “unclassified” pool. This is acceptable, since, based on preferences of HP’s global security operations, 100% or near-100% precision is preferred at the expense of recall. That is, preferred output of the IBFS is to have no false positives, as a false positive can lead investigators in a wrong direction. We are less concerned about unassigned samples, as they may aggregate precisely as more samples are added to the population.

The second major application of IBFS is the classification of samples when the only information available is a set of samples from one population. This is, in fact, the reason for the MWMD-based approach, and is a novel classification workflow. In this case, we are interested in whether samples cluster correctly when we have training data only for one of the classes. For an example of this workflow, we start with the following 4 classes:

Authentic (30 samples)
Real Counterfeits (29 samples)
In-house generated Inkjet (IJet) copies (59 samples)
In-house generated LaserJet (LJet) copies (60 samples)

Table 2, below, shows the actual results.

Group	Authentic	Counterfeit	IJet	LJet
Training set (constrained aggregate)	20			
Number of samples correctly classified	10	25	59	54
Number of samples incorrectly classified	0	0	0	0
Number of samples unclassified	0	4	0	6

Table 2. Classification of 4 classes in a presence of training data from one class (authentics).

For this problem, the trained data is classified with r=1.0, p =1.0, A = 1.0. For the untrained data, the classification statistics are r= .9324, p=1.0, and A = .965. The overall classification has the statistics r=.9367, p=1.0, and A=.9673. These results compare favorably with reported data on multi-class classification [2] and even binary image classification [3].

In all of our experiments, the IBFS is readily tuned to achieve 100% or near-100% precision. In sample populations where MWMD in the referent aggregate space is discontinuous (comprised of two or more logical groupings), logical groupings are often assigned to (multiple) aggregates that are actually subsets of the original grouping. While we can adjust the sensitivity to the MWMD discontinuity to limit the emergence of such sub-grouping, this has a tendency to lower the precision as it can cause two unrelated classes to merge together. For these situations, of course, the highest overall accuracy is normally achieved by breaking original groups into logical sub-groupings prior to initialization of the IBFS.

4. DISCUSSION

Counterfeit detection in printed materials can be approached in multiple ways. Buchanan et al. propose a unique “fingerprint” for each document to be authenticated based on its physical properties [4]. Clarkson et al. show a way to fingerprint a specific piece of paper [5], while Mikkilineni et al. concentrate on identification of the printer [6]. In our case, a more general approach is taken. Rather than examining a specific sample (though this can be also performed) and performing forensic-level analysis to confirm its authenticity, the IBFS is used to flag non-obvious similarities between samples, thus helping investigators interested in pursuing the largest counterfeiters at the onset.

Our approach to detection of counterfeits is not tied to a specific set of metrics. While our implementation uses 420 metrics and selects preferred set, other metrics can be readily used. The metric filtering process for the selection of aggregates can be based on partial independence as presented here. Additionally, filtering can be performed using an ontology or graph-driven selection (e.g. weighting the selection to ensure all histograms are represented).

The performance of the IBFS can be tuned to the needs of the end-user. Depending on the type of application, the IBFS can be tuned either for precision or recall. For example, in the case of pharmaceutical packaging—for which false negatives (not detecting counterfeits) should be minimized—the system is tuned to recall. In less critical applications, such as non-edible, non customer-threatening products (HP inkjet cartridges, for example), tuning for precision is generally more desirable to avoid unproductive investigations. In fact, tuning for precision is the deployment choice for HP’s printing supplies business--unassigned samples can be aggregated later when additional sample populations are entered into the IBFS.

The metric selection does not need to be automated as we have chosen here. Instead, the set of metrics used to describe the constrained aggregate can be selected based on a key encoded on the packaging, and can vary between printer/label converter, thereby allowing for print-shop specific “signatures” of metrics. Just as the metric winnowing can be customized for each of the clients, the actual metrics used could be changed. This flexibility is a consequence of our approach not being tied to specific metrics, and overall provides a novel approach to counterfeit detection by using minimum-training classification techniques. Our results to date, in fact, indicate that the IBFS performs as well (or better) when trained on a single aggregate (with an open-ended number of classes) rather than the more traditional training approach. For the latter, 96.7% accuracy was obtained. For the former, 88.0% accuracy—in line with 88-90% accuracy for multi-class classification [2]—was obtained.

Our approach is especially suitable for detecting non-compliant populations with low variance (i.e. presumably from a single source). High quality counterfeits (possibly from more sophisticated, larger-scale counterfeiters) will cluster more readily than poor reproductions on low quality paper. This is because in a low variance population, the MWMD metric will also have low variance, thus allowing for more exact clustering in the initial phase. For example, in Figure 2, the LaserJet samples have a large variance for the MWMD metric. Due to this high variance, several of the LaserJet samples cannot be classified and are left as unassigned in Example 2. In the case of the counterfeit samples (in the same Example 2), the sample population has a low variance in the MWMD for its metric population. Thus, the MWMD values in Figure 2 for counterfeit samples are very consistent, and are correctly clustered in the first aggregation attempt, even though several of the samples turn out to be outliers that are further away in the full metric space, and thus can be disaggregated . By manipulating how outliers are detected we can skew the system towards precision or recall as required.

Our current work is focused on optimizing the metric selection approach further, as well as further characterizing the impact of the classification parameters on aggregation precision, recall, and iterations to convergence. We are also integrating the technique into more traditional—inspection, quality assurance and authentication—brand protection image analysis workflows.

5. REFERENCES

[1] World Economic Forum, Update 2009: Threats to Security, http://www.weforum.org/en/knowledge/KN_SESS_SUMM_26734?url=/en/knowledge/KN_SESS_SUMM_26734.

[2] S.J. Simske, D. Li, J.S. Aronoff, ”Patterns for Using Multiple Classifiers in Image Classification”, IEEE ICIP 2009, submitted, 2009.

[3] A. Vailaya, A. Jain, and H.J. Zhang, "On image classification: city vs. landscape", Proc. IEEE Workshop Content-Based Access Image Video Lib., pp. 3-8, 1998.

[4] W.Clarkson, T. Weyrich, A. Finkelstein, N. Heninger,
J. A. Halderman and E. W. Felten, “Fingerprinting Blank Paper Using Commodity Scanners” IEEE Symposium on Security and Privacy 2009 May 17-20 2009.

[5] James D. R. Buchanan, Russell P. Cowburn, Ana-Vanessa Jausovec, Dorothée Petit, Peter Seem, Gang Xiong, Del Atkinson, Kate Fenton, Dan A. Allwood3 & Matthew T. Bryan, “Forgery: 'Fingerprinting' documents and packaging”, Nature 436,p. 475,28 July 2005.

[6] A. K. Mikkilineni, O. Arslan, P.Chiang, R. M. Kumontoy, J. P. Allebach, G. T.-C. Chiu, E. J. Delp, “Printer Forensics using SVM Techniques” IS&T's NIP21: International Conference on Digital Printing Technologies, Baltimore, MD; September 18, 2005; p. 223-226.