Machine assisted visual grading of rare collectibles over the Internet

Richard Bassett
Western Connecticut State University

Overview

Accurately identifying, grading and then determining the authenticity of rare collectible items such as coins, stamps, cards, comic books and artwork is a subjective non-automated process conducted by appraisers or graders. Appraisers and graders are usually human experts in their respective fields that draw on large established pools of domain knowledge, opinions of other experts in their field and make comprehensive comparisons to other 'works' in the field to assist them in arriving at their findings. As with many professions the credentials for a person establishing himself or herself as 'an expert' can range from being non-existent to possessing a long list of impressive industry certifications.

This study examines the human visual recognition process in the grading of rare collectibles and addresses a number of issues, limitations and constraints inherent to human visual recognition as a way of explaining grading variation. The study is then extended by the development of a machine-based system that removes the subjective human characteristics when grading, and is capable of grading rare collectibles by producing consistent and repeatable results. Finally, a machine-assisted human recognition system is developed that combines the technical and subjective attributes of human and machine grading.

The purpose of this research is to determine whether humans, machine systems, or hybrid human-machine systems are better at determining the condition, or grade, of collectible items from digital images over the Internet. The strengths and limitations of both human and machine based grading are studied and explored in a theoretical framework. The capabilities of humans and machine are being examined to explain why previous attempts at computer based grading were unsuccessful and to document why humans grade collectibles in an inconsistent fashion.

Significance of the research

The Internet has provided a major boon to the rare collectibles marketplace as dealers and auction houses are now able to reach vast numbers of collectors, investors and other potential buyers of their offerings. Collectors accumulate everything from rare coins and stamps to baseball cards, autographs, antiques, posters, comic books, beer cans and artwork. Investors seek to purchase rare collectibles that will appreciate over time.

The selling of collectible items isn't limited exclusively to dealers and auction houses as collectors are also able sell their duplicates or extra collectibles items by reaching large numbers of other collectors through the Internet. Collectors with wares to sell can access other collectors through auction sites such as eBay.com and Half.com as well newsgroups, chat rooms, email and their own personal websites.

Ultimately collectors, appraisers, dealers, auction houses, markets and insurance companies are concerned with the value, or worth, of a rare collectible as a basis to determining the proper wholesale and retail pricing.

Original Research

This study utilizes the Internet as a transport tool for improving the grading of collectibles, which has never been done before.

1. The formal study of how humans grades collectibles, why they come up with the grading assessments that they do and the inherent limitations that humans have in the grading process.

2. Why previous commercial attempts at fully automated machine-based grading were unsuccessful in a theoretical context.

3. The method of testing across the Internet was devised to decrease the amount of time required to complete testing, to span the geographic distance between the expert graders, and to maintain experimental consistency. The expert graders that participated in the study were located all over the country and the logistics of getting the expert graders into a single physical location for testing was not practical and also it was resource prohibitive.

Applicability to multiple collectible domains

While coins are typically graded on Sheldon's [16] 1 – 70 scale, comic books are graded on a 1 – 100 point scale [5] and collectible cards, such as baseball cards, are graded on a 1 – 10 point scale [1]. The tool is presently setup for the 1 – 70 range scale but in order to modify the tool for the evaluation of other domains the range scale, lowest to highest, would need to be changed, the input.txt control file which identifies the number of items and their label descriptions would need to be modified, the descriptions of the grades appropriate to the domain being evaluated would have to be loaded and the JPG images would have to be replaced.

The substantial increase in buyers, sellers and trading activity that the Internet has generated has magnified the problem of over-graded collectibles and thus has led to many to purchase items at improperly inflated valuations.

What makes a collectible item rare?

Some collectible items such as raw precious medals, like gold and silver, have only to contend with the variables of supply and demand when determining their values. However, the value determination of non-commodity type rare collectibles is more difficult to calculate as it is dictated by numerous complex factors that include: condition, authenticity, age, number originally produced, estimated surviving population, historical significance as well as market demand.

