Just Look: Knowing Peers with Image Representation*

Tomasz Kaczmarek† Kuntara Pukthuanthong‡

September 6, 2025

Abstract

We introduce Image-based Firm Similarity (IFS), a novel approach to industry classification that leverages machine learning and visual data. IFS leverages innate human visual processing capabilities to classify firms based on real-time investor perceptions. By analyzing over four million images—including visuals from Google, design patents, and 10-K filings—IFS captures firm similarities more dynamically and intuitively compared to traditional methods. It outperforms established classifications in pair trading, diversification, and industry momentum strategies, particularly in sectors with tangible products and rapid innovation. IFS aligns with investor perceptions, enhancing market applications by providing a more responsive and visually grounded understanding of firm relationships. IFS emphasizes vertical connections over horizontal ones.

Keywords: Images, Firm similarities, Diversification, Industry momentum.
JEL Codes: G00, G11, G12.

*Code and data are available upon request. We thank Fred Bereskin, Gerard Hoberg, Kate Holland, Markku Kaustia, Dhagash Mehta, Mike O'Doherty, Ville Rantala, and seminar participants at Technical University of Munich (TUM) School of Management, 2nd Workshop on Advances in NLP and Generative AI in Finance and Management, Munich, Germany, the University of Missouri Columbia, Missouri State University, University of Missouri St. Louis, the Fields Institute for Research in Mathematical Sciences at the University of Toronto, Poznan University of Economics and Business, and Blackrock. We also appreciate Ville Rantala for generously sharing his updated co-analysts' data.

†Department of Investment and Financial Markets, Institute of Finance, Poznan University of Economics and Business, Poland. tomasz.kaczmarek.sci@gmail.com. Polish National Agency partly funded Tomasz's research for Academic Exchange within the Bekker Programme, grant number BPN/BEK/2021/1/00404/U/DRAFT/00001, and National Science Centre, Poland, grant number 2021/41/N/HS4/02344.

‡Trulaske College of Business at the University of Missouri, Columbia, MO, USA. pukthuanthongk@missouri.edu

1 Introduction

What does an industry truly *look* like? Beyond financial statements and corporate descriptions lies a dimension largely untapped by traditional classification methods-visual representation. We introduce Image-based Firm Similarity (IFS), leveraging the brain's remarkable visual processing capabilities to redefine how we understand firm relationships.

Traditional industry classifications like SIC, NAICS, and GICS are foundational in finance research, but they struggle to keep pace with the rapid evolution and complexity of modern firms. Our study introduces Image-based Firm Similarity (IFS), which leverages the brain's unique capacity for visual processing to group firms by their real-world operations, not just their reported business lines.

As businesses rapidly evolve, classification methods that adapt swiftly to firms' operational focus changes are needed. An innovative clustering approach should offer the flexibility to reflect these rapid shifts and allow companies with diverse activities, such as Tesla, Amazon, and Walmart, to simultaneously belong to multiple industries, better capturing modern enterprises' multifaceted nature. In addition, the classification should capture both tangible and intangible aspects of companies, providing a more comprehensive representation of their business activities and value drivers.

IFS offers a unique approach to sector classification by mirroring investor-defined peer groups. It performs competitively against traditional systems like SIC, GICS, and NAICS, particularly in strategies such as pair trading and diversification. This competitive edge is due to its dynamic reclassification ability and high investor agreement within industries, leading to significant impacts on stock prices. Economically, IFS effectively captures vertical links—those between suppliers and customers—rather than horizontal links within the same industry. IFS effectively clusters firms whose value drivers are based on expected growth and intangibility.

Research shows that the human brain can process images in as little as 13 milliseconds, with 90% of information transmitted to the brain being visual. This results in images being processed 60,000 times faster than text, a phenomenon known as the picture superiority effect (Potter, Wyble, Hagmann, & McCourt, 2014). Our visual processing capabilities are deeply rooted in our evolutionary history. The brain devotes a substantial amount of its energy to visual processing, and the occipital lobe occupies about 20% of its overall

capacity.[1] To put this in perspective, this is a more significant proportion than what is dedicated to other crucial functions like language processing or decision-making. This disproportionate allocation suggests that our brains are naturally optimized for visual information processing, highlighting the evolutionary importance of vision for human survival and interaction.

Studies using Functional Magnetic Resonance Imaging (FMRI) reveal that the parietal cortex processes numeric values, particularly favoring symbolic notation over other representations in the left hemisphere. This neural bias underscores the brain's preference for visual stimuli even within numerical contexts (Cohen Kadosh, Cohen Kadosh, Kaas, Henik, & Goebel, 2007; Piazza & Izard, 2009). Extending beyond numbers, visual data captures attention more effectively and conveys complex narratives, emotions, and operational details with remarkable efficiency—qualities critical for investor decision-making in capital markets (Obaid and Pukthuanthong, 2022). These cognitive insights provide a compelling foundation for our Image-based Firm Similarity (IFS) methodology, which leverages visual representations to redefine industry classifications in a dynamic and intuitive manner.

Our IFS methodology leverages cognitive processes by using various graphical elements to group visually similar objects and define firm similarities, constituting an intransitive approach to define companies' peers. Furthermore, our similarity-based clustering algorithm categorizes firms into 45 and 73 transitive classes.[2] The distinctions between the 45 and 73 classes lie in the granularity of detail they capture, with more classes providing finer distinctions among firms.[3] This dual approach ensures comprehensive and adaptable classifications, accommodating the multifaceted nature of modern enterprises.

The effectiveness of IFS stems from its alignment with the natural tendency of the brain to group visually similar objects, supported by research highlighting the significant impact of visual communication on decision making (Branthwaite, 2002; Dewan, 2015; Shen, Horikawa, Majima, & Kamitani, 2019). This makes IFS particularly relevant in today's image-centric digital landscape, offering investors a more intuitive and relatable framework for stock categorization, especially in financial markets influenced by short-term

[1]The occipital lobe is one of the four major lobes of the cerebral cortex in the brain. It is located at the back of the head and is primarily responsible for visual processing.

[2]Transitivity implies that for any firm, A and B, in the same industry, a firm C in A's industry is also in B's. The transitive approach includes SICs, NAICS, and GICS. The intransitive approach includes the analyst approach. The text-based approach proposes both transitive and intransitive approaches. We explain the differences between transitive and intransitive classifications in detail in Section 2

[3]In the classification with 45 (73) classes, there are 25 (50) industries with at least five firms during the formation period (2009–2013).

fluctuations.

We primarily use Google Images for image collection due to its broad and dynamic coverage of firm-level visuals. Google aggregates market-driven, consumer-facing content that reflects current branding, product focus, and public perception. To assess robustness and expand representation, we supplement the core dataset with images extracted from annual reports and design patent filings.[4] While Google Images excel at identifying product-level similarity, images from 10-K filings and patents provide complementary firm-level visuals—particularly for small-cap and R&D-intensive firms. Incorporating additional image sources, such as patent filings and 10-K visuals, marginally enhances the balance and coverage of the IFS classification by broadening representation across firms. However, this expansion introduces challenges, including increased technical complexity in interpreting abstract visuals (e.g., schematics) and diminished relevance for investors, who may find such images less intuitive or engaging.

We introduce a novel algorithm for capturing entity similarities based on diverse image representations, particularly for companies depicted through varied visuals.[5] Employing Deep Convolutional Neural Networks (VGG19 model), transfer learning, and object recognition techniques, we analyze a large-scale image dataset to establish a statistically significant, visually based similarity measure. IFS utilizes various graphical representations, including product images, supply chain elements, raw materials, and other business-related visuals. This advanced technique is chosen due to the complex nature of our task, where companies are represented by diverse visuals, often including elements not directly related to their core business. Traditional methods struggle with this complexity, but our approach mimics the human brain's ability to process and categorize visual information rapidly and accurately.

Our method incorporates several targeted steps to ensure accuracy and relevance.

- Pre-processing images to filter out common but non-indicative objects.
- Extracting essential visual features that best represent a company's business.
- Applying dimensionality reduction to simplify image data and emphasize meaningful visual patterns.

[4] From 2009 to 2021, we collect approximately 1.63 million clean images from Google (covering 3,458 firms per year), 176,000 images from 10-K annual reports (covering 724–1,197 firms per year), and 181,000 images from design patent filings (covering 201–344 firms per year). The inclusion of these additional sources increases the total image pool by ∼20%, raises the average number of images per firm by ∼15%, and expands the number of firms eligible for IFS classification by 7–8% per period. For details on image processing and sample construction, see Section A.13.3.

[5] This study uses the terms photos, images, photographs, and pictures interchangeably. Our IFS encompasses all these visual representations.

- Defining firms' similarities based on these refined visual features.
- Clustering companies with defined similarities.
- Continuously adjusting for accuracy and validating against traditional classifications.

Our method functions as an efficient and impartial analyzer, processing millions of images to identify crucial visual elements representing the core business of a company while filtering out irrelevant details. For example, when analyzing a soda company like Coca-Cola, it distinguishes between images of soda bottles or cans (primary products) and delivery trucks or vending machines (ancillary visuals). By prioritizing the former for classification purposes, our method ensures that companies are categorized based on their core products rather than peripheral aspects of their operations.

Image processing is resource-intensive. We search for similarities among more than four million images downloaded from Google. Most do not provide valuable information due to their tangential relationship to the company's product offerings. Our methodology for identifying similar firms ensures an accurate visual representation of the products of a company. This process is essential given the challenges posed by the diverse nature of corporate imagery. Without this careful curation, we might mistakenly categorize Walmart in the transportation sector due to images of its delivery trucks or classify numerous NYSE-listed companies as electronics manufacturers due to the prevalence of iPads in corporate imagery.

Our rigorous approach, while demanding, results in a powerful new tool for firm classification that uses visual processing capabilities of the brain. This method provides a more intuitive, efficient, and potentially more accurate way to reveal firm similarities and industry dynamics in today's rapidly evolving business landscape. Although resource-intensive, it offers significant insights that traditional methods may overlook.

Out-of-sample (OOS) testing is a crucial component of our study, addressing concerns of look-ahead bias and data mining in finance. This approach ensures the predictive accuracy and economic relevance of Image Firm Similarities (IFS) in various market conditions and timeframes. We demonstrate the reliability of our image-based classifications by forecasting IFS for future periods using historical image data, such as predicting IFS for 2018–2019 with images from 2015–2017 and for 2020-2021 using images from 2017–2019.

We validate our IFS measure using R^2 values from financial ratio regressions of individual firms against their respective industries. We compare it with established classifications such as SIC, NAICS, GICS, and

the text-based classification of Hoberg and Phillips (2016, henceforth HP). IFS outperforms HP and GICS in most accounting-based ratios, showing the most robust performance in 9 of 16 financial ratios. The IFS achieves the best results in the beta, predicted returns, and PE ratios in market-based ratios.

In particular, IFS excels at capturing expectations and growth, surpassing HP and GICS by 3-5 percentage points in R&D growth and sales growth R^2 values. It also performs well in ratios that capture intangibles such as innovation and human capital. These results suggest that IFS captures both observable and unobservable factors, unlike traditional classification systems that primarily focus on observables.

IFS's performance remains robust even after we exclude small stocks from the analysis. Its use of images, which are more engaging and require less cognitive processing than text or numbers, makes it particularly effective for firms with distinctive products and well-suited for capturing innovation through new product offerings or enhancements. This demonstrates IFS's superior adaptability and responsiveness compared to traditional industry classifications.

Additionally, we explore three applications that leverage image-based similarities: pair trading, diversification, and momentum strategies. For the second and third applications, which require a transitive approach, we utilize the Image Industries 45 classification system. We opt for this classification rather than the more granular Image Industries 73 because it provides a sufficient number of firms within each industry category, ensuring robust analysis for our intended applications.

Our pair trading strategy invests in companies with similar profiles defined through imagery, text (HP), and peer analysts (KR). It ranks firms according to their earnings per share and sales growth, longing high-growth firms, and shorting low-growth ones. Sharpe and Calmar ratios consistently indicate the highest performance of our image-based similarity metric.

To assess diversification benefits, we construct portfolios by randomly selecting one stock from each industry classification for 500 trials. A good industry classification should yield distinct categories, maximizing the diversification benefits of investing across industries. IFS demonstrates superior performance. It produces the top three Sharpe and Calmar ratios across various portfolio constructions: equal-weighted and value-weighted portfolios, maximum Sharpe ratio optimized portfolios, and CVaR ratio optimized portfolios.[6] We also apply the industry momentum proposed by Moskowitz and Grinblatt (1999), short-term

[6]The closest competitor is six-digit GICs, which perform well for equal- and value-weighted portfolios but lag in optimized portfolios. These results validate that IFS is distinctive and offers potential diversification advantages.

reversal, and the volatility-adjusted momentum proposed by Barroso and Santa-Clara (2015). IFS delivers the highest Sharpe ratios for industries with equal and value-weighted industries.[7]

The effectiveness of IFS can be attributed to the fast dynamics and high investor agreement within industry categorizations. IFS shows the highest frequency of business reclassification between industries, with 17% to 22% of IFS-classified companies switching industries annually. IFS exhibits the fastest dynamics compared to other industry classification approaches, with over 200% faster than HP and NAICS and more than 400% that of other classifications.[8] This agility allows IFS to rapidly adapt to changes in companies' activities, strategies, and market positions through visual cues.

Examples of IFS agility include earlier reclassifications of companies such as Exxon Mobil Corp., Hornbeck Offshore Services Inc., FMC Corp, Oshkosh Corp., and SEACOR Holdings Inc. compared to traditional classification systems.[9] This responsiveness makes IFS particularly valuable for real-time investment decisions.

More investor agreement within an industry leads to more significant influences of aggregated demand and supply on stock prices, making that industry categorization more advantageous for investment applications. We demonstrate that IFS provides the highest agreement within industries, explaining its benefits in pair trading, diversification, and momentum strategies.

We show that our image industry classification presents vertical links more than horizontal links. The joint analysis of horizontal and vertical integration demonstrates that IFS industries map firm similarity by taxonomy and role in the economic system. Whereas HP classifications reflect revenue similarities, IFS captures firms' economic positioning, which is reflected in strong supply chain relationships even without

[7]Strategies built on IFS demonstrate superior robustness to changes in strategy parameters, including modifications to holding periods, company weighting within sectors, and using momentum or short-term reversal strategies. We also generate pseudo-random portfolios to show that the industry momentum effect associated with IFS is directly due to industry momentum rather than individual company momentum.

[8]This enhanced agility can be attributed to three key factors. First, IFS rapidly adapts to changes in companies' activities and market positions by analyzing real-time visual cues such as product offerings and branding. Second, while photos are downloaded annually and the model updated biennially, visual data still reflects a company's current state more immediately than delayed textual descriptions or financial reports. For instance, photos uploaded to Google are available almost instantly, whereas annual reports are published with significant lags. Finally, the human brain's ability to process images faster than text enables IFS to swiftly categorize industry dynamics based on visual similarities, offering a more responsive classification system than traditional methods like SIC or GICS.

[9]For example, SIC (NAICS) reclassified Exxon Mobil Corp from petroleum & coal products to oil & gas extraction in 2017. Our images demonstrate the same reclassification but two years earlier, based on photos from 2013 to 2015. SIC reclassified Hornbeck Offshore Services Inc. from water transportation to oil & gas extraction in 2018, while image-based similarity demonstrated a comparable change in peer structure from 2014 to 2016. Other examples of image agility in reclassification include, e.g., FMC Corp. reclassified from industrial machinery & equipment to chemical & allied products, Oshkosh Corp. from transportation equipment to industrial machinery & equipment, or SEACOR Holdings Inc. from water transportation to oil & gas extraction.

classification overlap.[10]

While our project aims to develop a novel industry categorization metric, we acknowledge certain limitations that highlight its specific scope rather than detract from its utility. First, IFS may be less effective for firms not easily represented by images, such as consulting firms or those dealing with abstract technologies. However, this limitation is mitigated by including images from patents and 10-K filings, which expand coverage to R&D-intensive and service-oriented sectors. As detailed in Section 5 (see Tables D4 and D5), incorporating these additional image sources does not significantly alter the results, confirming the robustness of IFS across diverse firm types. Second, while IFS cannot achieve the extreme granularity of traditional systems like SIC codes (which offer 300–400 clusters at finer levels), it provides meaningful classifications up to 100 clusters. Beyond this threshold, the sample size becomes too small to support reliable similarity measures. Nevertheless, IFS excels at capturing broader economic linkages and investor-defined peer groups, offering unique insights into firm relationships that complement traditional systems.

Third, IFS relies on recent visual data due to its dependence on image availability, limiting its applicability for historical analyses and long-term trend studies. However, this focus on contemporary data enhances its ability to reflect current market dynamics and firm operations, making it particularly valuable for real-time applications. Despite these limitations, we add to the industry classification literature by introducing a dynamic and visually grounded approach that aligns with investor perceptions. By leveraging Google Images alongside patent and 10-K visuals, IFS captures firm similarities in ways that traditional text-based classifications cannot. This methodology provides actionable insights for investment strategies such as pair trading, diversification, and momentum analysis.

Our contributions are fourfold: we add to the industry classification literature and join recent efforts in utilizing machine learning and images for investment applications. Obaid and Pukthuanthong (2022) extract sentiment from news images and show it surpasses text in predicting stock returns. Jiang, Kelly, and Xiu (2023) employ graphs to expect returns from trend strategies. We contribute to neuroscience literature by demonstrating the brain's superior ability to cluster photos compared to text and numbers and participate in behavioral economics research by showing how photos might cause overreaction, a critical explanation for

[10]IFS captures vertical links more effectively than horizontal ones, as demonstrated by Ameren Corp (a power company) and Kosmos Energy Ltd (an energy producer), both classified under Industry 45 due to shared visual features like power plants and energy infrastructure. Similarly, Exxon Mobil Corp (a petroleum producer) and FMC Corp (a chemical manufacturer) are clustered, reflecting their supply chain linkages where Exxon supplies raw materials for FMC's industrial processes. See Section 6 for our detailed discussion.

well-known momentum trading strategies (see Daniel, Hirshleifer, and Subrahmanyam, 1998).

2 Comparative analysis of industry classification metrics

Industry categorization has been a long-standing focus in financial economics research and can be broadly divided into transitive and intransitive approaches.

Transitive approaches, including the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), and the Global Industry Classification Standard (GICS), are based on business activities. These systems provide a hierarchical structure for categorizing companies in various sectors and industries. Intransitive approaches, on the other hand, focus on firm similarities based on various characteristics. These include text-based measures from 10-K filings (HP), firm characteristics (He, Wang, and Yu, 2021), analyst reports (KR), technological similarities (Lee, Sun, Wang, and Zhang, 2019), and Edgar search patterns (Lee, Ma, and Wang, 2015).

Our Image Firm Similarities (IFS) technique and the Hoberg and Phillips (HP) approach offer both transitive and intransitive methodologies. Previous research compares the informativeness of different classification systems. Kahle and Walkling (1996) examine SIC codes from the Center for Research in Security Prices (CRSP) and Compustat databases, while Fama and French (1997) create new industry classifications by grouping existing four-digit SIC codes. Krishnan and Press (2003) compare SIC codes to NAICS codes and Bhojraj, Lee, and Oler (2003) evaluate various classifications of the fixed industry. While these studies provide valuable insights into using existing static classifications, they do not explore fundamental improvements to the underlying methodology of industry categorization.

While convenient for research purposes, traditional industry categories like SIC or NAICS have at least two significant limitations. Firstly, they lack dynamism in reclassifying enterprises as product markets evolve. In contrast, IFS is inherently more dynamic, basing its classifications on product images. Anyone can upload photos to Google regularly throughout the year, and these photos are available immediately. In contrast, the text and numbers from annual reports are reported with lags and are not available in real-time. Although our methodology involves downloading image data once a year and updating the machine learning model biennially, the visual data still provides a more immediate reflection of a company's current state

than textual descriptions or financial reports, which are updated less frequently. Secondly, human brains can process images faster, whereas traditional classifications require time for human processing, digestion, and analysis. Thirdly, photographs provide a more direct and agile representation of firms' products and services, capturing advancements in offerings more quickly. For example, at the turn of the millennium, hundreds of new technology and web-based companies were broadly classified as "business services" under the SIC system, highlighting its inability to keep up with rapid technological changes. Moreover, while traditional approaches offer qualitative classifications, IFS provides quantitative outputs, presenting scores or rankings indicating the similarity between firms within an industry. This quantitative aspect improves the precision and applicability of the classification system.

Text- and analyst-based groupings offer advantages over SIC, NAICS, and GICS, primarily due to their intransitive nature and annual revision cycles. The intransitive approaches excel in areas like pair trading and comparable firm valuation. However, the transitive nature of SIC, NAICS, and GICS makes them more suitable for applications such as industry momentum analysis and portfolio optimization strategies.[11]

Text-based groups are pioneered by Lewellen (2012) and Rauh and Sufi (2012), and are formalized and extended by Hoberg and Phillips (2016) (HP). Ibriyamova, Kogan, Salganik-Shoshan, and Stolin (2019) extend the application to company brief descriptions. In this study, we compare our IFS with the text-based classifications by HP, as they are the only group that provides the data.

Turning to the analyst-based group proposed by Ramnath (2002) and Kaustia and Rantala (2021), we focus on the latter. The industry group is not a focus of Ramnath (2002). Kaustia and Rantala (2021) apply the methodology of Kaustia and Rantala (2015) to identify firms with common analysts to construct peer groups. Their classification is intransitive. Therefore, we only include their peer groups as our benchmark for the pair-trading application in this study.

2.1 Image-based Firm Similarity vs. Other Intransitive Approaches

Each has unique strengths and limitations that highlight their differences.

Data and Dimensionality: IFS leverages visual data, analyzing approximately four million images with an

[11] HP provides two types of classifications: transitive and non-transitive. Classifications that meet the transitivity condition are universal and widely applicable. For example, when diversifying a portfolio between industries, the classification must be transitive, and industries must be unrelated to maximize diversification benefits.

average of 75 images after cleaning per firm in each 3-year rolling period (see Table 2). Each image is a three-dimensional matrix (224x224x3).[12] This captures detailed visual information about a company's products and operations. This rich data set allows IFS to provide a comprehensive and nuanced representation of a company's offerings, which is particularly effective for industries with tangible products. In contrast, HP relies on textual data from 10-K filings, with a word count of a minimum of 1000 words, treating each word as a binary variable (0 or 1) to determine similarity. Other intransitive approaches use various data sources: analyst-based groupings rely on common analyst coverage, characteristic-based clustering uses firm-specific attributes like financial ratios, technology-based clustering focuses on technological similarities, and Edgar search-based classifications utilize search patterns from regulatory filings.

Dynamic Adaptability: IFS demonstrates superior dynamic adaptability. Images are uploaded frequently throughout the year, allowing IFS to quickly capture market trends and product innovations through visual cues, making it more responsive to real-time changes than text-based and numerical data methods. This dynamic nature is particularly beneficial in rapidly evolving industries. Other intransitive approaches, such as text-based, analyst-based, and characteristic-based classifications, may lag in adapting to market changes due to the need for periodic updates and reliance on textual or numerical data.

