

# Optimized Selection of Benchmark Test Parameters for Image Watermark Algorithms based on Taguchi Methods and Corresponding Influence on Design Decisions for Real-World Applications

Tony F. Rodriguez, David A. Cushman  
Digimarc Corporation, Tualatin, OR 97062

## ABSTRACT

With the growing commercialization of watermarking techniques in various application scenarios it has become increasingly important to quantify the performance of watermarking products. The quantification of relative merits of various products is not only essential in enabling further adoption of the technology by society as a whole, but will also drive the industry to develop testing plans/methodologies to ensure quality and minimize cost (to both vendors & customers.)

While the research community understands the theoretical need for a publicly available benchmarking system to quantify performance, there has been less discussion on the practical application of these systems. By providing a standard set of acceptance criteria, benchmarking systems can dramatically increase the quality of a particular watermarking solution, validating the product performances if they are used efficiently and frequently during the design process.

In this paper we describe how to leverage specific design of experiments techniques to increase the quality of a watermarking scheme, to be used with the benchmark tools being developed by the Ad-Hoc Watermark Verification Group. A Taguchi Loss Function is proposed for an application and orthogonal arrays used to isolate optimal levels for a multi-factor experimental situation. Finally, the results are generalized to a population of cover works and validated through an exhaustive test.

Keywords: Digital Watermarking, Design of Experiments, Taguchi

## 1. INTRODUCTION

The migration of watermarking techniques from theory to product has necessitated the need to quantify the performance of algorithms and products in a much more holistic fashion. A validation of performance against a single image, such as “Lena”, through a few transformations, provides little to no insight to potential customers attempting to understand the relative strengths/weaknesses of various products for their particular application. The problem is even more acute when a consumer of the technology is attempting to understand the impact a watermarking algorithm will have on fundamental system factors such as cost, reliability, ease of use and even security. A positive sign that Digital Watermarking has matured and is making the transition from research labs to applications in the field is evidenced by the increasing need of customers and researchers to quantify performance.

As watermarking technologies are packaged (hardware, software, etc.) and transitioned to the field, they are subject to traditional Software Quality Assurance (SQA) methodologies to ensure performance. While these techniques can yield the quantitative data desired by a customer, they are not ideally suited to the large volume of data that has to be run to accurately and reliably validate performance.

Traditional SQA techniques include black box testing and code inspections. Black box testing presumes that the individual running the test has little insight or knowledge of the source code, and that the technology can be decomposed into discrete functions which, in turn, are black boxes themselves. Code inspections are a practice-based review of system source code and, while generally acknowledged as best practice, are dependent on significant involvement by the engineering team. Additionally, they are ideally suited for situations where only a few known stimuli can be presented to the system, resulting in a limited number of code-paths.

Typical applications of testing practice in industry have revolved around testing the watermarking technology as a single, large and potentially cumbersome black box functionality. Large numbers of images are used to attempt to statistically indicate freedom from defects. From an industrial standpoint, this is achievable but requires an enormous effort in gathering and maintaining these images, and running tests on a regular basis. The approach is problematic from

the standpoint that one can only indicate absence of defects in the samples - one has to assume via the central limit theorem that the test images are sampled from a population in which image characteristics (structural, other) are normally distributed.

Basic test cases consist of using a large numbers of random imagery as input test cases. The primary focus of the test effort then becomes building a maintainable test harness that allows many images to be pushed through the watermarking application, capturing and processing of results and ensuring that the entire process is replicable.

The research community has understood the unique requirements placed on testing tools by watermark applications, and has generated tools (StirMark [2], UnZign [3], etc.) to assist researchers & potential customers. These tools have enabled black box testing in a limited form with small populations of images. However, these testing tools were not created to address the maintenance and automation issues that arise when attempting to commercialize the technology. Recently though, several efforts have been initiated to address these concerns and to provide a large population of cover-works for testing. Three noteworthy efforts are: StirMark Benchmark [4], Certimark [5] and the Public Watermarking Verification Site being developed by the Ad-Hoc group as presented in Security and Watermarking of Multimedia Content track at the 2002 SPIE Conference.

Of these three efforts, the authors are most familiar with the Public Watermarking Verification Site. The Site provides a mechanism and toolbox to create automated tests that will aid the researcher in validating the performance of various algorithms/products against large numbers of images, and to capture and summarize the resulting data. Traditionally these tools have been developed by commercial entities for specific products; by contrast, the Verification Site is a public workbench built entirely on open source tools. Once operational, the Site will provide users with a wide array of tools with which to perform black-box tests. The communicated goal of the Site is to provide a common place for benchmark testing and to share new concepts and their related implementation. Ultimately this will serve the community by fostering the development of commercially viable watermarking technology.

The onus of the responsibility for the correct application of these tools resides with the researcher. Tools themselves, regardless of how advanced, cannot dictate their optimal usage to a user. The most obvious approach to validate the performance of a watermarking algorithm is to use the tool to do an exhaustive test of all factors and levels and through sheer volume quantify the performance and insure freedom from defects. This approach will eventually yield the needed answers, but it may take considerable time and compute resources. Our research, over several years has indicated that to accurately assess the state of a watermarking technology, millions of unique images are required.

A test plan that uses a methodology other than through exhaustive testing needs to be utilized to yield results in a more expedient fashion. Ideally the methodology will yield statistically significant results that accurately predict the conclusion from the full factorial.

### **1.1. Rationale**

The goal of this paper is to provide an example of how a test planning strategy may be implemented by using Design of Experiments (DOE) techniques to create appropriate and accurate test cases that produce statistically reliable results. An example of a fictional but realistic customer application and its related factors (uncontrollable) will be considered. Test cases will be generated to optimize an implementation of a watermark technology by identifying optimal levels for various controllable factors specific to the algorithm. Test cases will consist of images from the Facial Recognition Technology [6](FERET) database of facial images collected under the FERET program. Finally, the results of the designed experiments will be compared to the full factorial evaluation of all variables, demonstrating a significantly minimized test environment that yields results that statistically match the results of the full run of testing. It is hoped that this will be viewed as a superior approach to commercial test and validation efforts for digital watermarking technology, rather than a shotgun approach to test case development.