Condition, which is often referred to as the grade, is significant, as the most discriminating collectors usually prefer their collectible holdings to be in the best condition/grade possible.

Determining the *authenticity* of a potentially rare collectible is a major concern due to the increased financial incentive to produce counterfeit collectibles and the emergence of the technological capabilities to do so. Counterfeiting rare collectibles dates back to the days that coins were first produced. [17]

With some collectibles *age* may be a significant driving factor in determining the value. But with certain collectibles age in itself is insignificant in determining value, for instance a 1995-W One Dollar Silver Eagle made at West Point is considered to be a great rarity with a present day value of just over $2,000 while an uncirculated Silver Dollar made in 1900, which is over 100 years old, is worth just $50. Another example of age having little influence on the value of a coin is the Ancient Greek coin identified as Tiberius 14-37 AD, in very good condition this 1900+ year old coin can be currently obtained for around $100.

The number or amount of a collectible that was *originally produced* is a value-contributing factor as it helps to give understanding to the largest potential size of the population.

In general terms the *surviving population* of a collectible item is thought to diminish with time, this population can also be thought of as the supply. If the documented surviving population exceeds the number production population then it is quite possible that a sufficient number of counterfeits may exist or that improper production record keeping occurred thus contributing to the excessive numbers. [18]

A collectible may be an object of *historical significance* that is tied closely to a time period, a person or a special event in history. Very often collectible items with historical significance may start as very ordinary products.

Market demand contributes to value fluxuations as collectible items fall in and out of favor with collectors over time. Newly discovered hoards can increase the supply in the market and thus drive down the demand. Newly revealed low supply numbers can suddenly cause a collectible to enjoy strong demand. Publicity of an item, the artist or production facility can create sudden demand for items. Unexpectedly high prices realized at auctions for similar items may cause a sudden surge in demand for a collectible. The passing of an artist or designer can create the realization that the supply is suddenly finite and thus drive up demand.

Collectibles that maintain their value and appreciate in value in the long term usually have several or more of these prime factors in common and are often the most sought after items in their respective categories or series. These items are referred to as rare collectibles. Only the minority of items in a particular collectible category is usually deemed rare and have high values attached to them. Thus, the vast majority of collectibles in a category or series are more readily available and easier to obtain at lower costs; these items are usually referred to as being common.

Condition (grade) and authenticity are the two major factors that sellers can fake, with condition (grade) being by in large the most widely abused factor. The value of a collectible item can be substantially more or less than its true value if the condition or grade of a collectible is improperly represented.

The issues associated with human grading

When humans grade items, such as coins, it is assumed that they carefully examine and visually recognize all of the technical and non-technical (subjective) detailed features on the item under consideration before arriving at a grade.

Depending upon the series; the number of technical detailed features to consider on a coin may range from 10 – 20 These 10 - 20 features then need to be translated to an industry-accepted ANA grade, which is on 1 – 70 scale.

The cognitive abilities of the grader also come into play with grading. [3, 9, 10] Grading is a visual pattern recognition process, as it requires humans to identify visualized items and match them against the stored knowledge in their long-term memory. If the graders' long-term memory doesn't contain the building blocks for the recognition of a particular collectible item then they are apt to misidentify or improperly grade it. A person with great skills in the Large Cent series (1793 – 1857) may possess considerable knowledge on the nuances of the series but lack any depth of knowledge in the Lincoln Cent series. When presented with a Lincoln Cent to grade the Large Cent expert would draw on their abilities in grading Large Cents and attempt to apply them to the Lincoln Cent. As so many features are different between the series it is unlikely that the Large Cent expert would be exactly correct in their assessment. However due to humans ability to do complex visual pattern matching they may come close and in extreme cases (such as a coin is a cull or a very nice specimen) their odds of getting the grade correct increase dramatically.