Furthermore, IFS relies on recent visual data, offering a dynamic and real-time reflection of firms' activities. Its applicability for historical analyses requires the series of images with timestamps and that is what we do in this paper. In comparison, traditional systems like SIC and NAICS excel in long-term trend studies but often lack the responsiveness needed to capture rapid changes in modern business environments

Limitations: Despite its strengths, IFS has certain limitations. It is most effective for businesses with tangible products or visually distinctive operations, making it more challenging for abstract or service-oriented firms such as consulting or high-tech companies. To address this, we expand the dataset by incorporating images from design patents and 10-K filings, which enhance coverage for R&D-intensive and small-cap firms while broadening IFS's applicability beyond consumer-facing companies. While IFS defines peers for approximately 1,100 firms per year—a sample size that balances classification accuracy with computational feasibility—the inclusion of these additional sources has improved granularity and representation. However, achieving more than 100 clusters remains challenging due to inherent data constraints. For details on the

[12]The first 224 is the height of the image in pixels. The second 224 represents the image's width in pixels, and 3 represents the number of color channels (typically representing Red, Green, and Blue.)

expanded dataset and its impact on classification performance, see Section **??**.

Investor Agreement and Adaptability: IFS shows high investor agreement within industry categorizations, significantly influencing stock prices. This makes it particularly advantageous for investment applications like pair trading, diversification, and momentum strategies. IFS's visual-based categorization aligns closely with how humans naturally process and group information through visual cues, effectively capturing intuitive cognitive responses. By leveraging the brain's innate preference for visual stimuli over text or numbers, IFS mirrors how investors perceive firm relationships as shown in Section 9.

While higher granularity in traditional classification systems like SIC or GICS provides enhanced precision for peer grouping and tailored investment strategies, it often comes at the cost of reduced adaptability to dynamic changes in firm operations. In contrast, IFS strikes a balance by offering sufficient granularity while maintaining flexibility through its dynamic reclassification capabilities.

3 Methodologies and Data

Interpretation of Economic Links

- High vertical linkages reveal strong supply chain dependencies.
- Some IFS industries show weak horizontal but strong vertical integration

  ⇒ *Better reflects functional economic roles. IFS identifies hidden linkages missed by traditional classifications.*

3.1 Conclusion

Our image-based industry classification methodology offers several advantages over traditional approaches. By leveraging visual data, we capture both tangible product characteristics and intangible branding elements. The approach is dynamic by design, automatically adapting to changes in firm operations without manual reclassification. This methodology provides a complementary perspective to text-based and financial classification systems, potentially revealing economic relationships that might otherwise remain undetected.

For a comprehensive description of each step-including data sources, preprocessing procedures, technical

specifications, and additional analyses-please refer to Appendix 10. We show the procedure for how we handle out-of-sample tests below.

4 Results

This section describes the relatedness of firms clustered with photos that illustrate their business activity. We start by reporting the characteristics of firms' similarities captured with images and industries formed with those similarities. Then, we examine the usefulness and limitations of IFS in explaining cross-sectional variations in firm-level stock returns, market-based valuation multiplies, and financial ratios by comparing them to other industry classification methods; this includes SIC, NAICS, GICS, Fama French, as well as Hoberg and Phillips (2016) industries.

4.1 Firm relatedness visualized by image

In exploring firm-relatedness through visual data, our approach harnesses the power of an image to form firm similarities. This methodology effectively captures the essence of companies' business activities, as illustrated by the visual representations of peers from notable large-cap companies: Exxon Mobil Corp, Walmart Inc., Citigroup Inc., and Johnson & Johnson. These examples highlight the diverse nature of objects that can underscore similarities within industries and are visualized in Figures 3, 4, 5, and 6.

Exxon Mobil Corp's range of images includes refineries, petroleum stations, drilling machines, and tankers, reflecting the broad spectrum of its operations. Walmart Inc. is visually represented through internal and external store views, grocery counters, consumer products, and delivery trucks bearing its logo, showcasing its retail dominance. Citigroup Inc. presents a more abstract connection, where images of glass skyscrapers, prominent building logotypes, and scenes of individuals engaging in financial activities, such as discussing in TV studios or using banking apps, depict the financial services industry's essence. Johnson & Johnson's image similarities are more straightforward, with visuals of cosmetics and standardized packaged chemicals highlighting its product offerings. This nuanced identification of similarities reveals how diverse objects and scenes can bind companies within the same industry, showcasing the sophistication of our image-based analysis. In Section 5, we include images from patents and 10-K filings, as these may provide additional

context that differentiates firms with overlapping visual elements. For example, technical schematics or operational visuals could clarify distinctions between financial institutions and energy companies.

However, it should be noted that, despite the overall success of this method, some instances of misclassification occur, such as the association of US Bancorp with Exxon Mobil Corp demonstrated in the last column of images in Figure 3. This is attributed to the visual overlap in the presentation of logotypes and the architectural resemblance between refineries and skyscrapers. Such examples underscore the challenges and complexities of defining industries solely based on visual data.

Building on the foundation of visual similarities, the next phase of our analysis delves into the visual features of Image Industries. We create the Image Industries with a large set of photos. Table 2 shows that the final average number of photos used to link firms to industries in a single period is more than 200 thousand. The typical industry is represented by at least several hundred images per period. The detailed human-made evaluation of all photos per industry is undoubtedly beyond the scope of this research.[13]

Figures 7 and 8 illustrate random representations of six sectors of Image Industries 73, characterized by high inter-industry correlation. The visual similarity of objects describing randomly selected companies between different industries is apparent. For example, many photos representing Industry 40 show electronic components and circuit boards. Similarly, Industry 15 displays images of packaging, bottles, and what appear to be chemical or cosmetic products. In contrast, Industry 1 consistently shows various vehicles, demonstrating the first characteristic of Image Industries: selective similarity aggregation. This approach groups visually similar objects across different industries rather than narrowly clustering identical objects within a single industry. For example, firms producing cars, trucks, and motorcycles may be grouped together based on shared visual features, capturing broader economic linkages. While this method excels in industries with tangible and visually distinctive products (e.g., manufacturing), earlier R^2 results indicate that IFS is also effective for ratios related to intangibility and growth—highlighting its complementary strengths across diverse dimensions of firm similarity.

In contrast, Figures 9 and 10 highlight industries with lower inter-industry correlations, demonstrating greater object diversity within each industry. Each of the six industries depicted in these figures contains

[13] A process where human experts manually review and assess a subset of images to identify key characteristics and ensure the accuracy of the automated classification. Nevertheless, we wish to indicate several observable characteristics even after looking at only dozens of images.

various objects, some of which may not seem intuitively related to human observers. This suggests that image industries may capture sophisticated object relations that are not immediately obvious. While this feature could be advantageous for Image Industries by identifying hidden common visual representations across different sectors, it may also pose a challenge when these visual similarities are unrelated to companies' actual product offerings or core business activities.[14]

Next, we shift our attention from the micro level to the bigger picture. Table 3 demonstrates the characteristics of Image Industries. First, our Image Industries classify 50.8% to 53.3% of our stock sample, translating into a monthly average of 1,084 to 1,110 firms. The photo representation of unclassified firms is of nondistinguishable quality to indicate which class it should be assigned to.[15] Second, many industries contain a small number of companies. This phenomenon is attributed to the fact that the similarity matrix used to construct industries is very sparse. Some companies have only one peer or at most a few. In clustering, such companies remain their sole peers, forming industries with few members. Therefore, our study employs two terms to describe the number of industries in a given classification. The first pertains to the total number of industries, including those composed of a single company. The second encompasses the number of industries with a minimum of five companies at the time of the first industry definition.[16] Consequently, a classification that has 25 (50) industries comprising at least five companies totals 45 (73) clusters. Using industries with a minimum of five companies is especially crucial for economic tests and proposed applications.

Table 4 presents the number of industries and the number of companies for the classifications used as benchmarks in our tests.

4.2 Economic homogeneity of firms clustered with image

To verify economic relations between companies classified into a single industry, we imply the methodology proposed by Bhojraj et al. (2003). We create equal-weighted industry portfolios and verify the ability of

[14]IFS might captures: 1. Hidden Visual Relationships: IFS identifies subtle visual connections that may not be apparent through traditional classifications. For example, firms grouped together may share underlying visual elements in their branding, packaging, or operational imagery that signify deeper economic or strategic links. 2. Broader Economic Linkages: By capturing these hidden relationships, IFS emphasizes broader economic positioning rather than narrowly defined product or revenue similarities. This aligns with its strength in identifying vertical links (e.g., supply chain relationships) and firms' roles within the economic system. 3. Sophistication Beyond Intuition: The diversity within industries depicted in Figures 8 and 9 indicates that IFS goes beyond intuitive clustering, uncovering connections that might be missed by human judgment or text-based methods.

[15]Image Industries classifies a similar number of stocks as the typical analyst approach introduced by Kaustia and Rantala (2021), who categorize an average of 1,075 stocks from 1983 to 2013.

[16]Table C2 demonstrates descriptive statistics of image industries defined in 2013. Industries marked in bold have at least five stocks.

these portfolios to explain contemporaneous firm-level indicators ($vble_t$). We estimate the average adjusted R^2 for all firms within each cluster. R^2 is calculated from a regression below for each firm i:

$$\text{v}ble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t, \tag{1}$$

where the dependent variable $vble$ represents ten ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t, and 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIV PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSETS; 14) RNOA; 15) ROE; and 16) ASSET to SALES, for each firm i at quarter t, the independent variable $vble_{ind,t}$ is the average of this variable for all firms in cluster ind excluding firm i at month t. To eliminate the possible effect of outliers, we winsorize at 1% for all ratios except MONTHLY RET and the binary variable DIV PAYMENT. Also, we perform only one-sided winsorization on SCALED R&D EXPENSE for the highest values because many firms report zero.

We use the same set of ratios as Kaustia and Rantala (2021) extended by FORECASTED MONTHLY RET, TOBIN'S Q, R&D GROWTH, SG&A to # EMPLOYEES, SG&A GROWTH, FORECASTED EPS, and EPS GROWTH. Table C1 in the Appendix shows details of the ratios calculation. Based on market information, we calculate $vble_{ind,t}$ monthly frequency for ten indicators. We estimate the quarterly industry averages for 16 ratios that use only the accountancy information, consistent with the reporting frequency. We calculate regression for industries that have at least five members. The average adjusted R^2 value shows the explanatory power of the industry average on the used ratio of the firm.

For comparability, Tables 5 and 6 present pairwise results comparisons derived from the *same* set of firms classified using Image Industries and other classification techniques. Table 5 compares Image Industries with market information-based variables. The average rank for the two image-based classifications ranges from first to last. Image classifications show extreme homogeneity when evaluating three variables: forecasted monthly returns, market leverage, and price-to-earnings ratio. Overall, the economic homogeneity based on

market-based ratios is average and comparable to SIC-based industries, including Fama-French Industries. This result is based on a sample of firms with data available for all classifications.

When considering all firms (see Table C3), the results for market-based indicators show more favorable overall rankings for Image Industries. Image-based classification performs comparably to GICS industries. The most significant improvements are observed in the market-to-book ratio, monthly returns, market leverage, beta, and enterprise value-to-sales ratio, with no declines in performance. Microcap stocks typically have lower liquidity. When we exclude them from the analysis (see Table C5), the ranking of Image Industries remains consistent.

We compare ratios based solely on accounting information, as reported in Table 6. The average rank of image-based classification varies widely, achieving the highest homogeneity for net sales, sales growth, R&D expense to sales, R&D growth, SG&A growth, EPS growth, dividend payout, profit margin, and debt to equity. However, it performs poorly in terms of asset-to-sales ratio. Comparing image and text-based classifications shows relative outperformance of similar classifications in 60% of observations. Image Industries outperforms GICS in 63% of cases.

Considering all firms (see Table C4), Image Industries is the most robust classification, ranking first or second in 75% of observations. Notably, excluding microcaps does not significantly alter Image Industries' accounting-based ratio ranking (see Table C6). This consistency underscores the robustness of our image-based classification across different firm sizes.

Overall, the homogeneity of firms clustered with images depends on the character of the data used to create ratios. It reaches the top-performing GICS classification for accountancy-based ratios and is equal to SIC-based industries for market-based ratios. Furthermore, the high variation in R^2 rankings suggests that image-based industries differ significantly from competitors. The following example illustrates the advantage of IFS as a classification based on product similarities. Ameren Corp is classified with IFS 73 to industry 45, with an exceptionally high inter-industry correlation of 0.335. Such a high correlation level means that companies classified in industry 45 are characterized by intense homogeneity. Ameren Corp is an American power company.

The IFS classified this company into industry 45 through photos of electric power plants. Close-up images also depict products of, e.g., Kosmos Energy Ltd, Mexco Energy Corp, or Noble Energy Inc—so

these firms are peers of Ameren. Instead, a two-digit SIC classifies Ameren Corp to electric, gas, & sanitary services, three-digit NAICS to utilities, and six-digit GICS to multi-utilities. Because of this—in the case of classifications not based on product similarities— Ameren Corp is not a peer of Kosmos Energy Ltd, Mexco Energy Corp, or Noble Energy Inc.

In contrast, HP classifications that are also product-related agree with IFS about the Ameren Corp peers. This example shows that product-oriented classifications categorize companies that need power plants to deliver electricity to the same industry as electricity producers. Meanwhile, other classifications break such companies down by different industries, where companies' performances are linked to a demand for different, unrelated products. In the next Section, we examine the proposed image-based classification's economic insights and sample applications.

5 Expanded Image Sets: Including images from 10-K and patents

The main findings of this paper rely heavily on Google Images, which predominantly feature consumer-facing firms with tangible products, such as those in retail and manufacturing. However, this approach has limitations when applied to service-oriented or technology-driven firms. To address these gaps and improve robustness, the analysis incorporates patent images and annual reports visuals to broaden coverage. Patent images are particularly valuable for capturing firms involved in abstract technologies or services, such as software and biotech. While Google Images excels in market-driven classification, real-time alignment, and investor consensus, patent images enhance technical sector representation. However, they introduce classification challenges due to their abstract and schematic nature, which often lacks consumer-facing context.

Similarly, images from annual reports focus on operational visuals—such as manufacturing plants and product lines — offering a broader firm-level perspective. These visuals help reduce reliance on market-driven Google Images but do not capture compliance-related or financial risk data. Additionally, the static nature of 10-K filings limits their ability to reflect rapid changes in firm operations or market conditions. Overall, Google Images provides granular product-level classifications but is less effective for abstract or service-oriented firms. Patent images improve coverage for R&D-intensive sectors but reduce precision due to their technical focus. Meanwhile, visuals from annual reports contribute to broader firm-level representation,

particularly for small-cap and innovation-driven firms that rely on abstract concepts often excluded from the initial sample.

By integrating these diverse image sources, the study achieves a more comprehensive representation of firms across various industries. For further details on sample selection, industry descriptions, and $\mathbf{R^2}$ results, refer to Section **??**.

Under the aggregated sample, we generate two classifications: IFS34(25), which includes 34 industries or 25 industries with at least five firms per class, and IFS65(50), encompassing 65 industries or 50 industries with at least five firms per class.[17] The results in Tables D4 and D5 show that under Image Industries 34 (25), the accuracy of *total assets*, *EPS growth*, and *RNOA* increases, while for Image Industries 65 (50), the R^2 is higher for *profit margin*. However, the R^2 of *debt-to-equity*, *forecasted EPS* (IFS 34 (25)), and *total assets*, *debt-to-equity*, and *ROE* (IFS 65 (50)) decline.

On the other hand, for IFS34(25), the R^2 of market-based ratios improves with the expanded sample. The R^2 of *market leverage* and *beta* is highest for the IFS classification. For Image Industries 65(50), the R^2 of *PE* becomes the second best.

We exclude patent and annual report images from the main results for several reasons. First, while patent images expand coverage to R&D-intensive sectors, they often lack consumer-facing context and primarily depict products or processes that are less relevant for clustering firms based on market-driven classifications. Patent images are more like technical schematics and dilute consumer-centric similarity signals, reducing homogeneity in financial metrics. While patent images improve technical sector coverage, they introduce too much noise compared to the market-driven, consumer-focused Google Images that form the core of the successful IFS classification system. Second, although 10-K images provide standardized operational visuals such as facilities and product lines, they fail to offer dynamic updates that reflect real-time market trends. The static nature of 10-K filings limits their ability to capture rapid changes in firms' operations or offerings. Additionally, images in annual reports are sparse and often consist of graphs, charts, or tables rather than visuals directly related to a firm's products or services. This lack of comprehensive visual representation reduces their utility for enhancing classification precision.

[17] It is noteworthy that without the images from patents and annual reports, we have IFS 45(25) and IFS 73(50). IFS 45(25) presents 45 classifications or 25 classifications in which each of the classifications has at least five firms. IFS 73(50) presents 73 classifications or 50 classifications, in which each of classification has at least five firms.

Consequently, including patent and 10-K images could compromise the detail and predictive accuracy of the primary classification outcomes. By focusing on Google Images, which provide dynamic, consumer-focused visuals, the IFS system achieves higher precision and relevance for market-driven applications.

6 Economic Links

Understanding how firms are clustered and interconnected is essential for analyzing the complex structure of economic interdependence. This entails recognizing *vertical linkages*, which connect suppliers and customers across different stages of production, and *horizontal relationships*, which arise from shared industry classifications. Our Image-based Firm Similarity (IFS) framework integrates these dimensions into a unified analysis, offering a more nuanced perspective on firm clustering.

The HP classification groups firms based on their business revenue streams, as reflected in Hoberg and Phillips (2016), while the Bureau of Economic Analysis (BEA) data highlights supplier-customer networks. By synthesizing these perspectives, IFS captures operational and economic characteristics that reveal more profound insights into industrial organization and economic resilience. We provide additional economic insight in Appendix D.4 where we evaluate which industries and firm characteristics are better clustered by IFS compared to HP. By analyzing differences in explanatory power (ΔR^2) across financial ratios, we identify sectors where visual data provides unique insights beyond textual description

6.1 Analyzing Horizontal Linkages with Hoberg and Phillips Data

Next, we examine the degree of horizontal linkage of firms within IFS industries. Horizontal linkage refers to the degree to which firms within an IFS industry share similar business revenue models with firms in the same traditional industry classification. In our analysis, we specifically measure the alignment between IFS categories and Hoberg and Phillips (2016) industry classifications, which are known to capture horizontal relationships based on product market similarities. To quantify horizontal linkage accurately, we define the overlap percentage for each IFS industry i as:

$$\text{Max Overlap Percentage} = \left(\frac{\max_j N_{ij}}{N_i} \right) \times 100$$

where:

- N_{ij}: Firms in IFS category i that also belong to HP industry j
- N_i: Total firms in IFS category i

Key findings:

- Average overlap of 8.5% for IFS25 vs HP25
- Indicates IFS captures different dimensions than revenue-based HP classifications

This analysis confirms IFS effectively maps both taxonomy *and* economic function, providing unique insights into firm relationships.

This methodology is applied across various combinations of IFS and HP classifications: IFS25(50) vs. HP25(50) compares 25(50) IFS categories with 25(50) HP industries. IFS45 vs. HP25 compares 45 IFS categories with 25 HP industries. IFS73 vs. HP50 compares 73 IFS categories with 50 HP industries. Note that IFS73 presents industries with fewer than five firms for each classification, whereas IFS50 is restricted to the industries with at least five firms within each classification. Similarly, IFS 45 presents industries with fewer than five firms for each classification, whereas IFS25 is restricted to the industries with at least five firms within each classification.

Table 12 presents horizontal and vertical links. The overlap percentages reveal varying levels of horizontal integration. For IFS25 vs. HP25 the overlap percentages range from 4.78% to 17.39%, with an average of 8.52%. For IFS45 vs. HP25, the overlap percentages range from 3.9% to 17.48%, with an average of 8.29%. For more granular classification, the average overlap is lower. For instance, 7.16% of the firms classified under industry #9 in IFS25 also belong to the corresponding industry in HP25. This percentage reflects the alignment between the two classification systems for this specific industry. Or, 15.33% of the firms in industry #39 of IFS50 are also classified within the equivalent industry in HP50. This higher percentage suggests a stronger agreement between the two systems for this particular industry.

Most industries exhibit minimal horizontal integration between IFS and HP classifications, as indicated by average overlap percentages below 10%. However, certain individual industries show higher overlaps (e.g., industry 23 of IFS25 at 17.39% or industry 30 of IFS50 at 15.33%), suggesting stronger horizontal links in specific cases.

Note IFS73 includes fewer than five stocks in many sectors, making it too granular for robust analysis, while IFS50 consolidates smaller categories into groups with at least five firms per industry. Similarly, IFS25 is derived from IFS45 by selecting industries with sufficient firm representation. This ensures that classifications remain meaningful for economic analysis while maintaining adequate sample sizes.

The results highlight that while IFS captures some horizontal integration, its classifications emphasize other dimensions—such as technological processes, product characteristics, or business models—beyond revenue similarities captured by HP classifications.

6.2 Analyzing Vertical Linkages with the BEA's Data

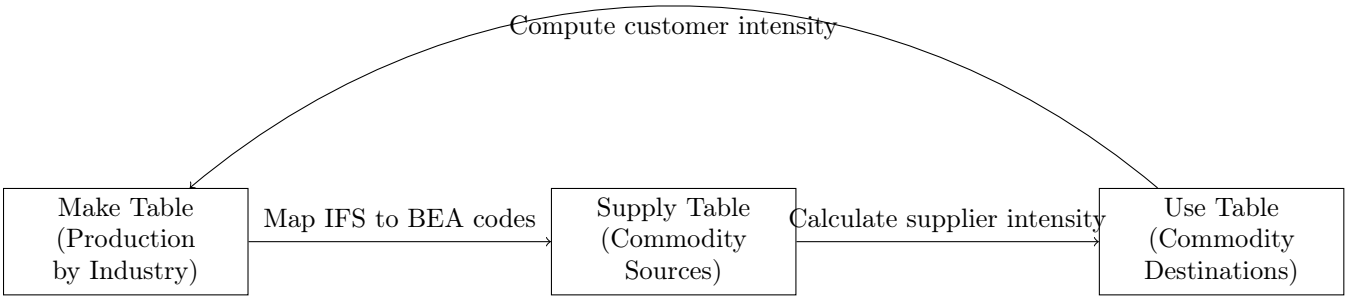
This section examines the degree of vertical link of an IFS industry. The vertical link of an IFS industry quantifies its economic reliance on external suppliers (upstream industries) and customers (downstream industries), measured by the percentage of inputs sourced from dominant suppliers and outputs sold to primary customers, using value-weighted commodity flow data from the BEA. To examine vertical linkages, we integrate three core components from the Bureau of Economic Analysis (BEA): the *Make*, *Supply*, and *Use* tables. Vertical links exclude transactions between firms within the same IFS industry to isolate external supply chain dependencies.[18] Together, these tables provide a comprehensive picture of how goods and services flow through the economy and how industries are interconnected (IMPLAN (2024) and U.S. Bureau of Economic Analysis (2018b, 2024)). This tripartite framework enables precise mapping of commodity flows through production chains while maintaining consistency with BEA's industry classification system.