A description of a sample application is given in Sec. 2, an overview of DOE is provided in Sec. 3, DOE test plan in Sec. 4, and DOE test results are documented in Sec 5, followed by a comparison against full-factorial results in Sec. 6. Finally, the conclusion is in Sec. 7.

## **2. SAMPLE APPLICATION DESCRIPTION**

Before an appropriate test plan can be developed to leverage a benchmarking toolset such as the Public Verification Site, a solid problem statement must be defined. Typically this is created through multiple iterations between customers and vendor to ensure a common understanding of the problem and its unique challenges. Not surprisingly this process usually uncovers nuances to the problem statement that were not evident to either party at the beginning of the process.

Once a common understanding of the problem statement is arrived at, next comes the process of identifying what metrics will be used to measure the performance of a specific technique/product. These metrics, combined with a Quality Criteria (QC) that defines the desired result (bigger is better, smaller is better, nominal is better), become the output of the test to be considered.

These metrics can consist of anything that is pertinent to application at hand. For watermarking applications they usually include visibility, robustness, false positive detection, payload size and computational cost. Once the metrics are identified along with their QC, a benchmarking tool can be used to collect the results.

After the results are collected, the fundamental challenge of turning test results data into meaningful quality assurance information can be undertaken. While this step in the process comes last, it is the most important. Without mutual agreement between all the parties as to what defines a success versus a failure before the test is undertaken, the process can devolve into a process based on opinion, not fact. Deriving the goals of testing via definition of what constitutes meaningful results information will substantially increase the likelihood that the testing effort will consist of the correct set of test cases

This process is relatively self-evident; unfortunately it is a common occurrence for it to not be followed. This can be problematic when the performance envelope that defines success for many of these metrics is small and dependent on the interaction of other factors and levels. Since there will always be tradeoffs, understanding those tradeoffs in terms that relate to cost incurred by the customer and the vendor are critical to understanding whether the algorithm under test is optimal for the application at hand.

## 2.1. Problem Statement – Fictional Customer Application

A customer is considering deploying a system to increase security at a large research facility that employs 800+ researchers. Their primary focus is ensuring that the facility is only accessible to employees with the appropriate security clearance. As a primary mechanism they have distributed badges to the workforce, with each badge containing a photograph of the employee it is assigned to. All doors to the facility are equipped with badge readers that validate the badge before unlocking the door.

The customer has determined that it is sufficiently difficult to attack the current system (either through counterfeit badges, hacking servers, etc.) but that the system cannot validate that a person is using their assigned badge. The system validates badges, not people.

The customer has decided to enlist the guards that are currently located at entrances to match badges to people. Additionally, new guards will be hired to perform random spot checks of employees during patrols of the facility. Analysis of the badges showed that the images were susceptible to alteration, nullifying the value of the image on the badge itself. A system to retrieve employee photographs and display them to the guard is desired.

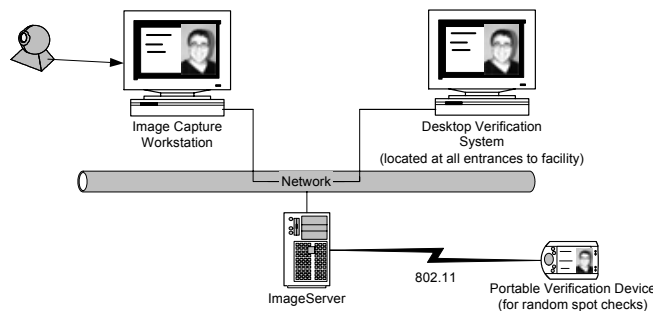


Figure 1. Employee Photo Retrieval System

A technology is needed to validate the authenticity of the image before being displayed. Of the various potential solutions, watermarking was identified as being a potential candidate; one that would be uniquely capable of identifying if an image had been tampered with.

The customer identified the following system constraints for the Portable Verification Device & the Desktop Verification System:

	<i>Desktop Verification System</i>	<i>Portable Verification Device</i>	<i>Details</i>
<i>Maximum Dimension for Image (in pixels)</i>	256x256 (dimensions of original image)	217x217 (when rotated 90 degrees)	Image is captured at 256x256, 8 bits per pixel. Display on portable device is smaller, forcing the image to be scaled to 85% of original and rotated 90 degrees.
<i>Image File Size</i>	No constraints	20KB	The portable device will use a 802.11 network that is heavily congested, so file size is a concern
<i>Cost of Device</i>	Must work with existing systems	< \$300	PC based systems are already located at entrances to the facility (Pentium 2 200 Mhz.) Preferred portable device contains a 100Mhz processor with no Floating Point Unit.

Table 1. System Constraints for Portable and Desktop Verification Systems

All image transformations will occur on the image server, with the exception of detection of the watermark. For security reasons, it was concluded that watermark detection would occur at the end nodes of the system (Desktop Verification System & Portable Verification Device) to ensure that images were not tampered with, refer to Figure 2.

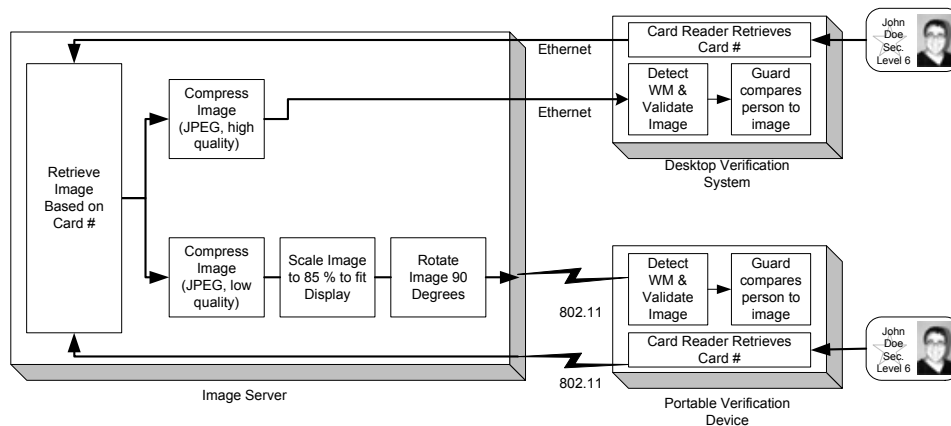


Figure 2. Functional Diagram of proposed Watermark System

## 2.2. Performance Metrics & Quality Criteria

As with many technologies it is a rare occurrence that one technology will completely displace another, even if the preferred technology is substantially better than what is currently deployed. In security applications the tendency is to layer technologies & solutions, resulting in a system that does not have a single point of failure.