Humans must draw on their long-term stored memory through their cognitive visual pattern recognition abilities while translating the visual input images that they have examined. Extensive documented research in the field of cognitive psychology supports the claim that humans have great difficulty processing more than 5 – 9 chunks of visual input data at one time. Thus by our very own cognitive nature humans loose, drop or fail to consider anywhere from 5 to 15 detailed technical features when grading coins. So in practice the process of a detailed grading examination by humans takes much more cognitive ability than the average grader is able to commit. As a result the grading opinion offered by humans is often incomplete and incorrect.

Exceptions to this process occur in very controlled situations, such as third party grading services and in high profile auctions, when more human resources are allocated, as financial stakes tend to be larger. In these situations multiple graders are often used for consensus grading. The grades of the multiple graders are combined with extended evaluation timeframes to come up with an average grade, which is assigned to the coin.

Humans have been grading coins since long before the 3rd party grading services came into existence. Prior to the adoption of the Sheldon 70 point scale by the ANA in the 1970's dealers

and collectors were assigning grades to coins as the basis of valuation on a less defined scale than Sheldon's [16]. As long as grading has existed so have differences of opinions [1, 8, 12, 13, 17] in the grades being assigned. Stu Miller's grading challenge website [11] is a very good demonstration of the differences in opinions that graders have when grading the same item.

Machine grading issues

Hypothetically, machines can be configured and trained to perform an evaluation of all of the detailed ***technical grading***[1] aspects that humans are supposed to consider in the grading process but rarely do. Such a should be able to consistently produce technical grades at a higher accuracy and consistency level than humans.

There have been several previously documented attempts at automated coin grading and two commercials attempts will be addressed. Two commercial companies, Professional Coin Grading Service (PCGS) [6] and CompuGrade [7] systems in the market in the 1990's but quickly withdrew them due to their lack of commercial success.

In 1990 PCGS announced a computerized system for grading coins. The system, which was known as the PCGS Expert, utilized robotics, image enhancement, image processing and an online image database for its integrated computer system. [6]

CompuGrade, a company run by Jim Diefenthal, had a system, which could consistently grade Morgan dollars to a standard of a tenth of a point. One aspect of the CompuGrade system was that the system could be used as a grading teaching tool as a method for learning grading. [4]

No formalized research could be located which detailed the approach of the CompuGrade system. However coin expert John Baumgart attended a presentation of CompuGrade at the 1991 Chicago ANA show and reported his observations of the CompuGrade system[2]. According to Baumgart, the CompuGrade system reportedly graded coins based on digital images taken in a multistep proprietary technology process within a controlled environment. First a defect map was generated by subtracting an ideal coin from the coin that was being graded and thus producing the defect map. This map contained all defects discovered such as bag marks and scratches, which appeared on the coin. Then an algorithm, which rated the marks contained in the defect map based on location and severity, came up with part of the grade. The next step required that several more images of the coin were taken to ascertain the eye-appeal by evaluating the coins luster. Lastly a human evaluator made sure there weren't any catastrophic errors.

Some of the technical problems reported [2] with the CompuGrade system included:

- Toning could not be measured accurately by the system, as measuring toning in with an automated system is a very complex task. Toning is a natural discoloration of a coin's

[1] ***Technical grading means strict adherence to certain grading rules without the inclusion of subjective qualities. A technical grade is determined solely by wear and defects that occur after striking.***

surface by the atmosphere over a long period of time. Many collectors often consider toning to be very attractive and desirable and they tend to prefer coins with natural toning. Toning is a subjective quality that is an important factor in determining the value of a coin. Toning colors of major concern are white, copper, nickel, and gold depending upon the metal composition of the coin. However these major colors can include: red, red-brown, brown, white, full white, original color, dark color, light tone, pleasing tone, rainbow tone, unusual tone, dark fields and light fields.

- Measuring abnormal die strikes created significant challenges. Die strikes are created during the minting process of impressing the design from a die into a planchet to make a coin. They are important because they indicate the completeness of detail (as in weak strike, full strike, etc.).

- One algorithm for grading didn't work for all coins in the Morgan Series due to frequent design changes in the series.