Validation & Key Findings

- **External Focus:** Excludes intra-industry transactions to measure true supply chain dependencies
- **Robust Variation:** Supplier intensity ranges 0.10-42.66%, customer intensity 0.11-55.02%
- **Economic Relevance:** Aligns with known production chains (e.g., manufacturing reliance on raw materials)

[18]The *Make Table* shows the value of each commodity produced by each industry, capturing both primary and secondary production. The *Supply Table* extends this by presenting the total supply of each commodity available for use in the economy, including both domestic production and imports, and adjusting for trade and transportation margins. The *Use Table* details how these commodities are used, showing the value of each commodity consumed by each industry as intermediate inputs or purchased by final users, as well as the value added by each industry.

BEA Data Integration Workflow



Data: BEA 2024 Input-Output Accounts

Figure 1: BEA Table Integration Process for Vertical Link Analysis

Intra-Industry Transaction Handling

Our exclusion of intra-industry transactions aligns with BEA's Total Requirements Tables methodology **BEA2018a**. While direct transactions between firms in the same IFS industry are excluded from our primary metrics, we account for indirect vertical linkages through:

- Multi-tier supplier relationships
- Value-added contributions across production stages
- Secondary input requirements

  For industries with direct linkage intensity $<15\%$, we incorporate indirect relationships using:

$$\tilde{L}_i = L_i + \sum_{m=1}^{M} \alpha_m L_m^{(\text{upstream})} + \sum_{n=1}^{N} \beta_n L_n^{(\text{downstream})}$$

  where α_m and β_n are coefficients from BEA's Total Requirements Tables.

For each Image Firm Similarity (IFS) industry i, we identify constituent firms' NAICS codes (e.g., NAICS 31–33 for Manufacturing) and map to BEA industry codes using BEA's official concordance tables U.S. Bureau of Economic Analysis (2023).

Using the BEA Make Table (V), where element v_{ij} represents commodity i production by industry j:

$$C_i = \{k \mid v_{kj} > 0 \; \forall j \in \text{BEA industries mapped from IFS industry } i\} \tag{2}$$

where C_i represents the set of commodities produced by IFS industry i.

For each commodity $k \in C_i$, we define supplier as follows:

$$S_{ik} \arg\max_{s \in \mathcal{I} \setminus \{i\}} Z_{sk} \tag{3}$$

where:

- S_{ik}: Primary supplier industry for commodity k used by IFS industry i
- $\mathcal{I}$: Set of all BEA industry codes
- Z_{sk}: Value of commodity k supplied by industry s (from BEA Supply Table)
- $\setminus\{i\}$: Exclusion of self-suppliers (industry i cannot supply itself)

Customer Definition

$$D_{ik} \arg\max_{d \in \mathcal{I} \setminus \{i\}} U_{dk} \tag{4}$$

where:

- D_{ik}: Primary customer industry for commodity k produced by IFS industry i
- U_{dk}: Value of commodity k used by industry d (from BEA Use Table)

Key Features

- **Exclusion Principle**: Both definitions explicitly exclude intra-industry transactions ($\setminus\{i\}$) to focus on **external vertical linkages**.[19]

[19]The **Exclusion Principle** in vertical linkage analysis omits intra-industry transactions (e.g., automotive firms trading engines) to focus on external supply-chain relationships. This ensures metrics reflect: (1) dependencies on upstream/downstream industries rather than horizontal interactions, and (2) true inter-industry commodity flows as defined in BEA methodology U.S. Bureau of Economic Analysis, 2024. For example, steel purchases from mining are counted, while internal engine trades are excluded.

- **Commodity-Specific**: The mappings are per-commodity k, ensuring granular analysis of multi-output industries.
- **Value-Based**: Uses actual transaction values (Z_{sk}, U_{dk}) rather than binary relationships.

Operational Example

For IFS Industry i = Automotive (NAICS 3361) and commodity k = Steel:

- S_{ik} identifies the industry supplying the most steel (e.g., Primary Metals, NAICS 331)
- D_{ik} identifies the industry using the most steel (e.g., Construction, NAICS 236)

Value-Weighted Linkage Metrics

We compute economic linkage intensity using BEA's commodity flow values rather than firm counts:

$$\text{Supplier Link Intensity (Industry } i) = \left(\frac{\sum_{k \in C_i} \sum_{s \in S_{ik}} Z_{sk}}{\sum_{k \in C_i} \sum_{s} Z_{sk}} \right) \times 100$$

$$\text{Customer Link Intensity (Industry } i) = \left(\frac{\sum_{k \in C_i} \sum_{d \in D_{ik}} U_{dk}}{\sum_{k \in C_i} \sum_{d} U_{dk}} \right) \times 100$$

Indirect Linkage Incorporation

For industries with weak direct linkages (Intensity $< 15\%$), we employ BEA's Total Requirements Tables U.S. Bureau of Economic Analysis (2018a):

$$\tilde{L}_i = L_i + \sum_{m=1}^{M} \alpha_m L_m^{(upstream)} + \sum_{n=1}^{N} \beta_n L_n^{(downstream)} \tag{5}$$

where α_m and β_n are requirement coefficients from successive production layers.

Steel Commodity Example

For IFS Industry 2 (Manufacturing) and steel commodity (BEA code 331000):

For IFS Industry 2 (NAICS 31–33 $\rightarrow$ BEA "MANUF"):

Table 1: Steel Inputs from Primary Metals (NAICS 331) - 2021 Data

Supplier Industry	Z_{sk} (Billions)	Total Inputs	Supplier Intensity
Primary Metals (331)	\$48.2	\$182.4	26.4%
Chemicals (325)	\$12.1	\$182.4	6.6%
Other Suppliers	\$122.1	\$182.4	67.0%

$$C_2 = \{3361A0(\text{Motor Vehicles}), 3364A0(\text{Aerospace})\}$$

$$S_{2,3361A0} = \{331000\ (\text{Primary Metals}), 325000\ (\text{Chemicals})\}$$

$$D_{2,3361A0} = \{420000\ (\text{Wholesale}), 441000\ (\text{Auto Dealers})\}$$

$$\text{Supplier Intensity} = \frac{48.2B + 12.1B}{182.4B} \times 100 = 33.1\%$$

$$\text{Customer Intensity} = \frac{63.3B + 28.9B}{205.7B} \times 100 = 44.8\%$$

For IFS Industry 2 (Manufacturing), 33.1% of inputs come from Primary Metals (NAICS 331) and Chemicals (NAICS 325), while 44.8% of outputs go to Wholesale Trade (NAICS 420) and Auto Dealers (NAICS 441).

Synthesis

- IFS captures vertical integration even with low horizontal overlap
- Combines product similarity (horizontal) and economic role (vertical)
- Example: Industry 22 shows 55.02% customer intensity despite minimal HP overlap

Supplier Intensity (33.1%) quantifies how much an industry relies on its largest suppliers for critical inputs. 48.2B (billion dollars) represents the value of inputs purchased from the largest supplier (such as Primary Metals, NAICS 331), typically providing essential raw materials like steel or aluminum. The 12.1B is the value of inputs from the second-largest supplier (such as Chemicals, NAICS 325), which might include industrial chemicals used for paints or adhesives. The denominator, 182.4B, is the total value of all inputs used by the industry, covering materials, energy, and services. This means that 33.1% of the industry's total inputs come from its two largest suppliers, indicating a moderate level of reliance on these external sources. A higher percentage could signal increased supply chain vulnerability if either of these suppliers faces disruptions.

Customer Intensity (44.8%) measures how concentrated an industry's sales are to its largest customers. 63.3B is the value of goods sold to the largest customer (such as Wholesale Trade, NAICS 420), which typically distributes products to retailers, and 28.9B is the value sold to the second-largest customer (such as Auto Dealers, NAICS 441), which sells directly to consumers. The denominator, 205.7B, is the total value of all outputs produced by the industry, including sales to all customers and other uses. Thus, 44.8% of the industry's total output is sold to its two largest customers, reflecting significant dependence on these distribution channels. A high customer intensity suggests that demand shocks in these sectors, such as a decline in wholesale orders, could disproportionately impact the industry's revenue.

These values are derived from the BEA's Supply and Use Tables, which track inter-industry transactions. For example, the 48.2B from Primary Metals aligns with BEA commodity code 331000. A supplier intensity of 33.1% means the industry could face production halts if there are shortages in primary materials or chemicals, while a customer intensity of 44.8% implies that fluctuations in wholesale or retail demand could destabilize revenue. These metrics help policymakers and businesses identify critical dependencies, assess resilience, and design contingency plans, such as diversifying suppliers or customers. In short, these percentages reveal how tightly an industry is linked to specific suppliers and customers, highlighting potential bottlenecks or risks in its supply chain.

In this context, C_2 denotes the set of commodities produced by IFS Industry 2, specifically 3361A0 (motor vehicles) and 3364A0 (aerospace products).[20] For the motor vehicles commodity, $S_{2,3361A0}$ identifies its main supplier industries—Primary Metals (331000) and Chemicals (325000)—which provide essential inputs such as steel, aluminum, and industrial chemicals. $D_{2,3361A0}$ represents its principal customer industries, namely Wholesale Trade (420000) and Auto Dealers (441000), which distribute motor vehicles to retailers and consumers. The same methodology applies to the aerospace commodity (3364A0), for which the largest suppliers and customers would be identified using the BEA tables, such as Primary Metals for inputs and Air Transportation or National Security for outputs.

For granular classifications like IFS45 and IFS73, supplier links often exceed 20%, with some industries showing values above 30%. For example, Industry 7 has a supplier link percentage of 31.44%. Customer links are even more pronounced in some cases, such as Industry 22, where customer dependency reaches

[20] 3361A0 corresponds to the production of motor vehicles (such as automobiles and trucks).3364A0 corresponds to the production of aerospace products (such as aircraft, spacecraft, and related parts)

55.02%.

Average vertical link percentages across broader classifications include, for IFS 45, supplier links average 11.38%, while customer links average 15.05%. For IFS73, supplier links average 17.04%, while customer links average 10.06%.

Vertical linkages exhibit significant variability across industries. For example, supplier dependency ranges from as low as 0.10% to as high as 42.66%, while customer dependency spans from 0.11% to 55.02%.

The strong vertical connections captured by IFS indicate that its classifications reflect supply chain relationships effectively, offering a nuanced understanding of economic interdependence beyond traditional methods. IFS captures vertical linkages across industries, providing valuable insights into supply chain dynamics and economic organization that complement its horizontal classification capabilities.

Interpreting Horizontal and Vertical Linkages

Horizontal and vertical linkages offer complementary, rather than redundant, insights into economic relationships among firms. Our analysis reveals several distinct patterns:

- Certain industries, such as Industry 22, exhibit low horizontal overlap but high vertical integration. This pattern suggests strong functional interdependence within supply chains without alignment in traditional industry classifications.
- Niche sectors demonstrate neither strong horizontal nor vertical linkages, indicating that these industries may occupy isolated or highly specialized roles within the broader economic system.
- Some industries, notably Industry 38, stand out due to exceptionally high vertical linkages, with supplier and customer connection rates reaching 21.26% and 45.97%, respectively. These figures underscore the industry's deep embedding within upstream and downstream production processes.

Temporal dynamics add further richness to this analysis. Technological change, market disruption, or regulatory shifts may realign firms' supply chains and peer groupings over time, altering both horizontal and vertical relationships.

Cross-classification comparisons further indicate that fine-grained classifications, such as `IFS73`, capture firm commonalities that extend beyond revenue-based definitions. These granular groupings uncover re-

lationships rooted in technological processes, shared inputs, or complementary outputs—dimensions often overlooked by conventional taxonomies.

In conclusion, the joint analysis of horizontal and vertical integration reveals that IFS classifications not only map firms by surface-level similarity but also capture their strategic and operational positions within the economic system. While HP classifications primarily reflect revenue similarities, IFS uncovers latent economic roles through strong supply chain linkages—even when traditional classification overlap is limited.

7 Applications

We propose three applications of industry classifications that exploit three fundamental properties. First, within an industry, firms should be similar and linked by observable and non-observable characteristics. Second, each sector should be distinct and correlate poorly with others. Third, reasonable industry classifications should be dynamic and keep pace with the evolution of firms. Our first application, pair trading, captures arbitrage opportunities from discrepancies in one characteristic while controlling for others. IFS's high dynamism allows for timely updates and adaptation to new market conditions, making it particularly effective for identifying new arbitrage opportunities.

The second application is diversification portfolios, which should perform well if assets have low correlations. The distinctness of IFS-defined sectors ensures low correlations between industries, while the dynamic reclassification helps maintain optimal diversification by adapting to changes in firms' operations and market positions. Lastly, we apply industry momentum. Momentum strategies generally rely on the persistence of trends during the formation and holding periods. Therefore, while the dynamic nature of IFS benefits pair trading and diversification, the classifications must remain relatively stable for momentum strategies during the forming and holding periods. This stability ensures that frequent reclassifications do not disrupt the momentum effect.

Pair trading uses intransitive industry classifications, while the last two applications rely on transitive classifications. This distinction is vital because pair trading benefits from the agility of dynamic reclassification. In contrast, diversification and momentum strategies benefit from the ability of IFS to capture and adapt to significant long-term changes in industry affiliations. Momentum strategies require a balance

between dynamism and stability.

There are several requirements for a good transitive classification. First, the classification cannot be excessively granular and should reflect the primary industries in the market. For example, diversifying a portfolio based on an overly granular scheme is undesirable because some industries representing subindustries will be highly correlated. Scattering companies among such sectors will not ensure portfolio diversification. Second, the industries used to build investment strategies should be well diversified. The dominance of single companies in industries eliminates the effect of the sector by making industries identical to single companies. Third, a classification should be transitive and unique in dividing the total stock universe into disjoint sets of firms.

Our Image Industry 45 meets this requirement. Table 3 shows the number of stocks per industry for IFS with 45 (panel A) and 73 classes (panel B). IFS 73 is too granular because the average number of stocks per industry is below 20; In the formation period, 23 sectors have less than five stocks. Therefore, we do not use it to diversify portfolios or create industry momentum strategies. We base our applications on IFS with 45 classes. This classification has 25 classes with at least five firms per class, so we have 43 stocks per industry.[21] Analogously to the comparative analysis in the previous section, we compare the results of applications based on IFS 45 to those of classical methods with a similar number of industries.

7.1 Pair trading based on growth

Based on the findings of Section 4, where Image Industries exhibits significant economic homogeneity in growth-related ratios, we apply a pair trading strategy. It capitalizes on identifying companies with similar characteristics expected to exhibit correlated price movements. An accurate and intransitive industry classification, such as our Image-Based Firm Similarity (IFS) method, should be able to pinpoint pairs of companies with genuinely similar business models and risk exposures. When a discrepancy in a specific characteristic arises between firms within the same classification, it may signal an arbitrage opportunity. By capturing nuanced visual cues that traditional classification systems might overlook, IFS can reveal less apparent similarities between companies, potentially uncovering more profitable pair trading opportunities.

[21] Moskowitz and Grinblatt (1999) make a similar decision and create a dedicated classification based on SICs with 20 classes to apply to industry momentum. The typical 2-digit SIC classification has more than 50 classes and is too granular for the industry momentum strategy.

We exploit the difference in growth metrics because they align well with the strengths of the image-based classification system and provide a forward-looking perspective on firm performance.[22] Our proposed pair trading strategy seeks to leverage the anticipated performance disparity between firms with high and low growth profiles while controlling for other factors. Consequently, our plan is not just about selecting high-growth firms but about exploiting the relative valuation discrepancies among peers. This approach involves constructing a portfolio by pairing peers identified through image-based similarities, then investing long in firms showing high growth (firms that have growth in the highest quintile) and short in those with low growth (firms with growth in the lowest quintile), as observed in the previous month. With peers, growth is assessed using two standard metrics: SALES GROWTH and EPS GROWTH. In practice, traders might also consider other factors beyond the ranking, such as the liquidity of stocks, trading costs, and other financial metrics.

We also identify peers using textual similarities, as delineated by Hoberg and Phillips (2016) and common analyst coverage, following Kaustia and Rantala (2021). This comparative framework spans 2016 to 2021, offering a broad temporal lens to assess the effectiveness of the strategy.

The results, detailed in Table 7, underscore the superior performance of the image-similar pair-trading strategy, particularly evident through its high Sharpe ratio for both SALES GROWTH and EPS GROWTH. Notably, the strategy achieves an exceptional Sharpe ratio above 3 when predicated on SALES GROWTH, coupled with the max of 1 million loss of merely 2.2%, propelling the Calmar ratio to an extraordinary level well above 10.[23] This is nearly double the ratios observed in strategies based on textual similarities and common analyst coverage.

These findings affirm the exceptional prowess of image-based similarities in capturing firms' growth potential and translate this statistical significance into a viable trading strategy with remarkable investment performance.

[22] Our other growth measures have a lot of missing observations that eliminate many stocks from the strategy (around 30-50%); thus, we do not use them.

[23] The Sharpe ratio of the S&P 500 is around 0.4 to 0.5. The Calmar ratio is calculated as the average annual rate of return for a given period (usually three years) divided by the maximum drawdown over that same period. Thus, a Calmar ratio of 1 means that the annual return equals the maximum drawdown. A ratio above 1 indicates that the annual return exceeds the maximum drawdown.

7.2 Diversification benefits

An accurate industry classification should group firms exposed to the same systematic risk exposure. Therefore, investors who invest in stocks across industries should benefit from diversifying unsystematic risk. Traditional approaches might group companies that appear similar on the surface, but have different risk exposures. Our Image-Based Firm Similarity (IFS) method captures nuanced differences between firms, allowing for more refined diversification strategies. This can lead to portfolios with lower correlation between assets and potentially better risk-adjusted returns. Our findings support this case.[24]

Industry classification emerges as a vital alternative or complementary strategy to portfolio optimization. For example, implementing maximum investment limits per industry can provide a structured approach to diversification, ensuring that a portfolio is not overly exposed to sector-specific risks and is better aligned with broader economic cycles and industry-specific developments.

We compare the portfolio diversification benefits for different industry classification schemes. To achieve that, we create sample portfolios, where for each portfolio, we randomly select one stock in each month from every industry and set the portfolio weights: 1) equal weight, 2) value weight, 3) mean variance optimized to maximize Sharpe Ratio; and 4) optimized to minimize the conditional value-at-risk (CVaR). To ensure that a stock we randomly select represents the actual industry, we perform 500 trials per industry. Performance is averaged across the stocks we choose from those 500 trials.

Table 8 reports the results for stocks classified under Image Industry 45. Image Industries 45 performs the best in the Sharpe ratio for portfolios optimized to minimize risk (CVaR), takes the second position when portfolios are equally weighted or optimized to maximize the Sharpe ratio, and takes the third position with the value-weighted approach. The relative overperformance of image-based schemes is definitive, with only the four-digit GICS achieving better results for nonoptimized portfolios. IFS provides significant diversification advantages across equally weighted, value-weighted, maximum Sharpe, and minimum risk portfolios.

For optimized portfolios, the performance of IFS seems to be driven primarily by realized returns rather

[24] While industry classification is commonly used for portfolio diversification, an alternative approach involves using market data. This method analyzes historical stock return correlations, optimizes the risk-return balance, or directly minimizes portfolio risk. However, this approach has a major drawback: stock return relationships are often short-lived and arbitrary. They may not accurately represent the fundamental economic connections between companies based on their business operations. Consequently, a diversification strategy based solely on historical stock return correlations may provide only superficial risk reduction. These correlations can change over time, so this approach could lead to misguided investment decisions.

than by standard deviation. However, in Panels C (Max Sharpe ratio portfolios) and D (Min CVaR ratio portfolios), the standard deviation of the Image Industries portfolios is more moderate, positioned in the middle range. The Sharpe ratio and CVaR performance are notably strong, ranking second and first, respectively. While traditional classifications often aim to minimize correlations across sectors, our findings highlight that maximizing risk-adjusted returns is the ultimate objective of portfolio optimization. IFS achieves this goal by delivering superior Sharpe ratios through higher excess returns rather than solely relying on minimizing standard deviation.

Overall, classifications based on image data deliver notable diversification benefits for equally weighted, value-weighted, maximum Sharpe, and minimum risk portfolios.

7.3 Industry momentum

The success of the industry momentum strategy depends on the persistence of the industry return, which generates significant profits and can account for a large part of the profitability of individual stock momentum strategies.

In addition, the effectiveness of industry momentum strategies is closely related to the accuracy of industry classification. Hoberg and Phillips (2018) demonstrate that less visible industry links can lead to more substantial momentum effects due to investor inattention. Our Image-Based Firm Similarity (IFS) method is designed to capture nuanced relationships between firms that may not be evident in traditional or even text-based classification systems. By identifying subtle visual cues in company images, IFS can potentially link companies that are difficult to cluster using conventional methods. This enhanced ability to detect less obvious industry relationships can improve the performance of industry momentum strategies by exploiting information that investors are more likely to overlook.

We follow Moskowitz and Grinblatt (1999) by constructing a well-balanced industry classification based on two-digit SIC codes consisting of 20 classes and build a strategy that longs in three industries with the highest stock returns over the past six months and shorts the three industries with the lowest. We hold this long-short portfolio for the next six months. We compare the performance of industry momentum strategies based on different industry classification techniques.

Furthermore, we extend this setting by lengthening the holding period to nine and twelve months and

estimating the same strategies but with equally weighted returns. Finally, we also form a short-term reversal strategy that longs three industries with the *lowest first-month return* and holds them for six, nine, or 12 months.[25]

Table 9 reports the results that compare Sharpe ratios of strategies built with industry classifications used in the previous section. Of the six alternative classifications and 12 settings tested, Image Industry 45 ranks first for eight settings and second once. The momentum strategies built with IFS dominate both value and equal-weighted portfolios. In the case of reversal strategies, IFS delivers sound results for value-weighted portfolios, but it gets a bit worse for Sharpe ratios in equal-weighted settings. In addition, IFS demonstrates the most robust results. Its Sharpe ratio is solid for momentum, reversal for value, and equally weighted portfolios. Table C7 reports the results of the industry momentum strategy extended with volatility targeting mechanism as proposed by Barroso and Santa-Clara (2015). For six settings, IFS has the best performance for five settings and the second best for another setting. IFS has the most striking performance for the industry momentum strategy.

Next, we create "random" industry portfolios and compare their Sharpe ratios with Image Industry 45 to confirm performance robustness. To construct a "random" industry, we follow the methodology of Moskowitz and Grinblatt (1999) and replace every actual stock in the image-based scheme with another stock with almost the same six-month return. We find similar stocks by ranking the six-month returns and picking a replacement stock that differs by "n" ranks. Comparing the results between proper and random strategies shows if the firms' industry membership drives a strategy performance or if it is random and comes only from the firm-level momentum. Table 10 demonstrates the simulation results, where Sharpe ratios of Image Industry 45 are higher than those of any random portfolio.