This is the case with the sample application described. Watermarking is being used to augment an existing system and, as such, can be measured against its ability to thwart specific attacks. The range of attacks it might encounter is quite large, but the value-add of watermarking can only be understood for this application if the threat is well understood. For the application under discussion, the customer analyzed the threats to the proposed system and developed a simple plot that defines dollars lost versus complexity of threat (Figure 3.)

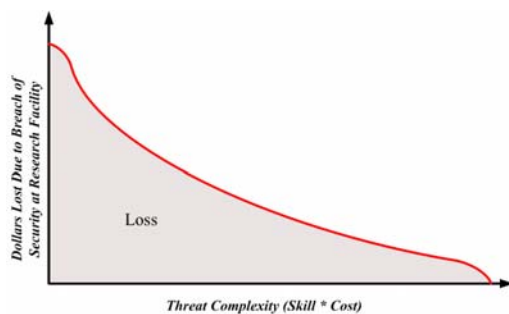


Figure 3. Currently losses incurred by breach of security at research facility using badges.

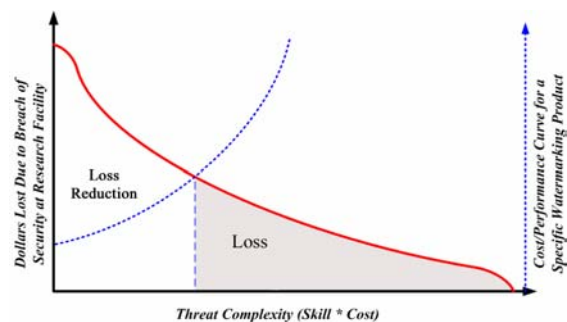


Figure 4. Reduced loss attributed to additional layer of security provided by watermarking.

A significant portion of the loss (area under the curve) is due to simple threats. Understanding the threats to the system, has allowed the customer to determine the value added by layering watermarking on top of their badge system, Figure 4. The dashed line in Figure 4, represents cost versus function for a proposed watermarking solution that meets the following (Table 2) performance criteria set forth by the customer.

	<b>Metric</b>	<b>Description</b>
<b>Visibility</b>	<=0.40	Average Watson metric for all employee photographs.  (Watson metric quantifies visibility of an image based on Just Noticeable Differences. A value of 1 implies that an expert viewer can determine a difference in the image)
<b>Detection Rate</b>	>=%99.8	Percent success when detecting a valid image
<b>Occurrence of False Positives</b>	<=1 in a million	Occurrence of an unmarked image detecting as valid
<b>Detection Time</b>	<=300msec	Maximum detect time for both desktop & portable applications.

Table 2. Customer Defined Watermark Performance Criteria

When operational the system will report one of two states after successfully detecting a watermarking in the image delivered by server.

1. Authentic Image
2. Suspicious Image

The states have a known cost to the customer: employee allowed access or additional investigation needed. For the sake of simplicity we will presume that the two states correlate directly to a metric that quantifies the confidence in the result generated by the watermark detector.

This metric is the Watermark Performance Metric (WPM) generated by the detector. The WPM is the performance metric that will be tracked and analyzed during testing with the explicit goal of maximizing the value to minimize the probability that a “Suspicious Image” result is incorrectly returned. It will also drive the Taguchi cost function that will quantify the cost to the customer and vendor of various technology options.

### 3. DESIGN OF EXPERIMENTS (DOE) AND TAGUCHI COST FUNCTION

Design of Experiments (DOE) consists of a combination of statistical techniques that were introduced in the book *Statistical Methods for Research Workers* by R.A. Fisher in 1925. Fisher was the first to develop a method that enabled the analysis of more than one effect from multiple variables at a time. Dr. Genechi Taguchi studied and perfected Fisher’s work while at Nippon Telephone and Telegraph Company during the 1940’s & 1950’s. By the 60’s & 70’s other Japanese companies had adopted the technique to optimize their design and production efforts. Of Taguchi’s many contributions to DOE, his work on test methodologies to compensate for noise factors and cost functions are utilized in this paper.

#### 3.1. Inner Array

Orthogonal arrays had been in existence prior to Taguchi’s involvement in DOE, but Taguchi simplified the use of arrays by defining standard arrays for various factor & level combinations. Orthogonal arrays can significantly reduce the number of trial conditions to be tested when compared to the full factorial of all factors at all levels. For a test that consists of 4 factors, each at 3 levels, the full factorial consists  $3^4 = 81$  trials. Using an L-9 array [7] (Table 3 ) results in only 9 unique trial conditions. The array defines which levels should be used for each of the trial conditions to ensure that there is equal opportunity for all levels to affect the outcome.

	<i>Factor A</i>	<i>Factor B</i>	<i>Factor C</i>	<i>Factor D</i>
<i>Trial 1</i>	1	1	1	1
<i>Trial 2</i>	1	2	2	2
<i>Trial 3</i>	1	3	3	3
<i>Trial 4</i>	2	1	2	3
<i>Trial 5</i>	2	2	3	1
<i>Trial 6</i>	2	3	1	2
<i>Trial 7</i>	3	1	3	2
<i>Trial 8</i>	3	2	1	3
<i>Trial 9</i>	3	3	2	1

Table 3. L-9 Orthogonal Array

Results from all trials (9 for an L-9 array) are averaged into the *grand average of performance*. This defines the baseline of performance with which to compare test results against. *Average effects* can be calculated for each level of a given factor by examining the orthogonal array and averaging all the results for that specific level. As an example, to determine the average effect for factor A at level 1, all results of factor A at level 1 are averaged, for an L-9 array this would consist of averaging the results for trials 1,2,3.