- The system didn't assure authenticity or detect counterfeits.

- The eye-appeal algorithm didn't provide the entire picture of eye-appeal, which is arguably subjective from person to person.

Avid collector and dealer Byron Reed pointed out that the CompuGrade system was "great from the standpoint of determining detail, but was really poor when determining pretty verses ugly." [14] Still an important benefit afforded by the system was the maintenance of a database, which included the previously graded coins. This database was potentially useful for identifying specific individual coins by contact marks, wear and for the purpose of insurance, theft protection, rarity evaluation, and provenance. [2, 14]

Both PCGS and CompuGrade attempted to build systems that they anticipated would become commercially viable and profitable. They soon discovered that the development of software could be a long and expensive process. In the wake of rising development costs, missed deadlines, ever increasingly complex rule sets, the hope of all profitability diminished and both companies quietly withdrew their systems from the market.

The need for human-machine grading

While not part of the technical grading process there are several subjective non-technical features that dealers, collectors and appraisers often wish to include in the grade of a coin. These subjective features include:

- Color
- Toning
- Defects
- Strike Quality
- Planchet Quality

The subjective features are not usually recognized to be part of the technical grade of a coin by the third party grading companies in the 1 – 70 grading scale. Grading companies seek to provide grades for 'sight unseen' trading and the inclusion of subjective features would contradict this sight unseen trading goal.

To many within the collecting profession these subjective features can heavily influence the value of the collectible. For instance a Morgan Silver Dollar in very good condition with certain toning can command more or less money in the market depending upon the hue of the toning but according to the third party grading services the coin is still technically in very good condition.

This is the point at which professionals disagree as many feel that the technical grade fails to reflect the ***market grade***[2] of the collectible. Collectors and dealers often layer their own opinions to incorporate subjective features onto professionally graded coins. Thus professionals wishing to buy or sell a coin that has been graded by a professional third party company start with the grade assigned by the company and then mark it up or down based on the subjective market features that the grading companies fail to incorporate.

The Experiments

Three experiments have been constructed to demonstrate the strengths and weaknesses in human, machine and machine assisted interactive grading. They are: Human, Machine, and Human/Machine Grading of coins from their digital images.

Producing non-biased consistent grades was an important consideration of these experiments. Tight experimental controls were put in place to reduce evaluation variables, to increase validity and to yield the greatest reliability. The controls planned for these experiment include:

- A minimum level of 3 years grading experience with Lincoln Cents was required for all graders. A person with 3 or more years of grading experience is usually considered an expert.
- The same expert graders were used for the human grading experiment as the detailed feature based grading experiment in order to measure variances in the experiments.
- All graders graded the same 25 coin images.
- The same evaluation procedure for the grading test was utilized for all graders.
- There was no financial incentive for overgrading & undergrading.
- The testing interface allowed the graders to enlarge the images for better viewing.
- No grades were provided to the graders in advance.

[2] ***Market Grade is a grade that takes into account strike, luster, die state, and overall eye appeal of a collectible item and greatly influences what price it will sell for in the market.***

- Ten sample coins were pregraded by independent third party certification services in order to arrive at a baseline grade for measurement in the experiments.

- The methodology of putting the experiment on the Internet overcame the issues of turnaround time and risk to the samples. The browser-based testing tool is capable of working with numerous expert graders simultaneously while preserving the integrity of the samples.

  An earlier approach that was considered, and quickly rejected, was to send the actual coins out to each of the graders. This approach was rejected on the basis of the time requirements and the potential loss/damage risk to the sample coins. By sending the physical coins out to the graders testing could only be done sequentially, one grader at a time, as the coins could only be in one location at any point in time. The time that it would take to perform the test on the grading would be burdened by the extra layers of overhead which included the speed of the US Postal Service (to and from the grader) and the amount of time that the grader maintained the test coins in their possession. If a grader was in California the mailing process alone could potentially add 10 days to the amount of time before another grader could test the coins. The projected extra time from shipping could easily have added 50 days to the study as a goal of the study was to have at least 5 expert graders participate in the experiment. Still a larger risk facing the experiment was the potential loss or damage of the physical samples. In order to ensure consistency for later result measurement and data analysis it was important that all graders graded the same coins and the all coins remained in the same condition throughout the entire experiment. By sending the samples out to the graders there was always the possibility that the samples could be lost in shipping, damaged anywhere along the way or kept by the grader. The probability of loss or damage increased substantially with each additional grader that was added to the test. The loss or alteration of the samples at any point would essentially mean that the experiment would have to stop with the results thus far received or that the entire process would have to be restarted with a new sample set.