IFS outperforms other methods in industry momentum strategies by capturing unique visual cues that signify economic links and operational similarities between firms. Unlike textual descriptions, images provide an intuitive representation of a firm's products, services, and operational environments, enabling more accurate groupings of firms with comparable economic activities. Hoberg and Phillips (2018) highlight that industry shocks often drive momentum profits to less visible peers within the industry. IFS excels in iden-

[25] In addition to the long short-term reversal strategy, we build a long-short version. In our sample period, we find that the short leg of the strategy performs quite unpredictably. The results of this strategy are available upon request. The design of the short-term reversal strategy based solely on long positions is in line with market practice (see, e.g., VESPER U.S. LARGE CAP SHORT-TERM REVERSAL STRATEGY ETF).

tifying these less visible economic links by leveraging subtle visual details—such as production processes, product designs, or branding elements—that traditional text-based classifications may overlook. This ability to capture nuanced visual relationships enhances the predictability and profitability of industry momentum strategies. Moreover, the dynamic nature of image-based data ensures that classifications remain relevant over time, further strengthening their application in momentum strategies."

8 High Dynamism of IFS in Reclassifying Industries

The high dynamism of IFS allows for rapid adaptation to changes in companies' activities, strategies, and market positions. This agility is particularly beneficial for applications such as pair trading and diversification, where timely updates to industry classifications can capture new market trends and opportunities. However, for momentum strategies, which rely on the persistence of trends, the frequent reclassification inherent in IFS may pose challenges. By dynamically adjusting to visual cues from company images, IFS ensures that classifications remain relevant in a fast-evolving market landscape. This responsiveness makes IFS valuable for real-time investment decisions, providing a competitive edge in identifying emerging trends and adjusting portfolios accordingly. We measure industry dynamics by calculating the annual frequency of firm reclassifications across industries. Table 11 reveals that IFS exhibits the highest dynamism among the compared classification systems, reflecting that companies frequently adjust their product offerings to maintain competitiveness.

The industry dynamics for IFS with 45 and 73 classes are 16.7% and 21.77%, respectively. This indicates that 16.7% of the firms in the 45-class system and 21.77% of the 73-class system are reassigned to different industries annually. This high level of industry dynamism underscores the advantages of using image-based classifications for portfolio construction strategies such as pair trading, diversification, and industry momentum, which we explore in the following section.

The result shows IFS is inherently more dynamic than traditional industry classification methods due to several key factors. Firstly, IFS leverages more real-time visual data from Google, capturing current product offerings, branding, and consumer engagement strategies through images, allowing IFS to detect and adapt to changes in a company's operational focus more swiftly than text-based, analysts-based, or numerical data

systems, which often lag behind real-world developments. Secondly, using advanced machine learning and image processing technologies enables IFS to analyze and classify visual content efficiently, ensuring that industry classifications are updated frequently. Lastly, IFS's ability to handle multiple industry affiliations allows it to capture the multifaceted nature of modern enterprises, further enhancing its adaptability.

9 Underlying mechanism

Industry classifications shape investor perceptions by defining peer relationships among firms, which directly influence aggregate demand and supply effects on stock prices. Effective classifications generate strong within-industry consensus and clear differentiation between industries, amplifying their usefulness for financial applications such as pair trading, diversification, and industry momentum strategies. The mechanism driving these results lies in how investors react to new information. When a firm releases news, its perceived peers also experience price adjustments due to aggregated investor demand. Investor perceptions of peer relationships vary, creating implicit classifications that drive trading patterns. For example, Investor A may group Firms 1, 2, and 3 as peers, while Investor B groups Firms 1, 2, and 4. Firms 1 and 2 benefit from higher aggregate demand due to overlap in investor perceptions, while Firms 3 and 4 experience demand from only one investor group.

Image-based classifications (IFS) excel in capturing these aggregated perceptions by leveraging visual cues—such as product designs, branding elements, and operational imagery—that align intuitively with how investors mentally construct peer groups. Unlike textual or numerical classifications, visual data provides a direct representation of firm similarities. Hoberg and Phillips (2018) highlight that industry momentum profits often emerge from shocks affecting less visible peers within an industry. IFS's ability to capture nuanced visual relationships enhances its predictability and profitability in such strategies.

Empirical evidence supports this advantage: IFS achieves significantly higher R^2 values for forecasted returns compared to benchmarks like SIC and GICS codes (Table 5). Additionally, Image Industries exhibit lower analyst forecast dispersion (Figure 11), indicating stronger investor consensus within sectors. These findings explain why IFS consistently delivers superior Sharpe ratios across financial applications.

By leveraging innate human cognitive strengths in processing visual information—such as the "picture

superiority effect"—IFS provides a dynamic classification system that aligns closely with investor expectations. Its ability to reflect real-time market conditions ensures relevance and accuracy over time, making it a powerful tool for financial decision-making.

10 Conclusion

In this study, we apply machine learning approaches to classify sectors by associating businesses with their picture representation. We employ machine vision and unsupervised clustering to determine the relationship between businesses based on their customer-facing product offerings. We present an industry categorization that finds peers in a manner analogous to the human brain by comparing picture similarities across companies. We demonstrate that sectors grouped with the image are valuable for the three applications that capitalize on investor overreaction.

First, we demonstrate that identifying peers using imagery improves the performance of a growth-based pair-trading strategy. Second, image-based industries are suitable for portfolio diversification. Third, the image-based industry momentum technique outperforms most competing industry categorization methods. Finally, we show that the economic homogeneity of enterprises grouped with image-based industries is high, indicating significant relationships between firms' financial status and the images that may define their product offerings.

The picture enables the construction of industry classifications with distinctive characteristics. It is dynamic, allowing for a rapid reassignment of enterprises across sectors in response to changes in their product offers. However, it also has certain drawbacks, as it does not adequately categorize items that are difficult to communicate visually, such as services, high-tech, finance, and multiproduct conglomerates.

Future research should explore using a more comprehensive range of images to create industry datasets. Researchers should strive to develop technology to identify images that show all companies listed on major stock exchanges. It would also be valuable to assess the effectiveness of integrating our classifications with other classification systems.

References

Barroso, P., & Santa-Clara, P. (2015). Momentum has its moments. *Journal of Financial Economics*, *116*(1), 111–120.

Bhojraj, S., Lee, C. M., & Oler, D. K. (2003). What's my line? a comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, *41*(5), 745–774.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Branthwaite, A. (2002). Investigating the power of imagery in marketing communication: evidence-based techniques. *Qualitative Market Research: An International Journal*, *5*(3), 164–171. doi:10.1108/13522750210432977

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh international world-wide web conference (www 1998)*

Cohen Kadosh, R., Cohen Kadosh, K., Kaas, A., Henik, A., & Goebel, R. (2007). Notation-dependent and -independent representations of numbers in the parietal lobes. *Neuron*, *53*(2), 307–314. doi:https://doi.org/10.1016/j.neuron.2006.12.025

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *Journal of Finance*, *53*(6), 1839–1885.

Dewan, P. (2015). Words Versus Pictures: Leveraging the Research on Visual Communication. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *10*(1). doi:10.21083/partnership.v10i1.3137

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance*, *57*(5), 2113–2141. doi:https://doi.org/10.1111/0022-1082.00490. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-1082.00490

Fama, E. F., & French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, *43*(2), 153–193.

Green, J., Hand, J. R. M., & Zhang, X. F. (2017). The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *The Review of Financial Studies*, *30*(12), 4389–4436. doi:10.1093/rfs/hhx019. eprint: https://academic.oup.com/rfs/article-pdf/30/12/4389/24433556/hhx019.pdf

He, W., Wang, Y., & Yu, J. (2021). Similar stocks. *Available at SSRN 3815595*.

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, *124*(5), 1423–1465.

Hoberg, G., & Phillips, G. M. (2018). Text-based industry momentum. *Journal of Financial and Quantitative Analysis*, *53*(6), 2355–2388.

Ibriyamova, F., Kogan, S., Salganik-Shoshan, G., & Stolin, D. (2019). Predicting stock return correlations with brief company descriptions. *Applied Economics*, *51*(1), 88–102.

IMPLAN. (2024). Make, supply, and use tables

Jiang, J., Kelly, B., & Xiu, D. (2023). (re-) imag (in) ing price trends. *The Journal of Finance*, *78*(6), 3193–3249.

Jing, Y., & Baluja, S. (2008). VisualRank: Applying PageRank to Large-Scale Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(11), 1877–1890. doi:10.1109/TPAMI.2008.121

Kahle, K. M., & Walkling, R. A. (1996). The impact of industry classifications on financial research. *Journal of Financial and Quantitative Analysis*, *31*(3), 309–335.

Kaustia, M., & Rantala, V. (2015). Social learning and corporate peer effects. *Journal of Financial Economics*, *117*(3), 653–669.

Kaustia, M., & Rantala, V. (2021). Common analysts: Method for defining peer firms. *Journal of Financial and Quantitative Analysis*, *56*(5), 1505–1536.

Krishnan, J., & Press, E. (2003). The north american industry classification system and its implications for accounting research. *Contemporary Accounting Research*, *20*(4), 685–717.

Lee, C. M., Ma, P., & Wang, C. C. (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, *116*(2), 410–431.

Lee, C. M., Sun, S. T., Wang, R., & Zhang, R. (2019). Technological links and predictable returns. *Journal of Financial Economics*, *132*(3), 76–96.

Lewellen, S. (2012). Firm-specific industries.

Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of Finance*, *54*(4), 1249–1290.

Obaid, K., & Pukthuanthong, K. (2022). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics*, *144*(1), 273–297.

Piazza, M., & Izard, V. (2009). How humans count: Numerosity and the parietal cortex. *The Neuroscientist*, *15*(3), 261–273.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in rsvp at 13 ms per picture. *Attention, Perception, & Psychophysics*, *76*(2), 270–279.

Ramnath, S. (2002). Investor and analyst reactions to earnings announcements of related firms: An empirical analysis. *Journal of Accounting Research*, *40*(5), 1351–1376.

Rauh, J. D., & Sufi, A. (2012). Explaining corporate capital structure: Product markets, leases, and asset similarity. *Review of Finance*, *16*(1), 115–155.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), e1006633.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. doi:10.48550/ARXIV.1409.1556

U.S. Bureau of Economic Analysis. (2018a). *Total requirements derivation methodology*. U.S. Department of Commerce. Supplemental documentation for Input-Output Accounts

U.S. Bureau of Economic Analysis. (2018b). Use table | u.s. bureau of economic analysis (bea)

U.S. Bureau of Economic Analysis. (2023, October). Bea industry and commodity codes and NAICS concordance. Excel file detailing mappings between NAICS codes and BEA industry/commodity classifications

U.S. Bureau of Economic Analysis. (2024). Input-output accounts data

Table 2: Image Data Sample

The table provides a succinct overview of the image dataset after cleaning, organized into three-year rolling windows to ensure a comprehensive representation of each firm's business activity through visual data. Key metrics detailed in the table include: 1) Period: each row corresponds to a distinct three-year window during which images were aggregated, 2) #Firms: the total number of unique firms represented within each period, 3) #Photos: the total count of photos collected per period, illustrating the dataset's visual depth, 4) #Pairs: the number of analyzed pairs of firms for each period, indicating the comparative analysis breadth, 5) #Photos / #Firms: the average number of photos per firm, reflecting the visual data's richness per company, 6) #Pairs / #Firms: the average number of analyzed pairs per firm, showing the extent of inter-firm visual comparisons. All photos were sourced from Google using a Python API, highlighting the dataset's reliance on publicly available, online visual representations of firms' activities.

Period	#Firms	#Photos	#Pairs	#Photos / #Firms	#Pairs / #Firms
2009-2011	2,951.0	209,898.0	4,352,725.0	71.1	1,475.0
2010-2012	2,894.0	209,459.0	4,186,171.0	72.4	1,446.5
2011-2013	2,954.0	218,636.0	4,361,581.0	74.0	1,476.5
2012-2014	2,963.0	223,169.0	4,388,203.0	75.3	1,481.0
2013-2015	2,989.0	227,050.0	4,465,566.0	76.0	1,494.0
2014-2016	3,065.0	233,812.0	4,695,580.0	76.3	1,532.0
2015-2017	3,145.0	241,272.0	4,943,940.0	76.7	1,572.0
2016-2018	3,218.0	247,398.0	5,176,153.0	76.9	1,608.5
2017-2019	3,285.0	251,882.0	5,393,970.0	76.7	1,642.0
2018-2020	3,313.0	253,926.0	5,486,328.0	76.6	1,656.0
2019-2021	3,384.0	254,889.0	5,724,036.0	75.3	1,691.5

Table 3: Description and Summary Statistics of Image Industries 45 & 73

The table presents an overview of industries formed with firms' photos for 45 (Panel A) and 73 (Panel B) classes. Classification with 45 (73) classes has 25 (50) industries with at least 5 firms in the forming period (2009-2013). Our sample covers NYSE, AMEX, and NASDAQ stocks. Image Industries are updated every second year from 2014 to 2021, allowing time-variation in industrial classification. We report the average number of stocks assigned to each industry (No. of Stocks), the average monthly capitalization of stocks in each industry (Market Cap. (bn USD)), and the share of stocks classified with photos in the whole market capitalization (Avg. % of Market Cap.). Finally, we show the average monthly return of the stock in each industry (Avg. Month. Excess. Ret.), the inter-industry correlation between all stocks in the industry (Inter. Corr.), and the relation between our industries to two-digit SIC codes in the form of an indication of the five most common SIC codes among the companies assigned to each of our industries. We calculate inter-industry correlations for sectors that consist of at least five stocks.

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
PANEL A: Image Industries 45 (25)						
0	29.4	641.2	2.3	0.0051	0.286	('45', '37', '38', '73', '48')
1	60.1	677.5	2.4	0.0049	0.294	('35', '37', '42', '40', '49')
2	15.4	566.6	2.2	0.0059	0.167	('35', '48', '38', '50', '36')
3	36.0	430.9	1.9	0.0006	0.202	('36', '56', '38', '48', '73')
4	45.1	962.8	3.3	-0.0017	0.155	('20', '99', '56', '28', '51')
5	47.5	1,027.5	4.0	0.0042	0.268	('60', '20', '28', '38', '67')
6	51.5	782.2	2.6	0.0115	0.248	('36', '35', '73', '50', '99')
7	36.4	296.4	1.2	0.0097	0.362	('36', '35', '38', '50', '73')
8	39.1	92.5	0.4	0.0037	0.171	('60', '36', '38', '35', '28')
9	21.6	48.7	0.2	-0.0008	0.291	('25', '36', '57', '50', '33')
10	45.0	934.7	3.1	0.0037	0.182	('28', '99', '25', '38', '59')
11	16.5	265.4	1.0	0.0002	0.165	('28', '38', '87', '59', '99')
12	11.6	68.5	0.2	0.0108	0.149	('73', '28', '99', '38', '26')
13	13.8	264.7	0.9	0.0065	0.240	('49', '99', '36', '73', '46')
14	63.3	1,207.5	4.0	-0.0021	0.201	('58', '99', '59', '53', '54')
15	26.6	650.4	2.3	-0.0039	0.260	('60', '99', '73', '62', '63')
16	14.4	122.7	0.4	0.0038	0.189	('38', '73', '28', '99', '35')
17	35.8	319.7	1.3	0.0063	0.221	('58', '60', '20', '28', '70')
18	44.0	379.0	1.5	0.0050	0.249	('60', '15', '70', '79', '73')
19	16.3	80.2	0.3	0.0103	0.245	('10', '14', '99', '60', '28')
20	12.6	52.8	0.2	0.0060	0.220	('60', '99', '20', '49', '28')
21	24.0	223.3	0.7	0.0078	0.215	('38', '35', '36', '99', '26')
22	32.5	278.0	0.8	0.0015	0.170	('55', '37', '60', '50', '99')
23	34.9	386.0	1.3	0.0051	0.276	('60', '73', '70', '65', '63')
24	3.3	2.5	0.0	-0.0005	-0.006	('10', '32', '49', '20', '14')
25	45.7	717.1	2.5	-0.0044	0.248	('13', '29', '49', '28', '44')
26	28.0	143.7	0.5	0.0010	0.203	('38', '15', '35', '50', '99')
27	34.0	137.0	0.5	0.0038	0.242	('35', '36', '37', '38', '30')
28	29.0	220.4	0.8	0.0048	0.292	('60', '63', '49', '28', '80')
29	49.1	555.2	1.9	0.0007	0.251	('36', '38', '73', '35', '99')
30	15.1	123.4	0.4	-0.0038	0.252	('56', '31', '30', '99', '37')
31	2.0	5.1	0.0	-0.0004	-	('34', '37', '38')
32	2.0	2.2	0.0	-0.0239	-	('10', '28')
33	38.4	920.6	4.2	-0.0255	0.278	('13', '29', '44', '49', '16')
34	13.3	56.3	0.3	-0.0056	0.182	('16', '20', '22', '25', '28')
35	2.0	16.0	0.1	-0.0223	-	('37',)
36	2.0	0.8	0.0	-0.0310	-	('35', '36')
37	1.9	8.9	0.0	0.0090	-	('59', '50')
38	1.0	0.0	0.0	0.0138	-	('60', '67')
39	2.0	8.9	0.0	-0.0159	-	('61', '63')
40	23.1	38.0	0.2	-0.0003	0.132	('60', '79', '73', '26', '28')
41	2.0	4.1	0.0	-0.0180	-	('13', '30')
42	13.0	101.5	0.5	-0.0072	0.212	('30', '31', '56', '60')
43	2.0	1.8	0.0	0.0060	-	('34',)
44	2.0	22.8	0.1	0.0152	-	('34',)
Sum	1,084.2	13,845.8	50.8			
PANEL B: Image Industries 73 (50)						
0	26.4	564.9	2.0	0.0054	0.282	('45', '37', '38', '73', '48')
1	41.2	477.6	1.7	0.0058	0.321	('42', '37', '40', '49', '99')
2	9.6	27.5	0.1	0.0086	0.208	('34', '38', '37', '99', '36')

Table 3: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
3	10.3	44.8	0.2	0.0030	0.211	('56', '23', '51', '38', '48')
4	7.4	36.3	0.1	0.0041	0.140	('38', '36', '56', '73', '48')
5	10.3	126.3	0.5	-0.0017	0.250	('28', '60', '99', '73', '67')
6	38.3	724.1	2.4	0.0102	0.237	('36', '35', '73', '50', '99')
7	12.7	138.8	0.5	0.0035	0.190	('38', '35', '28', '73', '36')
8	27.4	251.1	1.0	0.0060	0.288	('36', '38', '73', '50', '35')
9	20.1	65.1	0.3	0.0009	0.144	('60', '36', '28', '38', '35')
10	15.3	170.9	0.6	0.0072	0.337	('60', '36', '73', '35', '65')
11	23.3	247.2	0.9	0.0062	0.170	('28', '60', '59', '99', '56')
12	11.2	42.8	0.2	0.0112	0.177	('38', '35', '36', '73', '99')
13	31.9	1,181.7	4.2	0.0032	0.178	('20', '28', '38', '59', '99')
14	8.5	29.0	0.1	0.0054	0.304	('35', '36', '30', '37', '38')
15	23.1	276.3	1.2	0.0020	0.311	('20', '25', '57', '28', '50')
16	10.4	41.6	0.2	0.0047	0.241	('20', '37', '35', '33', '36')
17	15.6	244.8	1.0	0.0022	0.186	('28', '25', '38', '49', '57')
18	8.6	54.3	0.2	0.0090	0.162	('73', '99', '28', '20', '48')
19	11.5	31.0	0.1	0.0081	0.186	('36', '28', '73', '99', '13')
20	9.5	135.3	0.4	0.0084	0.272	('49', '99', '36', '73', '46')
21	19.4	83.7	0.3	0.0065	0.202	('36', '38', '35', '48', '73')
22	24.5	229.8	0.8	-0.0017	0.227	('58', '99', '20', '36', '28')
23	14.3	420.3	1.6	-0.0055	0.318	('60', '62', '73', '38', '99')
24	21.0	105.2	0.4	0.0053	0.176	('99', '36', '60', '28', '38')
25	9.8	64.8	0.2	0.0115	0.372	('60', '70', '79', '99', '24')
26	26.2	448.0	1.5	0.0054	0.155	('28', '58', '99', '38', '20')
27	14.3	511.7	1.7	0.0048	0.390	('60', '62', '67', '63', '73')
28	11.1	42.6	0.2	0.0058	0.257	('36', '73', '35', '50', '99')
29	12.9	113.5	0.4	-0.0054	0.211	('60', '78', '10', '23', '49')
30	12.5	20.5	0.1	-0.0024	0.178	('60', '99', '70', '67', '24')
31	9.2	29.6	0.1	0.0018	0.208	('60', '22', '24', '63', '50')
32	9.9	61.5	0.2	0.0102	0.251	('10', '14', '28', '38', '73')
33	5.7	17.6	0.1	-0.0099	0.180	('56', '60', '53', '48', '28')
34	10.6	60.3	0.2	0.0043	0.252	('38', '36', '33', '50', '26')
35	13.0	149.0	0.5	0.0062	0.248	('38', '35', '56', '23', '50')
36	17.7	221.6	0.7	-0.0021	0.289	('35', '38', '37', '50', '73')
37	27.1	273.4	0.8	0.0053	0.164	('55', '37', '50', '79', '99')
38	7.4	159.3	0.5	-0.0071	0.353	('49', '28', '10', '13', '15')
39	29.4	129.3	0.4	-0.0039	0.219	('60', '15', '63', '70', '65')
40	8.9	115.5	0.4	0.0096	0.449	('35', '36', '38', '33', '50')
41	8.8	18.0	0.1	0.0063	0.218	('38', '35', '60', '20', '36')
42	30.9	779.7	2.4	0.0064	0.182	('59', '53', '54', '55', '35')
43	23.0	403.6	1.6	-0.0111	0.191	('20', '59', '99', '53', '54')
44	17.8	147.6	0.6	0.0042	0.170	('20', '55', '37', '73', '36')
45	30.3	491.0	1.6	0.0041	0.335	('13', '29', '49', '44', '99')
46	32.5	322.4	1.3	0.0040	0.294	('15', '13', '70', '29', '25')
47	9.8	88.1	0.3	0.0076	0.227	('35', '36', '37', '73', '48')
48	10.9	157.9	0.5	-0.0009	0.300	('35', '63', '60', '36', '62')
49	17.3	317.0	1.0	0.0045	0.247	('36', '35', '73', '99', '60')
50	21.7	152.6	0.6	-0.0037	0.251	('60', '36', '35', '73', '63')
51	12.9	33.9	0.1	0.0104	0.236	('60', '38', '36', '35', '99')
52	13.2	268.9	1.0	0.0087	0.111	('60', '36', '38', '35', '99')
53	15.6	476.6	2.1	0.0033	0.293	('58', '99', '53', '54', '59')
54	15.1	142.6	0.5	-0.0083	0.184	('56', '31', '30', '99', '37')
55	5.7	30.7	0.1	-0.0005	0.149	('30', '31', '34', '56', '99')
56	20.9	417.6	1.6	-0.0060	0.292	('60', '13', '29', '79', '73')
57	17.3	170.9	0.7	0.0048	0.300	('70', '79', '60', '73', '63')
58	2.0	16.0	0.1	-0.0223	-	('37',)
59	2.0	0.8	0.0	-0.0310	-	('35', '36')
60	4.0	20.8	0.1	-0.0059	-	('15', '28', '30', '99')
61	45.9	405.7	1.9	-0.0050	0.157	('35', '36', '38', '73', '48')
62	1.9	8.9	0.0	0.0090	-	('59', '50')
63	1.0	0.0	0.0	0.0138	-	('60', '67')
64	2.0	8.9	0.0	-0.0159	-	('61', '63')

Table 3: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
65	5.2	21.3	0.1	0.0084	0.501	('60', '73', '59')
66	2.0	4.1	0.0	-0.0180	-	('13', '30')
67	29.3	1,155.5	5.3	-0.0038	0.183	('48', '73', '35', '36', '34')
68	8.0	24.2	0.1	-0.0018	0.054	('58', '63', '99')
69	13.0	101.5	0.5	-0.0072	0.212	('30', '31', '56', '60')
70	2.0	1.8	0.0	0.0060	-	('34',)
71	2.0	22.8	0.1	0.0152	-	('34',)
72	7.8	15.6	0.1	0.0046	0.051	('60', '37', '49', '73', '79')
Sum	1,109.7	14,395.7	53.3			

Table 4: Peer Groups' - Comparison

We present a comparison of different industry classification techniques by the average number of industries (Number of Industries) and the average monthly number of classified stocks (Number of Stocks). We compare two classifications based on Image Industries that have 45 (Panel A) and 73 classes (Panel B), with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 and 48 classes, NAICS industry classification with 20 classes, three digits NAICS codes, four and six digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 and 50 classes (Text 25 and Text 50). The sample covers stocks classified with Image Industries from 2014 to 2021.