The *main effect* for a factor is the difference between the average effect for different levels of the factor. The delta between the main effect for a given level and the grand average of performance for the whole population yields the *factor contribution*. The factor contribution is used to identify which levels for a given factor result in the maximum increase in the desired result as defined by the *Quality Characteristic (QC)*.

The quality characteristic defines what the desired direction for the outcome of the test is. The QC can be one of three values; bigger is better, nominal is better, smaller is better.

### 3.2. Outer Array – Noise Factor

When uncontrollable noise factors are present in a process, an additional orthogonal array can be utilized. The additional array, the *outer array*, describes trial conditions with which to sample the noise. For each trial condition prescribed by the inner array, multiple experiments can be performed as described by the trial conditions in the outer array. This combination of an inner array and an outer array was first used by Taguchi as a design method to reduce variation in the result of a process due to uncontrollable noise or variation in the system. For watermarking applications, this ability to develop test methodologies that can be used during design to derive implementations whose performance is less dependent on the cover-works is of interest.

When multiple experiments are performed for each inner array trial condition, tracking the standard deviation for the results becomes possible. It is both the adherence to a target value and reduction of variation that reduces cost. To track both mean & standard deviation, *Mean-Squared Deviation (MSD)* is used instead of a simple average. With samples  $y_1, y_2, y_3, \dots, y_n$ , and a target value defined as  $y_o$ , the MSD for nominal is best QC is calculated as follows:

$$MSD = \frac{(Y_1 - Y_o)^2 + (Y_2 - Y_o)^2 + (Y_3 - Y_o)^2 + \dots}{n} \quad (1)$$

The MSD is reduced by an equal amount by a reduction in either the variance ( $\sigma^2$ ) of the samples or the difference between the average value of the samples and the target value ( $Y_{avg} - Y_o$ ). It can be shown that Equation 1 is equivalent to the following:

$$MSD = \sigma^2 + (Y_{avg} - Y_o)^2 \quad (2)$$

For experimental situations where the QC is bigger is better, such as in this paper, the following MSD equation is used:

$$MSD = \frac{\frac{1}{Y_1^2} + \frac{1}{Y_2^2} + \frac{1}{Y_3^2} + \dots}{n} \quad (3)$$

For reasons of convenience and increased linearity, the MSD values are converted to *Signal to Noise* (SNR). Once converted, the desired result is always a bigger SNR regardless of the QC for the experimental result.

$$\frac{S}{N} = -10 \log_{10} MSD \quad (4)$$

In this representation, SNR captures the intent of reducing noise due to uncontrollable factors in the system while increasing the signal due to controllable factors. For watermarking applications, noise may not only include variation in the cover-works but also transformations or attacks the cover-works have undergone.

Analysis of Variance (ANOVA) can be performed on the resultant SNR for each of the factors within the internal array. This process identifies which factors are responsible for what percentage of the total variance within the system. If the contribution of factors is deemed small, then those factors can be eliminated from the analysis by “pooling.” After analysis and the potential pooling of insignificant factors, the final stages of analysis can be undertaken. These final stages are optimal level selection and application of cost function.

Identifying which levels are optimal for the controllable factors consists of identifying which combination of main effects for each factor yield maximal SNR. The expected result at this *optimal condition* can be predicted by summing the factor contributions for all factors and adding the sum to the grand average of performance.

### 3.3. Loss Function

Traditional *Statistical Process Control* (SPC) techniques are concerned with minimizing the quantity of products or output from a process that lay outside the specification limits. Cost is incurred when the output of a system is outside the specifications. This cost is usually thought of as an additional cost in running the process and is averaged over the full population produced, yielding a ratio *L* that is loss/unit in dollars. The loss is calculated from the view of the manufacturer or vendor of the product.

Taguchi has taken a more holistic view of cost. He has described the impact of poor quality as being harmful to society as a whole by negatively impacting not only the manufacturer, but also the customer and even the community for the lifetime of the product. Taguchi refers to this as *societal loss*. To quantify this loss he proposed the *loss function*

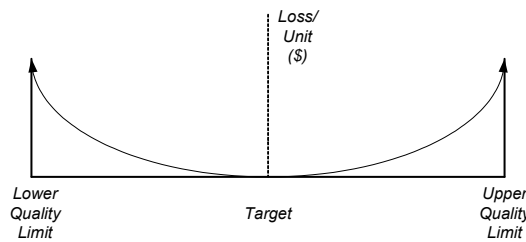


Figure 5. Loss function proposed by Taguchi

For a process that generates multiple outputs or products, the loss function is expressed as  $L=K(MSD)$ , where *K* is a constant that describes the loss per change in mean-squared deviation. The loss function allows an experimenter to quantify in dollars the decrease in loss due to a reduction in standard deviation and or by adherence to target value.

## 4. DOE TEST PLAN

With the problems statement clearly defined, metrics identified and criteria for success understood, the process of creating a test plan based on DOE is straightforward. The challenge to the vendor is to meet all the performance criteria set forth in Table 2, while operating in the environment described by the functional diagram in Figure 2, and maximizing the Watermark Performance Metric.

### 4.1. DOE Factors & Levels

To maximize the WPM, the vendor has identified four factors for their watermark algorithm that it believes impact the metric. These factors and related levels are described in Table 4. Of the four factors, two directly impact the implementation of the algorithm Pre-Filter and Interpolation method; the remaining two (Region of Interest & Bit-Planes) are based on requests by the customer. The first factor (“Region of Interest”) defines how much of the image is to be validated. The fourth factor (“Bit Planes”) is to investigate the possibility of reducing image size by eliminating bit planes after the images have been embedded.