Human Grading Experiment

In this experiment 5 – 10 expert human graders will be presented with the scanned obverse (front) images of 25 different sample Lincoln Cents. The scanned images spanned the entire spectrum of possible technical (sometimes referred to as business strike) strike grades in the Sheldon [16] 1 – 70 range. Each human grader will access the scanned images via a web browser across the Internet and record what they think the appropriate grade should be.

The mechanics of this test were designed to measure the template-matching model in human visual recognition [15] while capturing the overall grade assigned by the expert graders of the presented scanned images. The captured grading results for each participant would later be compared with the results of all graders to determine a baseline average grade for each item.

The approach of this experiment will be to utilize a browser-based software tool across the Internet. This tool is capable of presenting the human grader with a scanned image of a coin, a slider bar interface for grade selection and the ability to store and record the selected grade into a database.

Outcome Measurements:

- Number of graders
- Average years of experience of the graders
- Average grade
- Grade variation (range and standard deviation) among the graders for each all 25 coins
- Grade variation (range and standard deviation) among the graders for each all 10 coins that were graded by the certification services

Expected Outcomes:

Based on similar grading experiments done in the past, including the Stujoe Grading Challenge [11] it is expected that there will be considerable variation amongst the graders despite the fact that all graders are considered experienced experts. It is important to conduct this study for to establish baseline data for the next two experiments.

Machine Grading Experiment:

This experiment measures an expert system's ability to perform visual recognition on collectible images by using an automated template-matching model for visual grading. The mechanics of this test were designed to measure the template-matching model in machines recognition capabilities [15] while capturing the overall grade assigned by the expert system of the presented scanned images. The captured grading results for each image are compared with the results of the human graders to measure variances between the average human grade for each item against the machine grade.

In this experiment the machine was presented with the scanned obverse (front) images of 25 different sample Lincoln Cents. The scanned images presented to the expert system represent the spectrum of possible business strike grades in the Sheldon [16] 1 – 70 range. These images were identical to the images that the expert graders examined in the first experiment reviewed.

Outcome Measurements:

- Average years of experience of the graders
- Average grade
- Grade variation from the humans (range and standard deviation) for each all 25 coins
- Grade variation from the humans (range and standard deviation) for each all 10 coins that were graded by the certification services

Expected Outcomes:

It is expected that this experiment will yield consistent and repeatable results while producing grades on a technical grading level. Meaning the machine based system should be able to identify grades on the technical grading merits of the coin. The system will not be able to produce market grades, which take into account the subjective features identified earlier. The accuracy of the machine based system will be measured against the baseline grades of the coins which will be obtained from the 3rd party grading services.

Interactive Human/Machine Grading Experiment

The Human/Machine grading model is a hybrid model that provides the graders with the ability to evaluate all of the critical features on a collectible and the subjective features. This test was designed to combine the attributes of human visual recognition [15] with the capability of the computer to remind the humans to examine the recommended key features of the collectible item to produce a market grade, thus combining the art and science in the grading process.

This detailed grading method is more reliable and slower of the two methods since it must execute more instructions and algorithms to perform more detailed image comparisons. This experiment similar to the first utilizes a browser-based software tool across the Internet. This tool presents the human grader with a scanned image of a coin, a slider bar interface for grade selection, and records the selected grade in a database.