Industry Classification Name	Number of Industries	Number of Stocks
PANEL A: Image Industries 45 (25) - comparison		
Image Industries	34.9	1,011.0
Industry_MG	20.0	1,011.0
Fama-French 30 Industries	29.8	1,009.0
NAICS Industries	19.0	1,010.2
4-digit GICS	24.6	1,007.4
Icode 25 Industries	25.3	990.0
PANEL B: Image Industries 73 (50) - comparison		
Image Industries	60.9	1,011.0
2-digit SIC	61.4	1,011.0
Fama-French 48 Industries	46.8	1,009.0
3-digit NAICS	77.7	1,010.2
6-digit GICS	66.3	1,007.4
Text 50 Industries	45.3	990.0

Table 5: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Market Information

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable *vble* represents 10 ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry *ind* excluding firm i at month t. In Panel A we compare classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25). In Panel B, we compare classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes (Text 50). In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2014-2021. We calculate regression for industries that have at least five members. Table C1 in the Appendix shows details of ratios calculation.

Industry Classification Name	MARKET to BOOK	PRICE to BOOK	MONTHLY RET	FORECASTED MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q
PANEL A: Image Industries 45 (25) - comparison										
Image Industries	25.2	18.9	23.9	2.4	31.7	23.8	10.0	31.5	36.0	22.8
Industry_MG	26.1	21.7	26.7	2.1	31.7	25.7	10.1	31.8	36.5	24.5
Fama-French 30 Industries	25.4	20.5	26.8	2.1	31.1	25.9	8.3	30.5	37.6	23.8
NAICS Industries	26.2	21.6	26.3	2.2	31.5	24.8	9.3	30.7	36.7	25.0
4-digit GICS	27.5	20.7	28.0	2.1	32.2	26.2	10.5	31.2	39.8	25.7
Text 25 Industries	23.7	21.9	26.0	2.7	31.6	26.7	9.1	29.4	34.2	21.3
PANEL B: Image Industries 73 (50) - comparison										
Image Industries	26.5	18.7	23.1	2.6	31.5	25.1	10.5	32.1	34.6	23.6
2-digit SIC	28.1	22.4	26.7	2.2	32.5	24.5	11.2	30.9	35.0	26.3
Fama-French 48 Industries	27.4	21.3	26.2	2.3	31.3	25.9	10.5	31.0	35.2	25.6
3-digit NAICS	27.3	22.0	25.7	2.1	31.3	25.1	10.8	30.2	35.3	24.6
6-digit GICS	27.9	21.7	27.1	2.2	32.0	27.1	10.1	30.7	38.5	25.6
Text 50 Industries	24.5	20.2	26.3	2.9	32.3	25.7	10.7	30.1	32.9	22.0

Table 6: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Accountancy Information

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable *vble* represents 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIVIDEND PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSET; 14) RNOA; 15) ROE; and 16) ASSETS to SALES for each firm i at quarter t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry *ind* excluding firm i at month t. In Panel A we compare classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25). In Panel B, we compare classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes (Text 50). In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2014 to 2021. We calculate regression for industries that have at least five members. Table C1 in the Appendix shows details of ratios calculation.

Industry Classification Name	TOTAL ASSETS	NET SALES	DIV PAYOUT	PROFIT MARGIN	DEBT to EQUITY	SALES GROWTH	R&D EXPENSE to SALES	R&D GROWTH	SG&A to # EMPLOYEES	SG&A GROWTH	FORECASTED EPS	EPS GROWTH	DEBT to ASSETS	RNOA	ROE	ASSET to SALES
PANEL A: Image Industries 45 (25) - comparison																
Image Industries	41.9	43.4	6.0	27.9	16.6	39.1	22.2	22.9	23.5	33.4	42.5	15.8	26.1	16.7	15.6	22.9
Industry_MG	40.9	41.3	4.5	24.4	17.3	37.8	19.6	16.8	27.0	30.7	43.7	17.2	31.1	18.0	13.4	26.6
Fama-French 30 Industries	39.9	40.4	4.4	24.3	16.0	35.4	21.2	17.8	25.1	31.6	42.1	16.9	29.1	17.4	12.9	25.4
NAICS Industries	43.7	41.6	4.9	24.6	14.2	36.4	19.2	9.4	27.9	30.3	41.7	17.2	31.6	16.7	15.4	25.2
4-digit GICS	42.8	41.3	5.7	24.6	14.5	36.7	19.3	17.4	25.7	30.1	42.5	17.3	31.1	15.1	13.6	26.8
Text 25 Industries	37.7	38.2	5.8	26.7	16.5	35.9	22.1	18.5	26.0	29.9	40.5	18.0	30.0	17.7	16.1	27.6
PANEL B: Image Industries 73 (50) - comparison																
Image Industries	39.2	45.1	7.6	28.5	17.6	38.3	27.4	26.3	21.9	33.6	41.4	16.9	24.9	19.8	16.7	23.4
2-digit SIC	37.2	39.6	7.1	25.8	14.3	35.2	20.1	18.9	25.9	29.5	43.2	17.3	28.1	18.6	15.9	25.6
Fama-French 48 Industries	38.9	40.1	6.6	26.0	15.5	34.1	22.9	19.7	24.8	29.4	41.5	17.2	28.1	17.8	15.2	25.8
3-digit NAICS	35.3	40.1	6.0	27.6	15.9	34.0	19.8	21.4	26.1	28.2	42.8	17.8	27.8	19.1	18.3	25.6
6-digit GICS	40.7	39.7	6.3	26.1	15.7	35.6	20.5	19.0	27.7	28.8	41.8	17.7	28.2	18.3	16.4	26.5
Text 50 Industries	39.4	37.1	6.1	28.7	17.6	37.7	25.1	20.8	25.3	31.1	40.6	18.4	28.5	18.0	17.3	26.7

Table 7: Pair Trading Strategy on Growth

The table showcases the outcomes of pair trading strategies rooted in firm growth metrics, specifically SALES GROWTH and EPS GROWTH. Firms and their peers, identified through similarities in images (IFS), texts (HP), and shared analysts (KR), are sorted into quintiles based on growth observed in the preceding month. Investment positions are then held for one month, favoring firms with high growth (long) over those with low growth (short). The analysis, spanning 2014 to 2021, considers firms with at least five peers. Text similarity data is derived from Hoberg and Phillips (2016) (HP), while common analyst data is from Kaustia and Rantala (2021) (KR). This table elucidates the performance differential between the top and bottom growth quintiles, reflecting the efficacy of growth-based pair trading. Table C1 in the Appendix shows details of ratios calculation.

	SALES GROWTH			**EPS GROWTH**		
	IIC	HP	KR	IIC	HP	KR
Annual Ret.	0.260	0.219	0.192	0.136	0.114	0.039
Annual Std. Dev.	0.083	0.084	0.086	0.080	0.070	0.053
Max DrownDown	0.026	0.040	0.039	0.081	0.058	0.070
Max 1m. Loss	0.022	0.040	0.039	0.073	0.057	0.043
Sharpe Ratio	3.126	2.627	2.237	1.691	1.629	0.733
Calmar Ratio	10.107	5.511	4.886	1.671	1.960	0.555

Table 8: Industry diversification benefits - comparison with the Image Industry

We demonstrate the average performance of portfolios built with different industry classification techniques formed with 1) image industries with 45 classes (25 classes with at least five firms in 2013) (Image Industries), 2) Moskowitz and Grinblatt (1999) (Industry_MG), 3) Fama-French industries with 30 classes, 4) NAICS industry classification with 20 classes, 5) four digits GICS codes, and 6) transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25 Industries). To create a portfolio, we randomly select one stock from each industry monthly and measure the portfolio performance. We perform 500 trials per industry and report the average performance statistics for three different stock weighing techniques: 1) equal weights (Panel A); 2) market weights (Panel B); 3) mean-variance optimized portfolios to maximize Sharpe Ratio (Panel C); and 4) conditional value-at-risk (CVaR) optimized portfolios to minimize risk (Panel D). To optimize portfolios (Panels C and D), we use three years of historical monthly returns before the monthly optimization date. We use the same stock universe for every industry classification consisting of stocks with the Image Industry 45 classification. In each row, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers the years 2014 to 2021. We use sectors with at least five stocks.

Industry Classification Name	Image Industries	Industry_MG	Fama-French 30 Industries	NAICS Industries	4-digit GICS	Text 25 Industries
PANEL A: Equally weighted portfolios						
Annual Ret.	0.046	0.036	0.043	0.044	0.048	0.047
Annual Std. Dev.	0.231	0.243	0.237	0.265	0.214	0.241
Max DrownDown	0.485	0.521	0.501	0.550	0.444	0.497
Sharpe Ratio	0.199	0.149	0.185	0.167	0.226	0.197
Calmar Ratio	0.104	0.077	0.092	0.089	0.118	0.103
PANEL B: Value weighted portfolios						
Annual Ret.	0.106	0.086	0.090	0.092	0.100	0.107
Annual Std. Dev.	0.205	0.220	0.193	0.223	0.198	0.216
Max DrownDown	0.332	0.395	0.328	0.390	0.297	0.326
Sharpe Ratio	0.559	0.466	0.514	0.459	0.598	0.586
Calmar Ratio	0.382	0.284	0.323	0.295	0.426	0.409
PANEL C: Max. Sharpe Ratio portfolios						
Annual Ret.	0.128	0.117	0.138	0.125	0.116	0.120
Annual Std. Dev.	0.227	0.215	0.221	0.224	0.229	0.225
Max DrownDown	0.330	0.323	0.311	0.349	0.335	0.335
Sharpe Ratio	0.578	0.564	0.643	0.570	0.529	0.549
Calmar Ratio	0.441	0.428	0.497	0.403	0.420	0.415
PANEL D: Min. CVaR Ratio portfolios						
Annual Ret.	0.132	0.109	0.119	0.119	0.108	0.107
Annual Std. Dev.	0.213	0.203	0.212	0.205	0.214	0.211
Max DrownDown	0.313	0.321	0.317	0.312	0.327	0.330
Sharpe Ratio	0.634	0.571	0.587	0.585	0.532	0.526
Calmar Ratio	0.487	0.409	0.445	0.447	0.403	0.381

Table 9: Industry momentum & short-term reversal

The table compares Sharpe ratios of momentum and short-term reversal industry portfolios built with different industry classification techniques formed with 1) image industries with 45 classes (Image Industries), 2) Moskowitz and Grinblatt (1999) (Industry_MG), 3) Fama-French industries with 30 classes, 4) NAICS industry classification with 20 classes, 5) four digits GICS codes, and 6) transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25 Industries). We build an industry momentum strategy in the same way as Moskowitz and Grinblatt (1999) by investing long (short) for six, nine, or twelve months in three industries with the highest (lowest) six months momentum. The short-term reversal strategy is built by investing long in three industries for six, nine, or twelve months with the lowest one-month returns. The returns of industries are calculated as market-weighted (Panel A) or equally weighted (Panel B). The table has 12 rows representing different strategies. The first phrase demonstrates the name of the strategy (Momentum or Reversal), and the second phrase denotes the investment horizon (six, nine, or twelve months). In each row, the highest Sharpe ratio is marked as dark green, the second highest as light green, and the third highest as beige. The sample covers the years 2014 to 2021. We use sectors with at least five stocks.

	Image Industries	Industry_MG	Fama-French 30 Industries	NAICS Industries	4-digit GICS	Text 25 Industries
PANEL A: Market weighted industry returns						
Momentum 6m	0.606	0.189	0.786	0.060	0.398	0.436
Momentum 9m	0.683	0.117	0.590	-0.109	0.215	0.183
Momentum 12m	0.766	0.104	0.495	-0.192	0.119	0.029
Reversal 6m	1.178	0.941	0.814	0.990	1.144	0.911
Reversal 9m	1.178	0.974	0.823	0.941	1.153	0.923
Reversal 12m	1.206	0.995	0.836	0.967	1.190	0.964
PANEL B: Equally weighted industry returns						
Momentum 6m	0.650	0.088	0.283	0.282	0.235	0.093
Momentum 9m	0.524	-0.075	0.140	0.108	0.029	-0.059
Momentum 12m	0.383	-0.200	-0.059	0.072	-0.051	-0.147
Reversal 6m	0.498	0.567	0.577	0.596	0.668	0.616
Reversal 9m	0.500	0.583	0.567	0.624	0.694	0.611
Reversal 12m	0.533	0.607	0.611	0.664	0.766	0.709

Table 10: Industry momentum - "Random" Industry Portfolios

We compare Sharpe ratios of 'random' industry portfolios with Image Industries with 45 classes (25 classes with at least five firms in 2013). We build an industry momentum strategy in the same way as Moskowitz and Grinblatt (1999) by investing long (short) for six, nine, or twelve months in three industries with the highest (lowest) six months momentum. To construct a "random" industry, we replace every true stock in Image Industries with another stock with almost the same six-month return. We find similar stocks by ranking 6-month returns and picking a replacement stock that differs by "n" ranks. The table has three columns, where each column represents strategy with different "n" shifts, e.g., column Shift_-1 states replacing an actual stock with another whose 6-month momentum rank is one point lower. Column Shift_0 represents the original strategy. The returns of industries are calculated as market-weighted (Panel A) or equally weighted (Panel B). In each row, the highest Sharpe ratio is marked as dark green. The sample covers 2014-2021, where we use the first months to create stock 6-month momentum. We use sectors with at least five stocks.

	Shift_-1	Shift_0 or IIC	Shift_1
PANEL A: Market weighted industry returns			
Momentum 6m 6m	0.086	0.606	-0.042
Momentum 6m 9m	0.181	0.682	-0.074
Momentum 6m 12m	0.126	0.766	-0.127
PANEL B: Equally weighted industry returns			
Momentum 6m 6m	0.026	0.650	0.117
Momentum 6m 9m	0.192	0.524	0.115
Momentum 6m 12m	0.121	0.383	0.102

Table 11: Industry Dynamics

The table compares the statistics of industry dynamics. In Panel A we compare Image Industries with 45 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25). In Panel B, we compare Image Industries with 73 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes (Text 50). We estimate the industry dynamics as the frequency of reclassification of a given company among industries. For example, the dynamics for a company that is classified into one industry for the entire period is zero. The sample covers all stocks classified with Image Industries 45 (Panel A), Image Industries 73 (Panel B), and additionally with all other techniques from the years 2014-2021. We additionally require that each firm has classification in year t-1 to ensure that the captured changes in the industry are exclusively due to re-classifications. The reported statistics are the proportion of firms with a new classification to all observations. They are based on annual firm observations.

Industry Classification Name	Dynamics
PANEL A: Image Industries 45 (25) - comparison	
Image Industries	0.167
Industry_MG	0.020
Fama-French 30 Industries	0.021
NAICS Industries	0.033
4-digit GICS	0.000
Text 25 Industries	0.084
PANEL B: Image Industries 73 (50) - comparison	
Image Industries	0.217
2-digit SIC	0.027
Fama-French 48 Industries	0.025
3-digit NAICS	0.053
6-digit GICS	0.000
Text 50 Industries	0.092

Table 12: Vertical and Horizontal Link Data

This table presents the vertical and horizontal links of industry firm similarities for 45 and 73 industries (IFS 45 and IFS 73). The vertical link is captured by the number of firms that are suppliers and customers in the same sector within the corresponding IFS. Suppliers and customers are identified from the BEA tables. The horizontal link measures the percentage of firms in the IFS that overlap with those in HP industry classification. IFS 45 and IFS 73 consist of 45 and 73 industry classes, respectively. IFS 73 is more granular, with an average of fewer than 20 stocks per class. During the formation period, 23 of its sectors contain fewer than five stocks. To address this sparsity, IFS 50 is constructed by retaining only those IFS 73 classes with at least five stocks. Similarly, IFS 25 is derived from IFS 45 by selecting the 25 classes that include at least five firms each. See Section 9 for detailed estimation methodology.

	Vertical link (with BEA customers and suppliers data)				Horizontal link with HP industry classification			
IFS	IFS 45		IFS 73		IFS 25 vs. HP 25	IFS 45 vs. HP25	IFS 50 vs. HP 50	IFS73 vs. HP50
	Suppliers	Customers	Suppliers	Customers				
1	8.67	12.03	16.61	11.17	6.04	6.44	2.85	2.93
2	23.06	27.37	21.04	17.62	6.29	6.66	3.53	3.75
3	8.78	15.71	16.91	6.55	13.60	13.29	2.78	3.00
4	12.49	26.78	16.27	27.69	13.81	14.52	2.91	2.74
5	28.01	23.3	15.03	12.46	9.40	10.74	2.30	2.65
6	15.02	29.23	12.99	9.18	8.77	7.66	8.30	17.57
7	31.44	24.31	42.66	22.54	9.96	9.62	3.67	3.47
8	14.66	16.93	16.63	8.5	16.46	14.24	11.49	12.19
9	16.43	13.64	23.08	13.15	7.16	7.21	3.60	3.19
10	10.97	19.26	16.77	8.33	12.74	13.78	3.16	3.24
11	24.42	29.67	13.82	11.17	6.39	5.53	2.99	3.45
12	8.9	17.7	20.29	16.04	6.25	6.57	3.81	3.80
13	4.56	6.98	14.29	7.15	6.06	6.07	2.88	3.38
14	7.87	8.15	27.25	13.5	7.31	6.63	3.05	3.56
15	26.83	12.52	13.27	3.12	5.20	4.81	3.83	3.36
16	8.01	12.7	19.98	19.68	7.56	7.44	2.84	3.00
17	7.38	9.66	13.64	12.62	8.28	7.44	3.61	3.42
18	11.8	17.57	13.85	11.81	8.69	9.46	2.94	2.91
19	13.64	17.27	15.32	5.45	5.58	6.29	4.22	3.83
20	3.67	3.88	13.95	2.93	5.19	5.38	3.53	3.31
21	5.25	9.29	13.83	4.42	7.50	7.76	2.79	3.26
22	13.66	55.02	18.79	8.92	6.29	6.68	2.85	2.55
23	13.01	48.98	23.27	12.58	17.39	17.48	3.31	2.95
24	9.85	14.41	13.78	8.03	6.18	6.48	2.91	2.70
25	0.77	0.73	15.25	7.94	4.78	5.45	2.51	2.61
26	14.7	18.42	23.48	5.49		13.38	2.62	2.69
27	11.29	20.97	20.68	9.48		6.58	2.51	2.60
28	16.92	20.04	14.97	12.34		4.42	4.15	3.90
29	8.92	20.9	14.37	12.11		11.01	2.83	2.78
30	27.91	18.37	13.87	10.43		5.24	2.97	3.00
31	7.76	21.86	13.36	5.44		12.05	2.68	2.41
32	2.49	1.78	12.12	13.38		6.29	3.85	3.35
33	10.06	8.06	14.68	6.76		7.00	2.59	2.86
34	1.62	8.91	11.37	13.01		5.98	4.06	3.61
35	1.17	1.46	15.15	9.23		13.29	2.92	3.03
36	1.32	0.74	20.19	16.08		9.46	3.54	4.00
37	13.37	5.83	19.64	11.61		5.78	3.17	3.38
38	0.1	0.74	21.26	45.97		3.90	3.18	2.86
39	0.37	2.57	12.9	2.93		4.25	15.33	15.10
40	12.99	0.11	18.14	17.13		5.00	4.08	4.56
41	9.18	5.84	14.87	4.87		13.25	4.23	3.88
42	8.79	2.24	14.39	5.77		4.33	4.24	4.56
43			21.84	12.92		11.96	3.36	3.03
44			40.9	36.29		7.99	3.77	3.55
45			15.63	6.68			2.44	2.86
46			19.4	14.01			5.59	6.16
47			19.19	13.22			3.26	3.66
48			13.57	10.39			3.72	3.26
49			13.62	7.69			3.61	3.22
50			38.9	7.41			3.27	2.94
51			16.97	12.63				2.82
52			16.2	6.91				3.30
53			11.65	5.67				3.07
54			15.68	24.25				2.80
55			16.21	18.69				2.76
56			11.59	5.65				2.96
57			14.98	8.27				3.44
58			11.91	3.36				3.00
59			15.00	3.38				4.00
60			13.00	2.86				3.00
61			14.73	8.34				3.59
62			12.23	5.58				3.00
63			11.7	4.21				2.00
64			15.14	5.83				2.00
65			14.37	7.26				2.67
66			18.45	8.11				3.00
67			17.18	9.11				3.56
68			11.58	2.22				2.25
69			16.81	4.95				2.46
71			18.79	5.88				5.00
72			12.46	9.24				3.00
73			17.07	7.74				2.40
Avg	11.38	15.05	17.04	10.60	8.52	8.29	3.81	3.56
Min	0.10	0.11	11.37	2.22	4.78	3.90	2.30	2.00
Max	31.44	55.02	42.66	45.97	17.39	17.48	15.33	15.10
Median	9.96	14.03	15.20	8.71	7.31	6.84	3.27	3.05

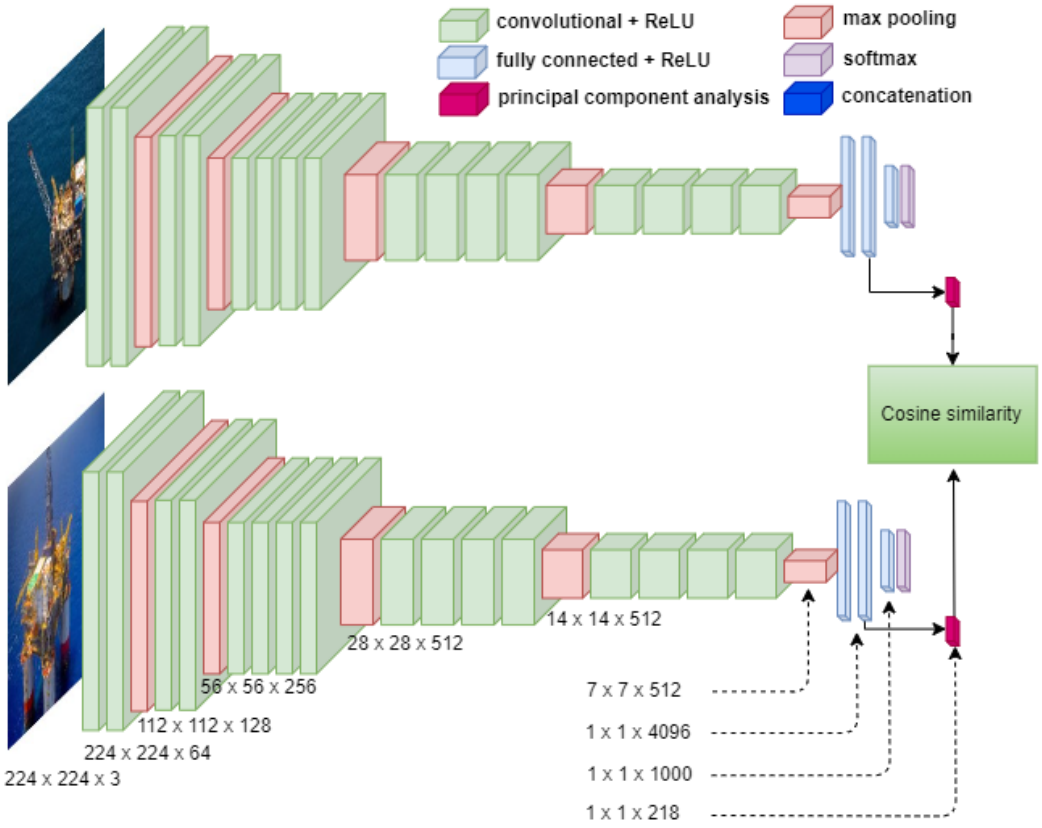


Figure 2: Image similarity

Note: The figure presents the architecture of image comparison. In the first step, images are standardized to dimensions 224x224x3, where the first dimension shows the height, the second the width, and the third the colors. Second, to detect objects on photos, we use a convolutional neural network VGG-19 that is 19 layers deep (Simonyan & Zisserman, 2014). The networks consist of convolutional layers that create a feature map, pooling layers that scale down the information generated by the convolutional layer, and fully connected layers that compile the data extracted by previous layers to form the final output. VGG-19 pretrained with more than 1m photos is designed to classify objects into 1,000 categories. We use the numerical representation of identified objects from the last but one fully connected layer with dimensions 1x1x4096. Third, we apply Principal Complement Analysis (PCA) to reduce this dimension and represent at least 70% of the variation. The reduced vector has a dimension of 1x1x218. Finally, feature vectors are the input to define cosine similarity between two photos.