		<i>Levels</i>		
		<i>1</i>	<i>2</i>	<i>3</i>
<i>Factors</i>	Region of Interest	Small – face only	Medium – face & hair	Large – whole image
	Pre-filter type	Type A	Type B	Type C
	Interpolation Method	Interpolation X	Interpolation Y	Interpolation Z
	Bit planes	8	6	4

Table 4. Watermark Algorithm Internal Factor & Associated Levels

With the internal factors defined, next comes the description of the external or “noise” factors. The external factors are readily apparent from the functional diagram in Figure 2. A pictorial representation of the factors/metrics and their relationship to the watermark detector under test is provided in Figure 6.

		<i>Levels</i>	
		<i>1</i>	<i>2</i>
<i>Factors</i>	JPEG Compression	Low	High
	Scale	100	85
	Rotation	0	90

Table 5. External (Noise) Factors as Defined by Figure 2, Functional Diagram

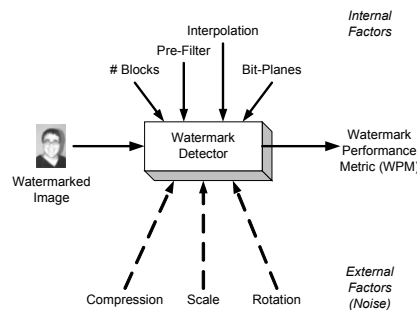


Figure 6. Relationship between factors & watermark detector.

### 4.2. Trial Conditions

The internal factors consist of 4 factors, each with 3 levels, yielding to a full factorial of  $3^4$  or 81 different permutations. A L-9 orthogonal array is appropriate for this experimental situation, resulting in 9 trial conditions. Those trial conditions are:

	<i>Region of Interest</i>	<i>Pre-Filter</i>	<i>Interpolation</i>	<i>Bit-Planes</i>
<i>Trial 1</i>	Small	A	X	8
<i>Trial 2</i>	Small	B	Y	6
<i>Trial 3</i>	Small	C	Z	4
<i>Trial 4</i>	Medium	A	Y	4
<i>Trial 5</i>	Medium	B	Z	8
<i>Trial 6</i>	Medium	C	X	6
<i>Trial 7</i>	Large	A	Z	6
<i>Trial 8</i>	Large	B	X	4
<i>Trial 9</i>	Large	C	Y	8

Table 6. L-9 Internal Factor Trial Conditions

For each of the internal trial conditions, the impact of the external factors needs to be accounted for. The external factors can be mapped into an L-4 orthogonal array, generating 4 different trial conditions.

	<i>Compression</i>	<i>Scale</i>	<i>Rotation</i>
<i>Trial 1</i>	Low	100	0
<i>Trial 2</i>	Low	85	90
<i>Trial 3</i>	High	100	90
<i>Trial 4</i>	High	85	0

Table 7. L-4 External Factor Trial Conditions

For each of the trial conditions from the L-9 array, data is collected as prescribed from the L-4 external factor trial conditions. This yields 36 unique tests that will be undertaken to determine the optimal set of internal factors to increase the S/N (and hence reduce system cost) for the WPM.

### 4.3. Image Sampling

To maximize accuracy, each of the 36 tests could be executed against the full population of employee photographs, 837 images. This would generate over 30,000 test results (WPM) to catalog and analyze, not mention the time to run the test. Instead, an approach of sub-sampling the full population is taken. Ten images are sampled from the population and used for each of the tests, resulting in 360 tests.

The ten images are sampled across the distribution of the Watson metric for the population. This is to insure independence between the Watson metric and the WPM. The customer's requirement that the average Watson metric be  $\leq 0.40$  is also validated as an initial condition for the test. The ten images selected have an average Watson metric of 0.377.

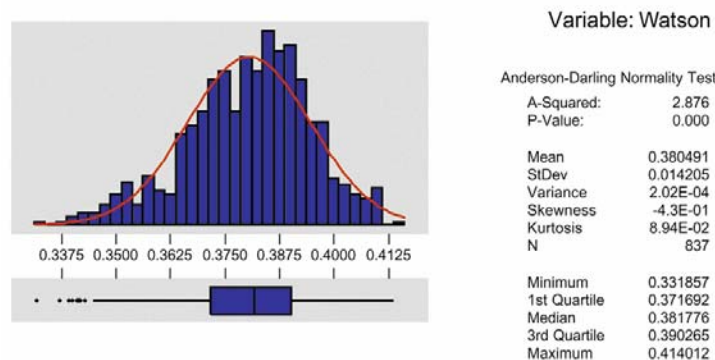


Figure 7. Distribution of Watson Metric Across all Employee Photographs (Embedded with a PSNR=47.09).

### 4.4. Test Execution

The implementation of the watermark detector is modified nine times as prescribed by the inner array. Each implementation is used to detect four groups of images. Each group represents a trial condition from the external array, accounting for the noise in the system (rotation, scale and compression). The same ten embedded images are used to generate each of the 4 groups, producing a total of 40 images for testing.

Utilizing an automated benchmarking system, detector is executed nine times across forty images, yielding 360 Watermark Performance Metrics to analyze. The results were entered into several commercial off-the-shelf statistical data analysis software packages to calculate and plot results.

### 5. DOE RESULTS

The use of an internal array with an external array can provide significant insight into the interplay between factors (internal & external), their contributions to variance (ANOVA), etc. The vendor is concerned with determining the optimal configuration of internal factors and levels that will increase the SNR for the external factors given while reducing cost. To achieve this, the main effects and factor contributions are collected, ANOVA is performed and optimal configuration predicted.