Humans would normally just look at several features, or chunks [10] when grading a coin. By using a graphical user interface (GUI) on the computer to do this we are able to present the graders with all possible features on every coin that should consider. In effect the GUI is minimizing the need for chunking in the visual recognition process by leaving the image up as long as the grader requires.

The tool allows the grader to visually inspect all detailed technical aspects of the coin, thus eliminating the chunking problem and it also gives the grader the ability to include their appropriate subjective interpretations, if they wish, thus arriving at a market grade. This machine-assisted approach to grading should yield more acceptable results as:

- Humans cannot process all of the detailed features required in the same fashion.
- Machines cannot apply the subjective interpretation to the coins that humans can.

Lastly by using the interactive tool designed for this study a grader has the ability to magnify the image into a very large high-resolution image for a detailed inspection. This magnification enhances the grading experience.

Outcome Measurements:

- Average years of experience of the graders
- Average grade

- Grade variation from the humans (range and standard deviation) for each all 25 coins
- Grade variation from the humans (range and standard deviation) for each all 10 coins that were graded by the certification services

Expected Outcomes:

The expectation is that this experiment will yield more accurate grading results than either the Human only or Machine only experiment. The accuracy will be measured relative to the baseline established by experiment #1 and the 10 pregraded coins.

Summary

A large problem with human grading is that it lacks accuracy and consistency. Machines might become good at technical grading but will clearly fail to take into account the subjective features that professionals feel are important. A hybrid human-machine grading process extends the technical grade with the addition of subjective features. These systems are currently not available within the coin-collecting domain. Third party grading companies have yet to see the value in such an approach as has been discussed extensively in the major coin collecting newsgroups such as rec.coin.collecting and PCGS Forums
.
The goals of this research include the study of the feasibility of grading coins over the Internet by viewing images over the Internet and examining the differences between human grading, machine grading and hybrid machine assisted human grading.

If the research indicates that such a hybrid human-machine effort has merit then perhaps it could be the catalyst for change within the third party grading services and for other web-based grading initiatives that involve others in the grading process.

References

[1] AGS, "Advanced Grading Standards for Collectible Cards". 2003, Advanced Grading Specialists, Findlay OH.

[2] Baumgart, J., "Compugrade Demonstration at the 1991 Chicago ANA show", N. rec.collecting.coins, Editor. 2001, rec.collecting.coins.

[3] Biederman, I., "Recognition-by-components: A theory of human image understanding". Psychological Review, 1987: p. 115-147.

[4] Ganz, D., "Computer Grading - Compu-Grade". 1996, rec.collecting.coins.

[5] Gottawiz.com, "Comic Book Grading Standards". 2003.

[6] Halperin, J., "CoinGrading.com". 1999, Dallas, TX: Heritage Capital Corporation.

[7] Hickmott, B., "CompuGrade Input - 5 Places to the right of the decimal", R. Bassett, Editor. 2002.

[8] Locke, M., "Professional Grading Accuracy". 1995, Locke, Mike.

[9] Marr, D., "Vision: a computational investigation into the human representation and processing of visual information". 1982, San Francisco: W.H. Freeman.

[10] Miller, G.A., "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Psychological Review, 1956. **63**: p. 81-97.

[11] Miller, S., "The Grading Challenge Polls". 2002.

[12] Numistrust, "The Coin Grading Process". 2002, Numistrust Corporation.

[13] PCGS, "History of Grading". 2001, Collectors Universe.

[14] Reed, B., "COMPUGRADE technology". 2001, rec.collecting.coins.

[15] Rueckl, J., "Cognitive Psychology Psych 256 Course Notes". 2000, University of Connecticut.

[16] Sheldon, W.H., "Penny Whimsy". 1976: Quarterman Publications, Incorporated.

[17] Travers, S.A., "The Fundamentals of Counterfeit Detection", in *Coin Grading and Counterfeit Detection*. 1997, Random House, Inc.: New York. p. 225.

[18] Wyman, R.J., "A Companion to Rare Coin Collecting". 1997, HeavyPen Productions.