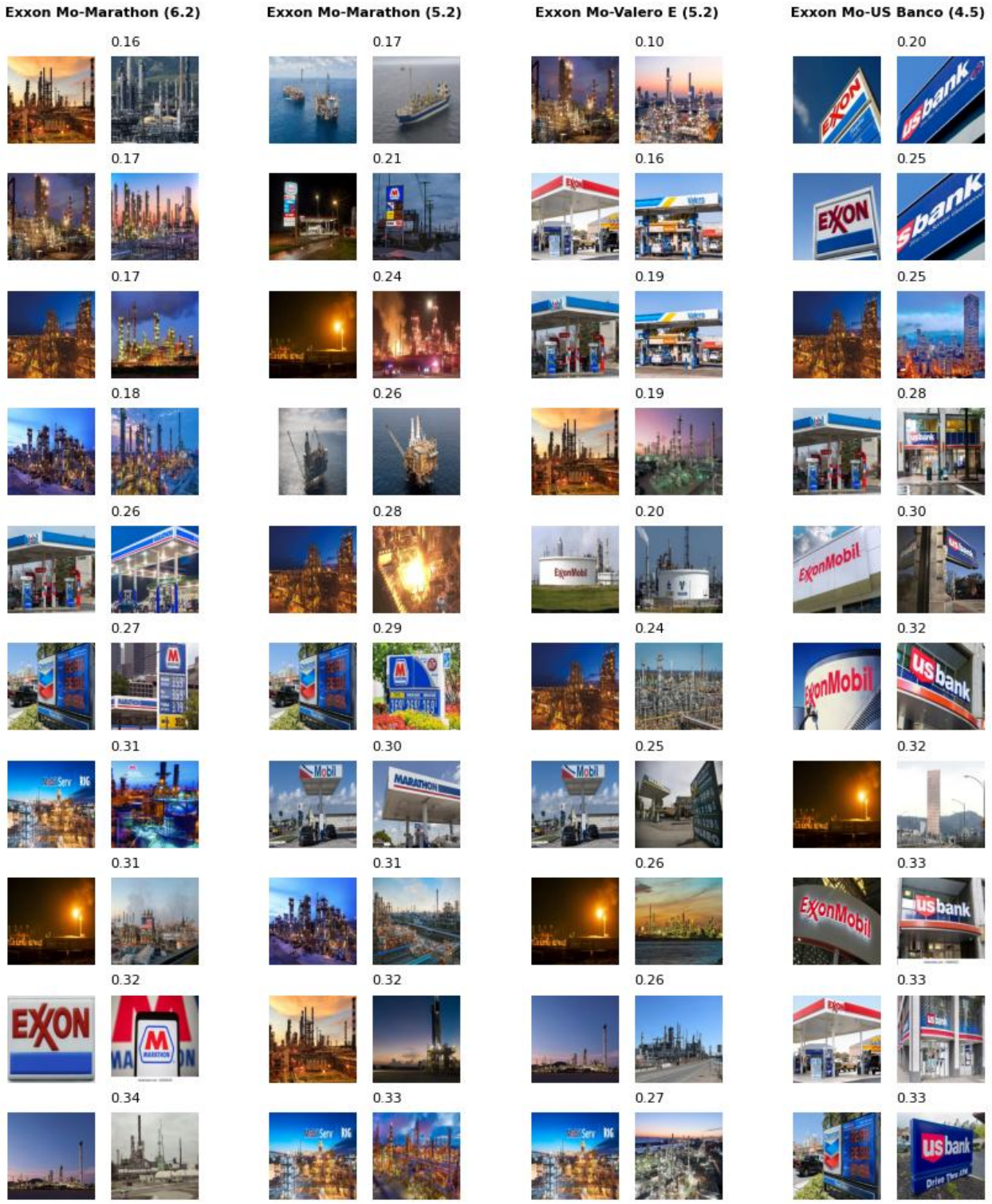


Figure 3: Photos representing four peers of Exxon Mobil Corp with the highest similarity score

This figure presents photos related with Exxon Mobil Corp and its four peers with the highest similarity score - from the left: Marathon Petroleum Corp, Marathon Oil Corp, Valero Energy Corp, and US Bancorp. Each set of photos is titled with peers names and similarity score in the brackets. Each pair of photos demonstrate the cosine similarity distance measure.

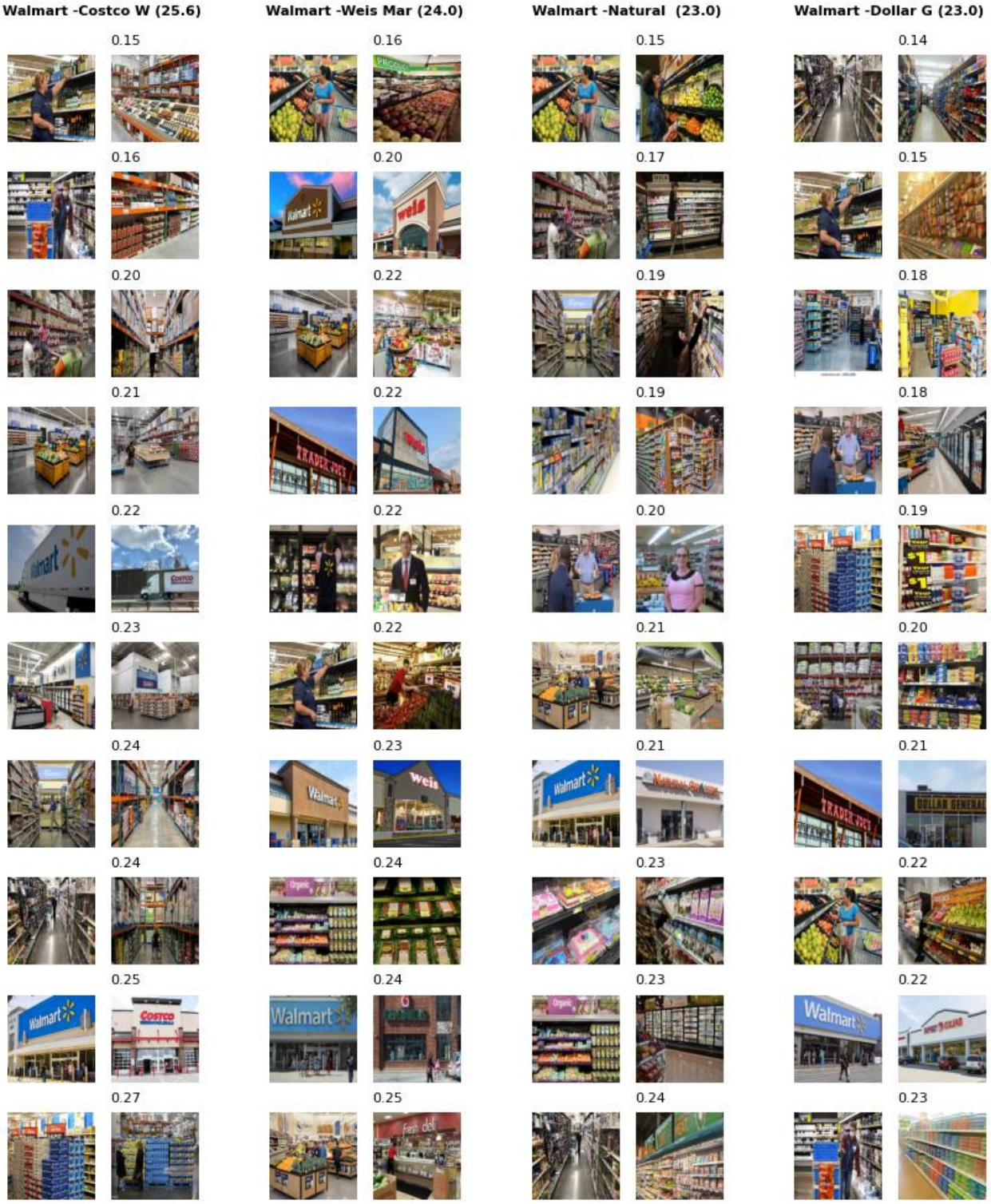


Figure 4: Photos representing four peers of Walmart Inc with the highest similarity score

This figure presents photos related with Walmart Inc and its four peers with the highest similarity score - from the left: Costco Wholesale Corp, Weis Markets, Natural Grocers By Vitamin Cottage Inc, and Dollar General Corp. Each set of photos is titled with peers names and similarity score in the brackets. Each pair of photos demonstrate the cosine similarity distance measure.

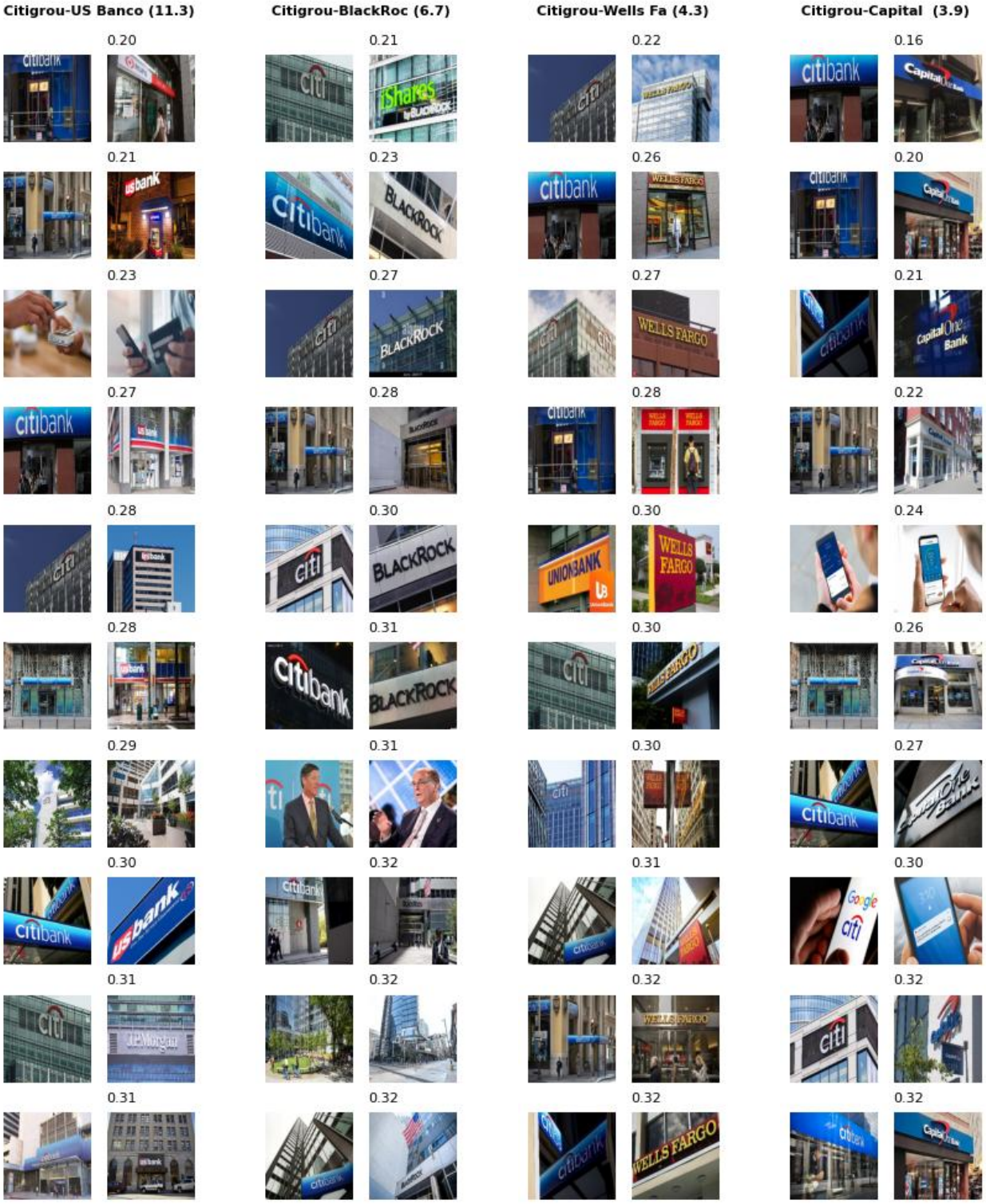


Figure 5: Photos representing four peers of Citigroup Inc with the highest similarity score

This figure presents photos related with Citigroup Inc and its four peers with the highest similarity score - from the left: US Bancorp, BlackRock, Wells Fargo & Co, and Capital One Financial Corp. Each set of photos is titled with peers names and similarity score in the brackets. Each pair of photos demonstrate the cosine similarity distance measure.

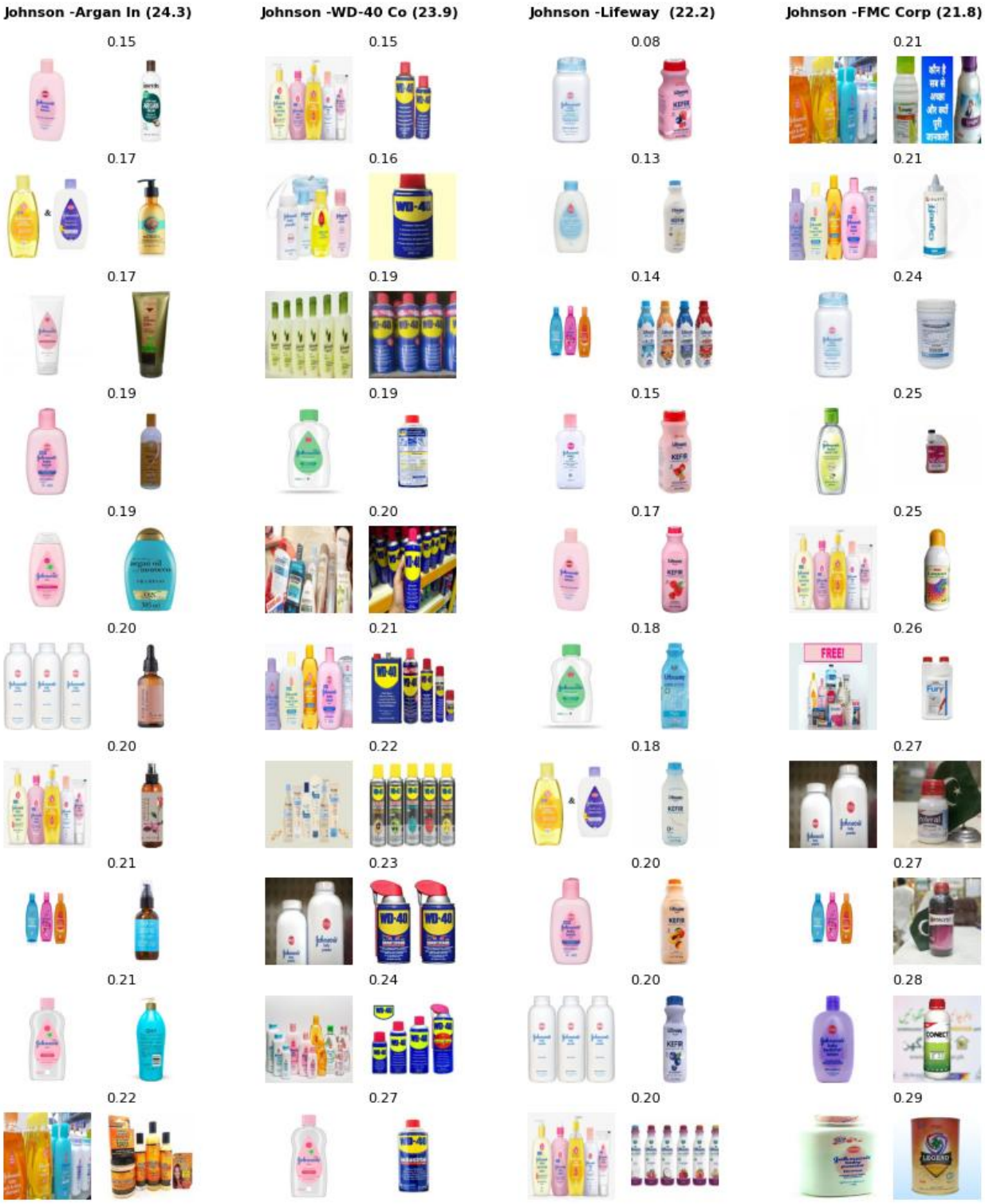


Figure 6: Photos representing four peers of Johnson & Johnson with the highest similarity score

This figure presents photos related with Johnson & Johnson and its four peers with the highest similarity score - from the left: Argan Inc, WD-40 Co, Lifeway Foods Inc, and FMC Corp. Each set of photos is titled with peers names and similarity score in the brackets. Each pair of photos demonstrate the cosine similarity distance measure.

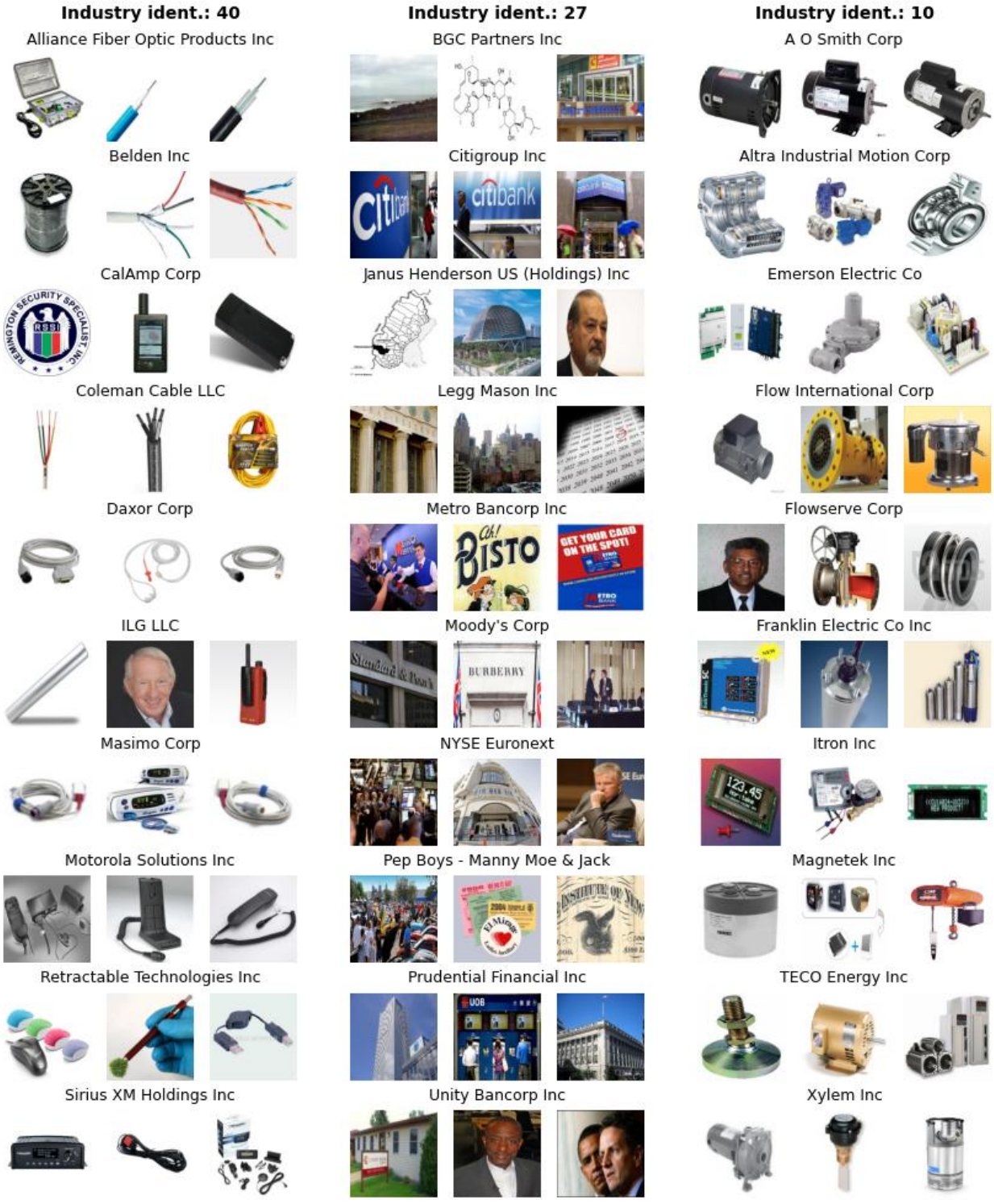


Figure 7: Photos representing industries with highest inter-industry correlations (1)

This figure presents photos from Image Industries 73 presenting three industries with high inter-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos are downloaded from Google from the years 2009 to 2013 and are used to form industries. We excluded industries represented with less than 10 stocks.