#### 5.1. Inner Array Main Effects

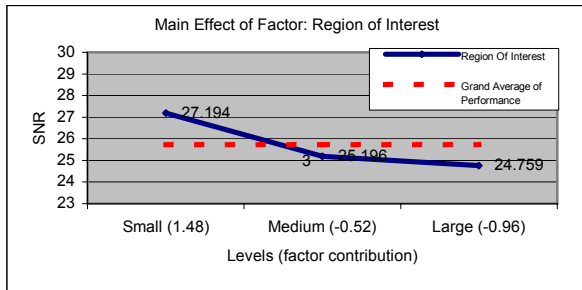


Figure 8. Main Effect: Region of Interest

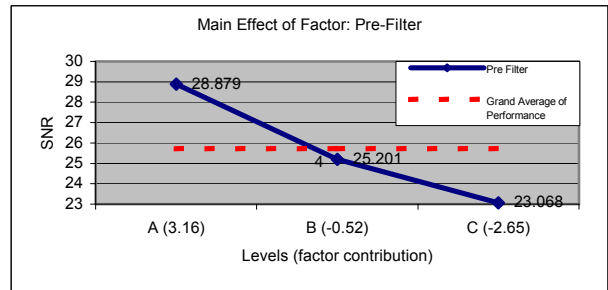


Figure 9. Main Effect: Pre-Filter

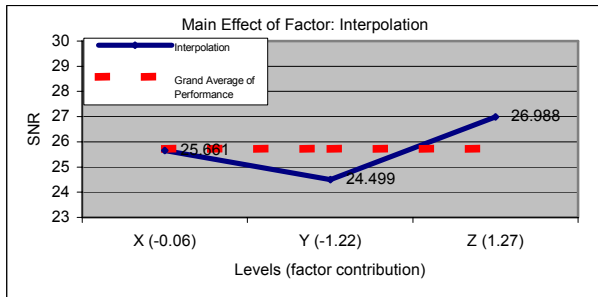


Figure 10. Main Effect: Interpolation

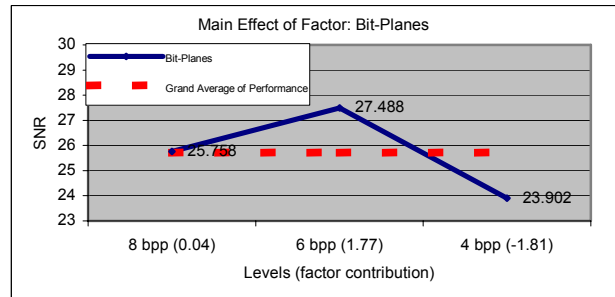


Figure 11. Main Effect: Bit-Planes

Main effects for all four factors are plotted above, with factor contribution provided for each level. Of the results, the trend for Region of Interest and Bit-Planes are the most counterintuitive. Further investigation into Region of Interest, shows that as the region grows, the mean for the WPM actually increases, but so does the standard deviation, which in turn reduces the SNR. The mean for Bit-Planes tracks the SNR, peaking at 6 bits per pixel. The watermark algorithm performs better when the dynamic range of the image has been reduced from 256 to 64 levels after the embedding of the watermark. Pre-filter type A and Interpolation method Z resulted in the best SNR for those factors.

#### 5.2. Outer Array Main Effects

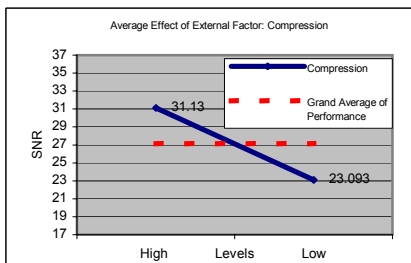


Figure 12. Compression

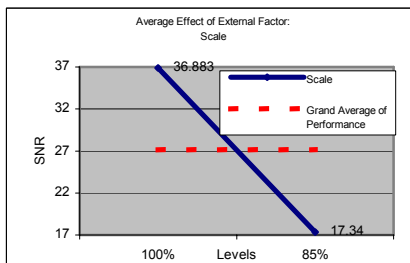


Figure 13. Scale

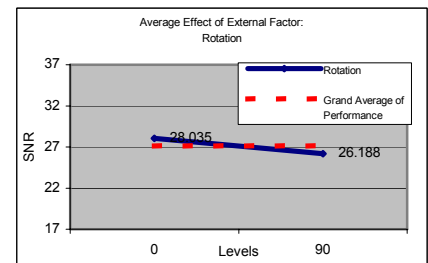


Figure 14. Rotation

Of the three external factors, Compression and Scale have the most dramatic effect on the SNR of WPM (reducing the SNR). The influence of rotation is negligible.

### 5.3. DOE: ANOVA

An analysis of variance will determine which of the internal factors has largest impact on the SNR of the system. The variance introduced by Region of Interest and Interpolation is relatively small compared to remaining factors (refer to Table 8 and Figure 15.)

Factor	Degrees of Freedom	Sum of Squares	Variance	Sum	Percent Variance
Region of Interest	2	10.11	5.06	10.11	11.17%
Pre-Filter	2	51.84	25.92	51.84	57.25%
Interpolation	2	9.31	4.652	9.305	10.28%
Bit-Planes	2	19.291	9.645	19.291	21.30%

Table 8. ANOVA: Internal Factors

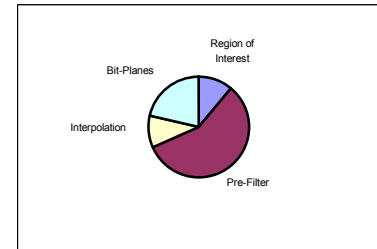


Figure 15. ANOVA

Analysis can continue with all factors to the stage of selecting optimal levels. Before this step is taken, the impact of the ANOVA should be considered as it relates to design decisions and their tradeoffs. Factors can be eliminated from the analysis, or “pooled.” When pooled, the contribution to the variance by the factor to be pooled is combined with the experimental error term. Once pooled, the factor is eliminated from subsequent analysis and is not used in the selection of levels for optimal performance.

Pooling factors simplifies remaining analysis steps and allows for flexibility in the selection of levels for the factors pooled. It is decided to pool Region of Interest due to its small contribution to the variance and the potential performance benefits of using a smaller region of interest. Interpolation has less of an impact on performance, but is also pooled, as its contribution is relatively small.

### 5.4. Optimal Performance

Determining optimal levels for the remaining factors is done by identifying which levels provided maximum factor contribution. For Pre-Filter, the factor contributions were 3.16, -0.52, -2.65 (Figure 9). The contributions for each Bit-Plane were 0.04, 1.77, -1.81, respectively (Figure 11). By inspection, the optimal configuration is Pre-Filter at level 1 (filter A) and Bit-Planes at level 2 (6 bpp). The predicted new performance is determined by adding the contributions to the grand average. The grand average is the mean of the results from all 9 trial conditions presented in Table 9, or 25.72. This yields a predicted SNR of 25.72 + 3.16 + 1.77, or 30.65.

	SNR
<i>Trial 1</i>	30.34
<i>Trial 2</i>	27.23
<i>Trial 3</i>	24.00
<i>Trial 4</i>	25.33
<i>Trial 5</i>	25.99
<i>Trial 6</i>	24.27
<i>Trial 7</i>	30.97
<i>Trial 8</i>	22.37
<i>Trial 9</i>	20.94
<b>Grand Average of Performance</b>	<b>25.72</b>

Table 9. SNR for each L-9 Trial Condition

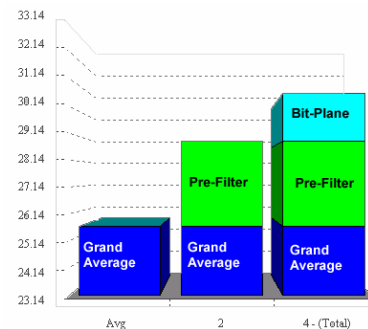


Figure 16. Factor contributions and predicted optimal SNR

### 5.5. Loss Function

Utilizing the loss function, the increase in SNR between the grand average (assumed initial condition) and the optimal levels can be roughly estimated in dollars. The function relies on input from the customer in the form of units of production (for 1 month), cost of rejection and tolerance. Unit of production is calculated by the number of images that pass through the detector in 20 working days. With an average of 800 researchers showing up for work each day and entering the facility on average twice a day, the number of detections is estimated to be 32,000 per month. Additionally the customer and vendor identify a Watermark Performance Metric value of 25.0 as being the threshold between the

detector returning “Authentic” or “Suspicious”. Using these assumptions the follow plot can be generated to show reduction in loss due to the optimal levels (dashed line) compared to the grand average of performance (solid line).

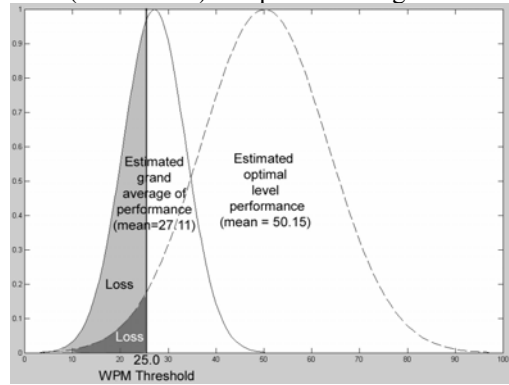


Figure 17. Predicted distribution of performance (grand average, optimal level )

The decrease in loss is the ratio of the area under the curves below the threshold of 25. The loss can also be calculated in dollars using a version of the loss function 8 for a bigger is better QC. The loss ( $L$ ) for a single output ( $y$ ) of the system is calculated with Equation 5 and for multiple outputs ( $\bar{y}$  as mean and  $\sigma^2$  as variance of population) with Equation 6.

$$L = \left( k \frac{1}{y^2} \right) \quad (5)$$

$$L = k \left[ \frac{1}{\bar{y}^2} \right] \left[ 1 + \left( \frac{3\sigma^2}{\bar{y}^2} \right) \right] \quad (6)$$

The constant  $k$  is calculated based on the loss at the specification limits and the acceptable tolerance for the output  $y$ . To calculate  $k$ , one can use Equation 5 to calculate the constant when  $y$  is at the threshold of acceptable tolerance and the loss is known.

The customer has provided the threshold of performance as a Watermark Performance Metric result of 25. The cost to the customer of a valid image that returns a WPM less than 25, is calculated to be \$5 based on the time and expense involved in verifying the image via other methods. The constant  $k$  is calculated by setting  $L=\$5$  and  $y=25$ , and solving for  $k$ , which yields  $k=625$ .

With  $k$  known and variances for the grand average and the optimal performance estimated, a predicted average loss per watermark detection can be calculated using Equation 6. The loss for grand average of performance is calculated to be \$4.24 per watermark detection. The loss when optimal levels are used is calculated to be \$0.53. The estimated savings is calculated as \$3.71, per watermark detection. Over the period of a one-month, or 32,000 watermark detections, the savings is considerable.

## 6. VALIDATION OF RESULTS

To validate the experimental design methodology, the DOE results were compared to the results of the full factorial set of tests with an objective of seeing whether the DOE sub sample (360 test cases, refer to section 4.3) was truly representative of the larger, complete run of tests. A complete test run is defined by  $3^4$ , or 81 experimental watermark reader applications, tested with 837 input test images using 8 combinations of external noise factors (Table 7).

$$81 * 837 * 8 = 542,376 \text{ unique test cases}$$

As a primary indicator of how well the DOE samples compared to reality, SNR values were calculated for all possible experiments (each of the unique test cases). This is effectively sampling the entire population.

For each of the 9 trial conditions recommended by the DOE process, the corresponding SNR across both the full population of 837 images & 8 external factors were calculated against the DOE sub sample, Table 9.