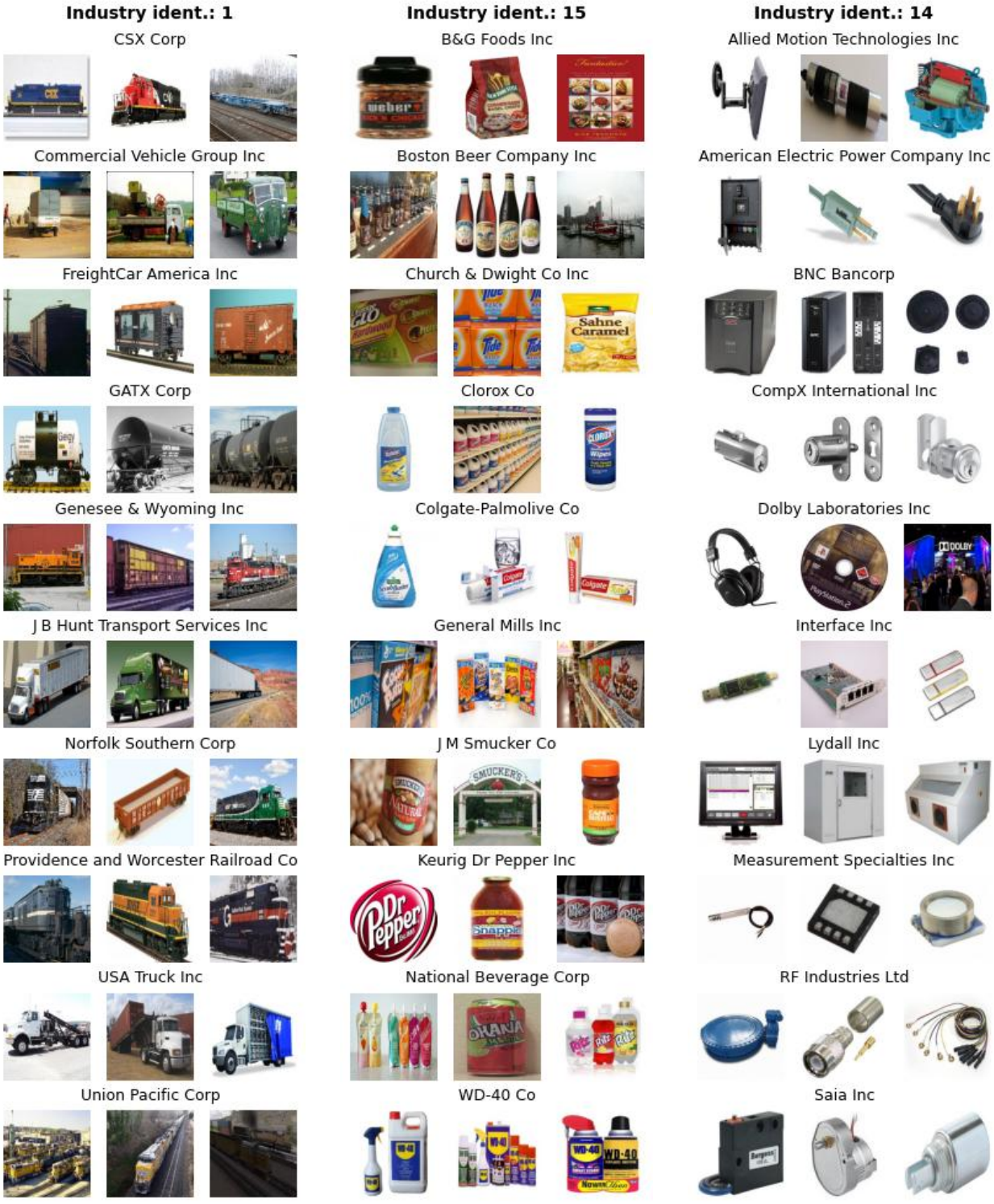


Figure 8: Photos representing industries with highest inter-industry correlations (1)

This figure presents photos from Image Industries 73 presenting three industries with high inter-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos are downloaded from Google from the years 2009 to 2013 and are used to form industries. We excluded industries represented with less than 10 stocks.

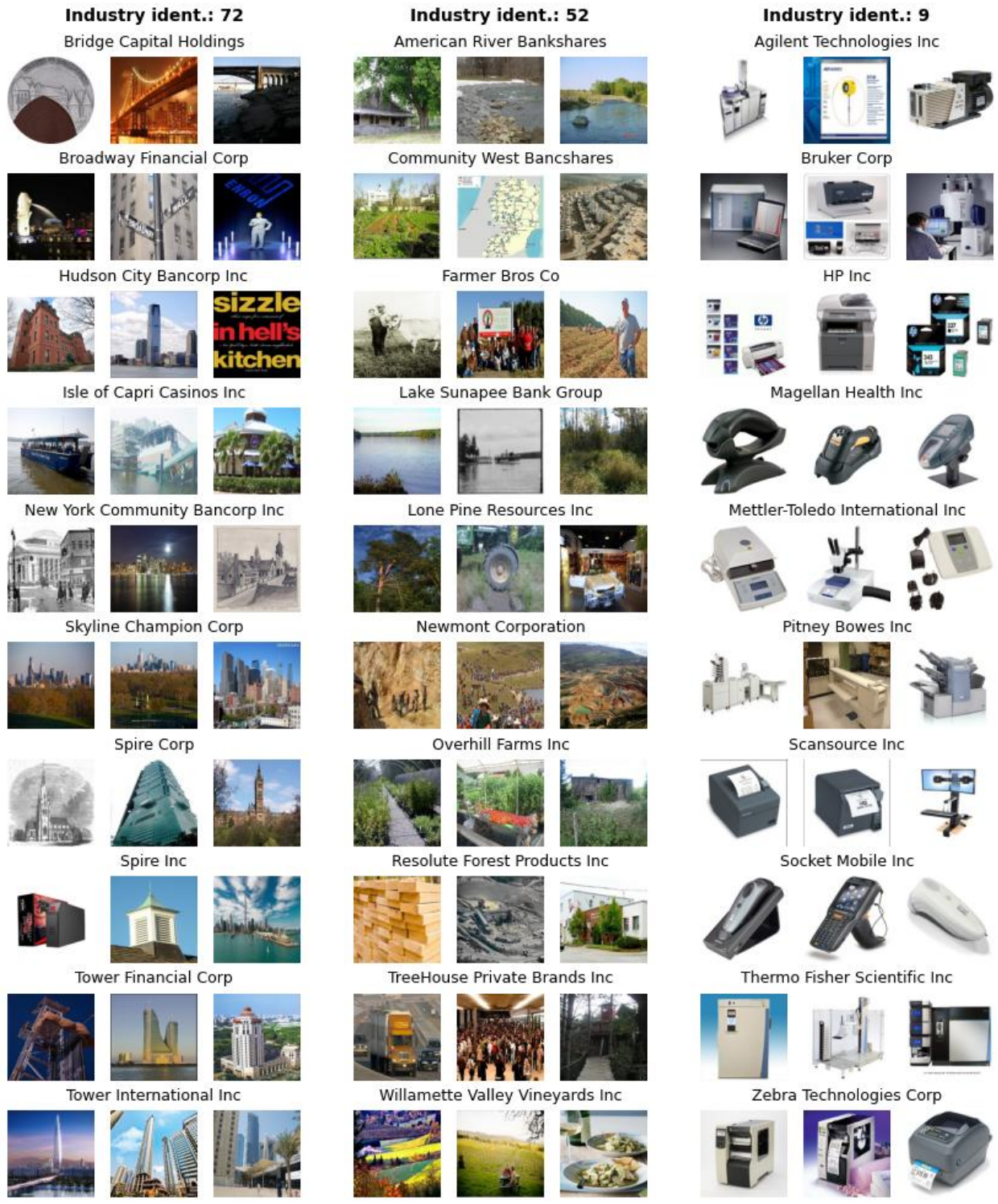


Figure 9: Photos representing industries with the lowest inter-industry correlations (1)

This figure presents photos from Image Industries 73 presenting three industries with low inter-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos are downloaded from Google from the years 2009 to 2013 and are used to form industries. We excluded industries represented with less than 10 stocks.

Figure 10: Photos representing industries with the lowest inter-industry correlations (2)

This figure presents photos from Image Industries 73 presenting three industries with low inter-industry correlations. Each industry is represented with 10 randomly selected stocks and their three randomly selected photos. Photos are downloaded from Google from the years 2009 to 2013 and are used to form industries. We excluded industries represented with less than 10 stocks.

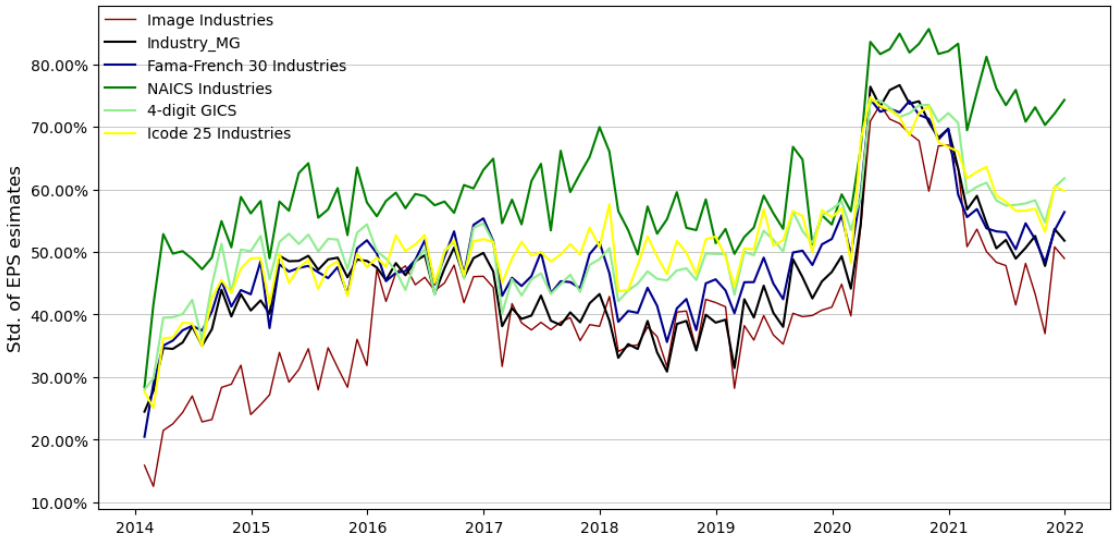


Figure 11: Investor's agreement

The chart illustrates the level of investor agreement regarding the behavior of companies within the same industries. We adopt the measure of agreement proposed by Diether, Malloy, and Scherbina (2002). Investor opinion diversity regarding a company i at time t is computed as the analysts' forecasts in the month prior to the fiscal period end date divided by the absolute value of the mean forecast. Subsequently, investor opinion diversity regarding all companies within a given industry j at time t is expressed as the standard deviation of the estimated ratios, while investor opinion divergence regarding industry classification at time t is the mean of these standard deviations. Forecast data is sourced from I/B/E/S detail files. The dataset covers the period from 2014 to 2021 and pertains to industries comprising at least 5 companies.

Appendix A: Methodology and Data

A.1 Methodology: Image-Based Industry Classification

Our study introduces a novel approach to industry classification using Image-Based Firm Similarity (IFS), which leverages visual data from company images to group firms into industries. This methodology unfolds in three key stages: calculating firm similarities, clustering firms into industries, and dynamically updating classifications as new data becomes available.

Step 1: Calculating Firm Similarities

The foundation of our method lies in measuring how visually similar two firms are based on their image sets. Each firm is represented by a collection of images (e.g., product photos, logos, or operational visuals) sourced from publicly available platforms like Google Images. To quantify similarity:

Pairwise Image Comparison: For each pair of images a_i (from Firm A) and b_j (from Firm B), we calculate a similarity score:

$$\text{similarity}(a_i, b_j) = 1 - \text{distance}(a_i, b_j),$$

where $\text{distance}(a_i, b_j)$ measures how different the two images are using advanced image recognition techniques.

Aggregating Firm-Level Similarities: The overall similarity between two firms, $S(A, B)$, is computed by summing the similarity scores across all image pairs:

$$S(A, B) = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \text{similarity}(a_i, b_j),$$

where n_A and n_B are the number of images representing Firms A and B, respectively. This step ensures that firms with visually similar products or operations have higher similarity scores.

Step 2: Clustering Firms into Industries

Once pairwise similarities are calculated for all firm pairs, we use clustering algorithms to group firms into industries. The clustering process works as follows:

Similarity Matrix: We organize the similarity scores $S(A, B)$ into a matrix where each cell represents the similarity between two firms.

Clustering Algorithm: A clustering algorithm processes this matrix to form industry groups. Firms with higher mutual similarity scores are more likely to be grouped together:

$$C = \text{cluster}(S(A, B)),$$

where C represents the set of industry clusters. This step translates complex pairwise relationships into coherent industry classifications.

Step 3: Dynamic Updates

Industries evolve as companies change their operations or introduce new products. To reflect these changes:

- We periodically update the image sets for each firm and recalculate similarities.
- The updated similarity matrix is used to reassign firms to clusters, ensuring that classifications remain relevant over time.

This dynamic updating process captures real-time shifts in firm activities and market conditions.

Our methodology leverages visual data to capture both tangible (e.g., product designs) and intangible (e.g., branding elements) aspects of firms. By focusing on images—a medium humans naturally process quickly and intuitively—IFS provides a dynamic and nuanced view of industry relationships that complements traditional classification methods based on text or numerical data.

A.2 Data

We define the stock universe as all firms from NYSE, AMEX, and NASDAQ that we have obtained from CRSP. We take shares with codes 10 and 11 and collect all photos representing firms directly from the Google search engine via the Python API. The API allows the collection of 100 photos for a single query. The set of images retrieved from Google for each query depends on the ranking that Google defines. The algorithm ranks photos based on the reliability of their upload sources, with images displayed first by Google assessed to originate from the most credible sources. Given that Google indexes thousands of images for each listed company, the first 100 photos each year come only from highly evaluated sources; the majority originate from national newspapers, company websites, or Wikipedia.[26,27]

[26]The order of images displayed by Google depends on two ranks. The first is a universal PageRank that estimates the ranks of the most reliable content sources and displays them accordingly. The second is VisualRank, which combines PageRank information with visual similarity analysis to rank images based on their relevance and visual features (e.g., color, texture, shape). This ensures that the top-ranked images are both credible and visually representative of the query (Brin & Page, 1998; Jing & Baluja, 2008). By leveraging these algorithms, our methodology benefits from a curated set of high-quality, market-driven visuals that reflect firms' operations and public perceptions.

[27]Google ranks the credibility of image sources using several key factors, including E-A-T (Expertise, Authoritativeness, Trustworthiness), domain authority, and the quality of content. Websites with high-quality backlinks, user engagement metrics,

To achieve a time-varying classification, we download photos for each year separately. We use the following search phrase: '{Company Common Name} products after: {year}-01-01 before: {year}-12-31.' Refinitiv retrieves data for firm names with the field 'Company Common Name.' Google provides the history of indexed photos from 2008. We retrieve photos for each year from 2009 to 2021. Finally, we collect close to four million photos and group our data sample into 3-year rolling windows. We create groups to achieve a more comprehensive photo representation for each period. The sample covers, on average, 2,250 stocks per year.

Photos downloaded from Google need a thorough cleaning. Pictures do not always convey meaningful information about a firm's business activity (e.g., faces, logotypes, or landscapes). We perform a cleaning procedure that eliminates most photos, but significantly improves photo quality. We demonstrate a detailed procedure of photo cleaning in Appendix **??**.

The timestamp on each image is a crucial component that allows us to cluster businesses and develop a dynamic measure of our categorization. Our photographs contain time stamps based on the upload time.

In addition, we build a collection of 26 financial ratios at the stock level based on Kaustia and Rantala (2021) and group them into ratios using market information and ratios based only on accounting data. Table C1 in the Appendix **??** demonstrates a detailed calculation methodology for each ratio.[28]

We download the SIC and NAICS codes from CRSP, GICS from Compustat, similarity scores with industries based on text classification from Hoberg and Phillips (2016), and similarity scores on common analysts from Kaustia and Rantala (2021) to compare our Image Industries with other classification techniques. Classifications from CRSP are time-varying, making them more comparable to our technique.[29]

Our total data sample covers the years 2009 to 2021. We divide the 13 years of data into the 5-year training sample (2009 to 2013) and eight-years of the out-of-sample testing sample (2014 to 2021). During the eight-year out-of-sample period (2014-2021), we use rolling three-year windows of historical data to make biennial predictions about future industry classifications. This approach allows us to test the predictive power of our model, update industry compositions to reflect changing company offerings, and avoid look-ahead bias by using only past data to forecast future classifications.

We build image-based industries in four significant steps. First, we start with the feature extraction procedure from each image. This step is important in defining the numerical representation of objects from photographs. The process is described in Section A.3. Second, we move to identify yearly updated firms'

and secure connections (HTTPS) are often deemed more credible. Additionally, the expertise of authors, content freshness, and transparency about the site's purpose contribute to how Google assesses the reliability of a source. Overall, a strong online reputation and consistent delivery of valuable content are crucial for ranking well.

[28] Our set of 26 indicators includes the 19 indicators used by Kaustia and Rantala (2021) plus seven additional ratios. A detailed breakdown is discussed in Section 4.2.

[29] Compustat delivers static classifications. CRSP gives dynamic but only for SIC and NAICS. Therefore, we have dynamic data from CRSP (SIC, NAICS) and static data from Compustat (GICS). See Table 11.

similarities based on their image representation from all 3-year rolling window periods. Our procedure of peer definition is demonstrated in Section A.7. Third, we use data from the image definition training sample to create IFS. The established industries are constant for the entire out-of-sample testing period.[30] Lastly, as companies dynamically change their product offerings depending on market conditions, we update the composition of each industry on a biennial basis. We describe industries' definition and their updates in Section A.13.

A.3 Extracting Content from Images

To classify firms into industries based on visual data, we first extract meaningful features from company-related images. This step is essential for identifying patterns and objects that represent a firm's business activities, enabling systematic comparisons between firms.

A.4 Overview of Image Processing

We use the VGG19 model, a convolutional neural network (CNN) pre-trained on the ImageNet dataset, to process and analyze images. VGG19 is widely recognized for its ability to identify visual features, ranging from simple shapes to complex objects. It transforms each image into a numerical representation—a feature vector—that captures its key visual characteristics.

A.5 Steps in Feature Extraction

- **Resizing Images:** All images are resized to 224×224 pixels to match VGG19's input requirements. This ensures consistency across the dataset.
- **Layer-by-Layer Analysis:** The VGG19 model processes each image through multiple layers:
  - *Convolutional Layers:* Detect features such as edges, textures, and patterns. These layers are represented mathematically as:

$$F_{l+1} = \text{ReLU}\left(W_l * F_l + b_l\right) \tag{6}$$

where F_{l+1} is the feature map obtained from layer $l+1$, W_l and b_l are the weights and biases of layer l, F_l is the input feature map to layer l, and ReLU denotes the Rectified Linear Unit activation function that introduces non-linearity, enhancing the network's learning capability.

[30] Given that access to the photos' history is limited and the testing period lasts eight years, changes in the design of the industries would not significantly affect the study results. Nevertheless, our methodology allows us to update the composition of industries as the period of photo availability lengthens.

- *Pooling Layers:* Simplify the data by focusing on the most important features.
- *Fully Connected Layers:* After convolutional and pooling operations, the image data is flattened into a single feature vector with 4,096 dimensions.

- **Feature Vector Creation:** The resulting feature vector serves as a compact numerical summary of the image's visual content, making it possible to compare images systematically.

A.6 Why Feature Extraction Matters?

The feature vectors generated by VGG19 allow us to quantify similarities between images from different firms. For example, a soda company like Coca-Cola might have feature vectors representing bottles, cans, or vending machines. A car manufacturer like Ford might have feature vectors representing vehicles or assembly lines.

These numerical representations form the foundation for comparing firms based on their visual data.

This process transforms raw images into structured data that can be analyzed mathematically. By focusing on visual features, we capture aspects of firms that traditional text-based methods might overlook, such as product designs or operational visuals.

A.7 Identification of Firms' Similarity

To classify firms into industries based on their images, we need to measure how similar their visual representations are. This process builds on the feature extraction step discussed in Section A.3 and refines it to identify meaningful similarities between firms.

A.7.1 Cosine Similarity: Measuring Image Similarity

At the heart of our methodology is cosine similarity, a mathematical tool that measures how similar two high-dimensional vectors are by comparing their orientation in space. In our case, these vectors represent the visual features of images extracted using the VGG19 model. Cosine similarity is calculated as:

$$\text{Cosine Similarity}(a, b) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}},$$

where a and b are feature vectors representing two images (e.g., one from Firm A and one from Firm B). These vectors encode various visual attributes such as color, texture, shape, and patterns. Cosine similarity ranges from 1 (completely similar) to 0 (completely dissimilar), providing a precise measure of how visually alike two images are.

A.7.2 Dimensionality Reduction with PCA

The feature vectors extracted by VGG19 are 4,096 dimensions long, which can make comparisons computationally intensive and less effective due to the *curse of dimensionality*. This phenomenon occurs when high-dimensional spaces become sparse, making it harder to distinguish meaningful patterns.

To address this issue, we use Principal Component Analysis (PCA) to reduce the dimensionality of these vectors while retaining at least 70% of the original variability in the data. PCA transforms the data into a smaller set of uncorrelated variables (principal components), which capture the most important features of the images. Mathematically:

$$Z = XP,$$

where X is the matrix of original feature vectors, P is the matrix of principal components, and Z is the transformed data in reduced dimensions.

This step ensures that our similarity calculations focus on the most significant visual features while eliminating noise and redundancy.

A.7.3 Refining Similarities

Once we reduce dimensionality, we calculate cosine similarity between all pairs of images from two firms. However, simply summing all similarities between image pairs can dilute the impact of highly similar images due to noise from less relevant matches. To address this:

- **Thresholding:** We apply a threshold to filter out image pairs with cosine distances above 0.4 (i.e., less similar pairs). This ensures that only meaningful matches contribute to the overall similarity score.
- **Optimal Pairing:** We use an optimization technique called the Linear Sum Assignment Problem (LSAP) to pair images from two firms in a way that minimizes the total distance between them. This approach prioritizes the strongest visual correlations between firms.

The result is a refined similarity score that accurately captures how visually similar two firms are based on their image sets.

This process allows us to quantify firm similarity in a way that reflects their visual characteristics—such as product designs or operational visuals—rather than relying solely on textual or numerical data. By focusing on meaningful visual connections, we create a robust foundation for grouping firms into industries that better reflect their economic roles and market positioning.