	<b>SNR</b> <b>837 Images</b>	<b>SNR</b> <b>10 images</b>
	<b>All 8 external trial conditions</b>	<b>4 external trial conditions</b>
<b>Trial #1</b>	30.69	30.34
<b>Trial #2</b>	29.31	27.23
<b>Trial #3</b>	24.45	24.00
<b>Trial #4</b>	25.76	25.33
<b>Trial #5</b>	31.13	25.99
<b>Trial #6</b>	24.87	24.27
<b>Trial #7</b>	<b>33.81</b>	<b>30.97</b>
<b>Trial #8</b>	23.54	22.37
<b>Trial #9</b>	22.14	20.94

Table 10. SNR Comparison – DOE Trial Conditions

It can be seen that in both sets of results that the most successful experiment is Trial #7 (in bold), with Trial #1 being a close second.

It was shown earlier that the factors having the largest impact on the WPM were the image Pre-Filter type and the Bit-Plane level. In the case of Trial #7 seen in Table 9, above, the image Pre-Filter used was Type-A with 6 Bit-Plane image data. ANOVA performed in section 5.3 resulted in pooling Region of Interest and Interpolation factors. With these factors pooled, a more interesting validation point is shown by comparing the predicted optimal performance calculated in section 5.4 against the actual performance calculated with the full data set.

To achieve this the SNR for the full factorial data set were calculated by combining results from all 9 experimental variants that use the Type-A Pre-Filter and 6 Bit-Plane image data. This data is compared to the DOE prediction for the Type-A Pre-Filter / 6 Bit Plane experiment, as shown in Table 10, below.

	<b>All Type-A Pre-Filter &amp; Bit-Plane 6 variants (9 variations). Across all 837 images &amp; 8 outer array trial conditions</b>	<b>DOE, Predicted optimal performance (1 variant). Across 10 images &amp; 4 outer array trial conditions</b>
<b>SNR</b>	32.49	30.65

Table 11. SNR Comparison

The question then arises as to whether an exhaustive evaluation of the full factorial results data would recommend a better combination of levels for the remaining Pre-Filter & Bit-Plane factors than the DOE recommendation. SNR calculations for each experimental grouping were completed, and tabulated by image Pre-Filter type and Bit-Plane level.

<b>Pre-Filter Type / Bit Plane Combinations</b>	<b>Avg. SNR</b>	<b>Pre-Filter Type / Bit Plane Combinations</b>	<b>Avg. SNR</b>	<b>Pre-Filter Type / Bit Plane Combinations</b>	<b>Avg. SNR</b>
Type-A / 8 Bit Plane	32.27	Type-B / 8 Bit Plane	30.82	Type-C / 8 Bit Plane	23.82
<b>Type-A / 6 Bit Plane</b>	<b>32.49</b>	Type-B / 6 Bit Plane	31.11	Type-C / 6 Bit Plane	24.60
Type-A / 4 Bit Plane	25.82	Type-B / 4 Bit Plane	23.77	Type-C / 4 Bit Plane	23.12

Table 11. SNR Values – Challenging the Optimal Configuration

The calculations based on the full population of images recommend the same levels for Pre-Filter and Bit-Plane factors (in bold above) as was predicted by the DOE experimental process. The result is larger, 32.49 versus 30.65 for the DOE predicted optimal performance.

## 7. CONCLUSION

The strong correlation between the results from the DOE recommended experiments and the full factorial data from all possible experiments suggests that substantial time and work effort can be saved during the testing process with high confidence that the results will closely match. The impacts for the research community in terms of using industrial process control tools such as Taguchi design of experiments methods to advance watermarking technology to the next level of commercial readiness is evident, and given consistent application of these methodologies in conjunction with standardized benchmarking practices will increase the commercial viability of digital watermarking technology significantly.

The authors recognize that this single test run example is only a small part of the iterative process of invention and transitioning those inventions into a commercially ready product solution. However, using this methodology throughout the development process will provide more concise and predictive indications of the stability, performance capabilities and overall quality of the watermarking technology, while minimizing the generally monumental effort in collection and maintenance of test cases.

- 
- 1 Beizer, Boris, *Black-Box Testing: Techniques for Functional Testing of Software and Systems*, chapter 1, John Wiley and Sons Inc., New York, 1995
  - 2 Martin Kutter and Fabien A.P. Petitcolas. "A fair benchmark for image watermarking systems". Ping Wah Wong and Edward J. Delp, editors, proceedings of *Electronic Imaging, Security and Watermarking of Multimedia Contents*, vol. 3675, pp. 226-239, San Jose, California, USA, January 1999. The Society for Imaging Science and Technology (I.S.&T.) and the International Society for Optical Engineering (SPIE). ISBN 0-8194-3128-1
  - 3 Fabien A. P. Petitcolas, Ross J. Anderson and Markus G. Kuhn. "Attacks on copyright marking systems". David Aucsmith, editor, *Second Workshop on Information Hiding*, in vol. 1525 of *Lecture Notes in Computer Science*, pp. 218-238, Portland, Oregon, USA.
  - 4 Fabien A. P. Petitcolas, Martin Steinbach, Frédéric Raynal, Jana Dittmann, Caroline Fontaine, Nazim Fatés. "A public automated web-based evaluation service for watermarking schemes: StirMark Benchmark". Ping Wah Wong and Edward J. Delp, editors, proceedings of *Security and Watermarking of Multimedia Contents*, vol. 4314, pp. 575-584, San Jose, California, USA, January 2001. The Society for Imaging Science and Technology (I.S.&T.) and the International Society for Optical Engineering (SPIE). ISBN 0-8194-3992-4
  - 5 SACD, INA, PHILIPS, TUD, TCC, NETIMAGE, UVIGO, "Watermarking applications and requirements for benchmarking", [online], (October, 2000), Available at HTTP: [http://vision.unige.ch/certimark/public/CMK\\_D21.pdf](http://vision.unige.ch/certimark/public/CMK_D21.pdf)
  - 6 P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image and Vision Computing J*, Vol. 16, No. 5, pp 295-306, 1998
  - 7 Ranjit K Roy, *Design of Experiments using Taguchi Approach*, chapter 6, John Wiley & Sons Inc., New York, 2001
  - 8 Phillip J. Ross, *Taguchi Techniques for Quality Engineering*, chapter 1, McGraw-Hill Book Company, New York, 1988