A.8 Linear Sum Assignment Problem (LSAP)

The Linear Sum Assignment Problem (LSAP) is a combinatorial optimization technique used to determine the optimal one-to-one matching between two sets—in our context, the images from Firm A and Firm B—to minimize the overall distance between matched pairs. A smaller distance implies higher visual similarity between firms.

Formally, the LSAP can be defined as follows:

$$\min_{x_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Distance}_{ij} \cdot x_{ij} \tag{7}$$

subject to:

$$\sum_{j=1}^{m} x_{ij} \leq 1, \quad \forall i \tag{8}$$

$$\sum_{i=1}^{n} x_{ij} \leq 1, \quad \forall j \tag{9}$$

$$x_{ij} \in \{0, 1\} \tag{10}$$

Here, Distance_{ij} represents the visual dissimilarity (typically $1-$ cosine similarity) between image i from Firm A and image j from Firm B. The variable x_{ij} is a binary indicator that equals 1 if image i is paired with image j, and 0 otherwise. These constraints ensure that each image can be matched at most once, forming an optimal assignment.[31]

By solving this optimization problem, we identify the optimal matches that minimize the overall distance or maximize the overall similarity, ensuring that the most similar pairs are prioritized.

Upon solving the LSAP, we obtain $\hat{x}_{i,j}$, which represents the estimated optimal matches between images from firm A and firm B. These optimal matches are then used to calculate the total similarity score between the firms, considering only the pairs that meet our similarity threshold.

A.9 Expected vs. Excess Similarity

The total similarity between firms A and B includes both systematic and idiosyncratic components. Expected similarity is driven by common photos that frequently appear in sets of many companies, such as iPhones,

[31] $x_{i,j}$ is a binary decision variable that indicates whether image i of firm A is paired with image j of firm B in the optimal assignment. It takes the value of 1 if image i is paired with image j, and 0 otherwise. $x_{i,j}$ is not directly computed but is determined by the optimization algorithm. The algorithm aims to minimize the total *Distance* by deciding which images should be paired ($x_{i,j} = 1$) and which should not ($x_{i,j} = 0$). The optimization problem seeks to find $x_{i,j}$ that minimizes the total *Distance*, subject to constraints ensuring that each image from firm A is paired with at most one image from firm B and vice versa. This approach accounts for cases where the number of photos per company A and B are not the same, allowing for some images to remain unmatched.

humans, or buildings. These common elements cause every two sets of photos to exhibit a certain degree of baseline similarity. In contrast, excess similarity represents the portion of similarity that is not explained by the expected similarity. It is driven by photos that are not typical across all companies but are characteristic of specific industries, as well as by an unusually high number of common photos within certain sets. For example, two cell phone producers may have an unnaturally high number of images featuring mobile phones, or two construction companies might be visualized by a very high number of building images. This atypical relatedness demonstrates excess similarity, which is the type of similarity we aim to identify.

The degree of expected similarity depends on the number of photos in set A and set B. The higher the maximum number of pairs that can be defined by those two sets the higher the expected similarity. To disentangle these, we decompose the overall similarity into two parts:

- **Expected Similarity:** Arises from common visual elements across firms, such as generic objects or shared backgrounds (e.g., smartphones, buildings), which generate a baseline similarity.
- **Excess Similarity:** Represents the deviation from this expected baseline. It captures unique, industry-specific visual cues that are not explainable by chance or common imagery.

A.10 Weighted Least Squares (WLS) Regression

To account for heteroskedasticity in the similarity data, we employ a Weighted Least Squares (WLS) regression model. Unlike Ordinary Least Squares (OLS), WLS assigns weights to each observation inversely proportional to its variance. This ensures that observations with higher variability exert less influence on the estimation of model parameters.

The WLS model is given by:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma), \tag{11}$$

where Y is the $n \times 1$ vector of total similarity scores, X is the $n \times p$ design matrix (in our case, $p = 1$, since we use a single explanatory variable representing the number of image pairs), β is the regression coefficient, and Σ is an $n \times n$ diagonal matrix containing observation-specific weights.

After estimating the regression coefficient $\widehat{\beta}$, the expected similarity scores are computed as:

$$\widehat{Y} = X\widehat{\beta}. \tag{12}$$

A.11 Significant Excess Similarities

To determine whether excess similarity is statistically significant, we conduct a one-sided hypothesis test at the 99% confidence level:

$$H_0 : Y - \widehat{Y} = 0.$$

We compute the upper bound of the 99% confidence interval for $\widehat{Y}$ and define excess similarity as:

$$\text{Excess Similarity} = Y - \widehat{Y}_{99\%}. \tag{13}$$

Only firm pairs with similarity scores above this threshold are classified as exhibiting statistically robust visual similarity.

Statistically Significant Similarity Over Time

To ensure temporal stability and mitigate short-term noise, we impose two robustness conditions:

Type I Error Control (False Positives): A firm pair is considered visually similar in year t only if statistically significant excess similarity is observed in at least two out of the three years: $t-2$, $t-1$, and t. This rule filters out transient or spurious connections.

Type II Error Control (False Negatives): Once a firm pair is classified as similar in year t, we extend that similarity into year $t+1$. This update rule improves temporal consistency by preventing the classification from overreacting to brief perturbations.

A.12 Similarity Matrix Construction

Based on the filtered excess similarity scores, we construct a sparse similarity matrix for each year. In practice, only about 1% of all possible firm pairs meet the threshold for statistical significance.

Example: Suppose we assess the similarity between Firms A and B using images from 2009–2013. If significant similarity is observed in 2011 and 2013 (but not in 2012), we classify the pair as peers in 2013. To ensure stability, we extend this classification into 2014. This process relies on a five-year rolling window to define firm-level peer groups, which are then used for out-of-sample testing.

This WLS-based approach balances sensitivity (capturing meaningful visual relationships) with stability (avoiding noise), resulting in a robust classification of firm similarities over time. By isolating statistically

significant excess similarity, our method focuses on economically informative visual connections that complement traditional classification techniques.

We introduce a biannual updating process to accommodate the dynamic nature of firm similarities in response to new images. This process relies on a similarity matrix predominantly composed of unseen photos, given that it is based on a rolling 3-year observation period, with approximately 66% of the images being new in each update. Starting with an individual firm, we identify the cluster to which it and its peers are previously assigned as defined by the new similarity matrix. We calculate the mean similarity for each potential cluster from the previous period and select the one with the highest mean similarity for reassignment. This procedure is replicated for all companies, followed by a reclassification step. Here, we assess whether the current cluster assignment remains optimal for each firm by calculating the mean distance to each cluster. If a more suitable cluster is identified, we reclassify the firm accordingly. This iterative process ensures that each firm is associated with the cluster that best reflects its current visual representation.

A.13 Image Industry Definition

The definition of Image Firm Similarities (IFS) is a crucial step in our methodology, where we cluster firms into industries based on their visual similarities. This process involves two key components: scaling similarity scores and applying a clustering algorithm designed for our sparse similarity matrix.

A.13.1 Scaling Similarity Scores

To prepare similarity scores for clustering, we adopt a tailored scaling approach that preserves the granularity of statistically significant relationships while filtering out noise from non-significant ones:

- **Statistical Significance:** Only statistically significant similarities are retained; non-significant firm pairs are assigned a similarity score of zero.
- **Rescaling Significant Scores:** Significant similarity values are rescaled to fall within the range of $[0.5, 1]$, ensuring stronger relationships are emphasized while preserving separation from weaker, insignificant connections.

This scaling approach enhances the clustering process by focusing attention on economically meaningful visual relationships and reducing the impact of noisy data.

A.13.2 Clustering Methodology

Given the sparsity of the similarity matrix, standard clustering techniques often perform poorly. Instead, we adopt an iterative clustering approach inspired by Hoberg and Phillips (2016):

- **Initial Clustering:** Each firm begins in its own individual cluster.
- **Iterative Merging:** At each step, the two clusters with the highest mutual similarity score are merged. This process continues until a pre-specified number of clusters is achieved.
- **Reclassification Step:** After each merge, firms are reassigned to the cluster to which they exhibit the highest average similarity.

This iterative refinement ensures that each firm is ultimately grouped with others exhibiting the most visually coherent features.

A.13.3 Dynamic Classification Update

Industries evolve as firms introduce new products or shift strategic direction. To account for such evolution, we update IFS classifications every two years using a rolling three-year observation window:

- **Similarity Matrix Update:** Roughly 66% of images in each update are newly scraped, allowing the model to capture recent developments in firm operations or branding.
- **Reassignment Process:**
  - For each firm, we compute the mean similarity to each cluster using the updated matrix.
  - Each firm is reassigned to the cluster with which it has the highest average similarity.
- **Reclassification Step:** After reassignment, we verify whether the firm's cluster remains optimal. If not, the firm is moved to a better-fitting cluster.

This dynamic update procedure ensures that industry definitions reflect contemporary firm behaviors and remain robust over time.

By combining tailored similarity scaling, iterative clustering, and dynamic classification updates, our methodology produces visually grounded industry groupings that evolve with the market. These classifications provide a flexible and robust foundation for a variety of asset pricing and strategy applications, including pair trading, diversification, and momentum investing.

Appendix B: Photo Cleaning Procedure

This section demonstrates our photo-cleaning procedure. Our database, upon downloading images from Google, includes 4.2 million images. Many of these are unsuitable for processing in our study. This section elucidates the methodology used to clean up the dataset by removing extraneous images.

The cleaning process begins with identifying and removing images that contain text. Photos with text are predominantly composed of notes, logotypes, or other elements unrelated to the firm's business operations. This step is pivotal in the data cleaning process and accounts for the most considerable reduction in image count. We utilize the Python implementation of the open-source Tesseract 5.0 software for text detection on photos. The algorithm identifies text in 2.2 million images, representing over half of the downloaded images.

Subsequently, we eliminate images dominated by human faces. While some photos featuring human faces can convey certain business characteristics of firms, images overwhelmed by faces can occur across any company, irrespective of its industry. These often include photos of company employees, frequently showcased for marketing purposes. We employ the Open Source Computer Vision Library (OpenCV) (Bradski, 2000) to detect images dominated by faces, specifically utilizing the CascadeClassifier. The algorithm flags 0.2 million images as being dominated by human faces.

Further analysis reveals that photos with a predominantly white background typically depict graphs, and those with extreme size proportions often represent logos still present in the database. We remove images dominated by a white background (where 90% of pixels are white) and those with aspect ratios exceeding 2.5:1 or 1:2.5.

After the comprehensive image-cleaning process, we are left with a sample of 1,626,175 images, averaging between 120,000 and 150,000 per year. This refined dataset forms the basis for our subsequent analysis, ensuring that the visual data accurately reflects the firms' business activities and characteristics.

Appendix C: Tables and Figures

Table C1: Ratios definitions

The table provides definitions of variables used in this study. Panel A presents the ratios that use market data, which we calculate monthly. Panel B shows ratios based only on accountancy information that we update with quarterly frequency. In Panel C, we show some additional variables used as interim steps to calculate some ratios from Panel A or B. We download all financial information from Compustat and market data from CRSP. All lowercase variables in column Definition present a symbol in Compustat.

Variable Name	Full Ratio Name	Updates	Definition
PANEL A: Set of Ratios Using Market Information			
MARKET to BOOK	MARKET to BOOK ratio	monthly	Book assets (atq) minus BOOK_EQUITY plus MARKET_CAP all divided by TOTAL_ASSETS.
PRICE to BOOK	Price-to-Book	monthly	MARKET_CAP divided by total common equity (ceqq).
MONTHLY RET	Monthly return	monthly	Monthly return from CRSP
FORECASTED MONTHLY RET	Forecasted monthly return	monthly	Monthly return one month in the future from CRSP.
MARKETLEVG	Market Leverage	monthly	BOOK_DEBT divided by TOTAL_ASSETS minus BOOK_EQUITY plus MARKET_CAP.
EV to SALES	Enterprise Value-To-Sales	monthly	The sum of MARKET_CAP, long-term debt (lltq), and debt in current liabilities (dlcq) all divided by NET_SALES.
PE	Price-to-Earnings	monthly	MARKET_CAP divided by the sum of the latest four quarter reported net income before extraordinary items (ibq).
BETA	Beta (36 months)	monthly	Beta from the single index model based on monthly returns over the previous 36 months downloaded from WRDS.
MARKET CAP	Market capitalization	monthly	Average price times shares outstanding (SHROUT), prices as average from bid offer (BID) and ask (ASK), all divided by 1,000. Data from CRSP.
TOBIN Q	Tobin's Q	monthly	MARKET_CAP plus long-term debt total (dlttq) plus debt in current liabilities (dlcq) all divided by TOTAL_ASSETS
PANEL B: Set of Ratios Using Only Accountancy Information			
TOTAL ASSETS	Total assets	quarterly	atq
NET SALES	Net Sales	quarterly	Sum of the latest four quarterly reported net sales (saleq) from Compustat.
DIV PAYOUT	Dividend Payout	quarterly	Sum of the latest four quarter reported dividend per share (dvpsxq) divided by sum of the latest four quarter reported earnings per share (epspxq).
PROFIT MARGIN	Profit Margin	quarterly	Sum of the latest four quarter reported net operating income after depreciation (oiadpq) divided by NET_SALES.
DEBT to EQUITY	Leverage	quarterly	Total liabilities (ltq) divided by total stockholders' equity (seqq).
SALES GROWTH	Sales Growth	quarterly	The logarithm of (net sales 1 year in the future divided by current value NET_SALES).
R&D to SALES	Scaled R&D Expense	quarterly	Sum of the latest four quarter reported research and development expense (xrdq) divided by NET_SALES.

Table C1: (continued)

Variable Name	Full Ratio Name	Updates	Definition
R&D GROWTH	R&D Growth	quarterly	The logarithm of (sum of the next four quarter reported research and development expense (xrdq) divided by sum of the latest four quarter reported research and development expense (xrdq)).
SG&A to EMPLOYEES	SG&A Expansion	quarterly	Sum of the latest four quarter reported selling, general and administrative expenses (xsgaq) divided by a number of employees (emp).
SG&A GROWTH	SG&A Growth	quarterly	The logarithm of (sum of the next four quarter reported selling, general and administrative expenses (xsgaq) divided by the sum of the latest four quarter reported selling, general and administrative expenses (xsgaq)).
FORECASTED EPS	Forward EPS (next fiscal year)	quarterly	Earning per share for fiscal year 1 from IBES
EPS GROWTH	Forward to trailing EPS	quarterly	The logarithm of (earning per share for fiscal year 1 from IBES divided by the sum of the latest four quarter reported earnings per share (epspxq)).
DEBT to ASSETS	Book Leverage	quarterly	BOOK_DEBT to TOTAL_ASSETS.
RNOA	Return on Net Operating Assets	quarterly	Sum of the latest four quarterly reported net operating income after depreciation (oiadpq) divided by the sum of property, plant, and equipment (ppentq) and current assets (actq), less current liabilities (lctq).
ROE	Return on Equity	quarterly	Sum of the latest four quarter reported net income before extraordinary items (ibq) divided by COMMON_EQUITY.
ASSETS to SALES	Asset Turnover	quarterly	TOTAL_ASSETS divided by NET_SALES.
PANEL C: Other Variables Used as Input to Calculate Ratios			
COMMON EQUITY	Common Equity	quarterly	ceqq
BOOK EQUITY	Book Equity	quarterly	Stockholders' equity (seqq) minus preferred stock liquidating value (pstkq) plus balance sheet deferred taxes and investment tax credit (txditcq).
BOOK DEBT	Book Debt	quarterly	TOTAL_ASSETS minus BOOK_EQUITY.
PANEL D: Other variables we use in Table E2			
Change in inventory		quarterly	Change in inventory scaled by average total assets
Absolute accruals		quarterly	Annual income before extraordinary items (ib) minus operating cash flows (oancf) divided by average total assets (at); if oancf is missing then set to change in act - change in che - change in lct + change in dlc + change in txp-dp.
Current ratios		quarterly	Current assets divided by current liabilities
Real estate holdings		quarterly	Buildings and capitalized leases divided by gross PP&E
CAPEX and inventories		quarterly	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at).
Growth in CAPEX		quarterly	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at).
Cash productivity		quarterly	Fiscal year end market capitalization plus long term debt (dltt) minus total assets (at) divided by cash and equivalents (che)

Table C1: (continued)

Variable Name	Full Ratio Name	Updates	Definition
Earnings volatility		quarterly	Standard deviaiton for 16 quarters of income before extraordinary items (ibq) divided by average total assets (atq).
Revenue surprise		quarterly	Sales from quarter t minus sales from quarter t-4 (saleq) divided by fiscalquarter-end market capitalization (cshoq $\times$ *prccq*).
Sales to price		quarterly	Annual revenue (sale) divided by fiscal-year-end market capitalization.
Operating profitability		quarterly	Revenue minus cost of goods sold - SG&A expense - interest expense divided by lagged common shareholders' equity.
Cash flow to price		quarterly	Operating cash flows divided by fiscal-year-end market capitalization.
Zero trading days		quarterly	Turnover weighted number of zero trading days for most recent 1 month.
Industry-adj sales to price		quarterly	Industrial-adjusted annual revenue (sale) divided by fiscal-year-end market capitalization.
Growth in CSE		quarterly	Annual percent change in book value of equity (ceq).
Depreciation/PP&E		quarterly	Depreciation divided by PP&E.
Quick ratio		quarterly	(current assets - inventory) / current liabilities.
Cash holdings		quarterly	Cash and cash equivalents divided by average total assets.
Dispersion in forecasted EPS		quarterly	Standard deviation of analyst forecasts in month prior to fiscal period end date divided by the absolute value of the mean forecast; if meanest = 0, then scalar set to 1. Forecast data from I/B/E/S summary files.
Age		quarterly	Number of years since first Compustat coverage.
Earnings to price		quarterly	Annual income before extraordinary items (ib) divided by end of fiscal year market cap.
Growth to LTNOA		quarterly	Growth in long term net operating assets.
Sales to receivables		quarterly	Annual sales divided by accounts receivable.
% change in CAPEX		quarterly	Percent change in capital expenditures from year t-2 to year t.
Leverage		quarterly	Total liabilities (lt) divided by fiscal year end market capitalization.

Table C2: Description and Summary Statistics of Image Industries 45 & 73 in industry formation period

The table presents a static overview of industries at the time of the first industry definition (year 2013) formed with firms' photos. Classification with 45 (73) classes has 25 (50) industries with at least five firms. Our sample covers NYSE, AMEX, and NASDAQ stocks. We report the average number of stocks assigned to each industry (No. of Stocks), the average monthly capitalization of stocks in each industry (Market Cap. (bn USD)), and the share of stocks classified with photos in the whole market capitalization (Avg. % of Market Cap.). Finally, we report the relationship between our industries to two-digit SIC codes by indicating the five most common SIC codes among the companies assigned to each of our industries. Industries marked in bold consist of at least five stocks.

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
PANEL A: Image Industries 45 (25)				
0	30	463.5	2.3	('45', '37', '38', '35', '36')
1	63	406.1	2	('42', '35', '37', '40', '38')
2	2	2.4	0	('30', '36')
3	107	1,254.8	6.3	('36', '38', '48', '73', '35')
4	15	80.1	0.4	('56', '23', '51', '38', '60')
5	97	1,864.1	9.4	('20', '28', '38', '51', '59')
6	21	66.9	0.3	('60', '63', '67', '73', '82')
7	103	574.8	2.9	('36', '35', '50', '73', '38')
8	36	165.7	0.8	('35', '36', '38', '37', '34')
9	35	94.2	0.5	('36', '33', '35', '37', '30')
10	21	35.4	0.2	('25', '57', '24', '37', '52')
11	8	32.8	0.2	('99', '20', '53', '54', '59')
12	17	60.4	0.3	('28', '38', '26', '39', '73')
13	3	15.7	0.1	('73',)
14	6	35.8	0.2	('36', '49', '87', '96')
15	2	1.8	0	('20', '99')
16	2	41.3	0.2	('49', '73')
17	71	855.2	4.3	('58', '20', '53', '54', '99')
18	29	604.6	3	('73', '60', '62', '63', '99')
19	2	3.8	0	('73', '79')
20	4	7.4	0	('17', '73', '80', '99')
21	12	37.7	0.2	('26', '15', '27', '28', '30')
22	30	29.0	0.1	('60', '10', '13', '49', '67')
23	59	298.7	1.5	('60', '70', '15', '49', '63')
24	2	2.0	0	('10', '49')
25	21	31.2	0.2	('60', '20', '28', '49', '73')
26	36	202.0	1	('38', '35', '73', '36', '50')
27	39	180.9	0.9	('55', '37', '36', '50', '73')
28	2	74.7	0.4	('49',)
29	2	2.5	0	('64', '73')
30	1	0.1	0	('99',)
31	2	13.5	0.1	('37', '38')
32	2	1.8	0	('10', '28')
33	39	968.6	4.9	('13', '29', '44', '49', '16')
34	14	49.9	0.3	('16', '20', '22', '21', '25')
35	2	17.2	0.1	('37',)
36	2	0.9	0	('35', '36')
37	2	8.7	0	('50', '59')
38	1	0.0	0	('67',)
39	2	8.7	0	('61', '63')
40	24	33.9	0.2	('60', '79', '73', '26', '28')
41	2	4.4	0	('13', '30')
42	13	78.7	0.4	('30', '31', '56', '60')
43	2	2.2	0	('34',)
44	2	19.5	0.1	('34',)
Sum	987	8,733.6	43.9	
PANEL B: Image Industries 73 (50)				
0	24	328.1	1.6	('45', '37', '38', '35', '47')
1	43	320.9	1.6	('42', '37', '40', '35', '47')
2	2	2.4	0	('30', '36')
3	9	29.4	0.1	('38', '48', '34', '37', '59')
4	8	22.9	0.1	('56', '51', '23', '38')
5	2	0.3	0	('28',)
6	9	22.3	0.1	('73', '35', '38', '36', '37')

Table C2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
7	8	56.1	0.3	('63', '67', '73', '82')
8	76	373.7	1.9	('36', '35', '50', '73', '10')
9	12	145.8	0.7	('35', '38', '36', '50', '80')
10	14	95.9	0.5	('35', '36', '38', '49', '50')
11	22	35.4	0.2	('60', '22', '13', '34', '49')
12	10	16.1	0.1	('73', '79', '26', '28', '35')
13	43	809.9	4.1	('28', '38', '59', '51', '36')
14	16	48.0	0.2	('36', '38', '73', '22', '35')
15	46	918.9	4.6	('20', '28', '54', '26', '29')
16	21	78.3	0.4	('35', '37', '33', '36', '30')
17	21	35.4	0.2	('25', '57', '24', '37', '52')
18	5	5.4	0	('20', '59', '67', '73', '99')
19	11	54.9	0.3	('28', '26', '38', '99')
20	3	15.7	0.1	('73',)
21	10	76.0	0.4	('36', '37', '38', '13', '23')
22	6	35.8	0.2	('36', '49', '87', '96')
23	2	1.8	0	('20', '99')
24	31	22.2	0.1	('36', '38', '99', '50', '73')
25	2	41.3	0.2	('49', '73')
26	37	243.0	1.2	('58', '20', '99', '25', '28')
27	14	427.1	2.1	('60', '62', '73', '55', '63')
28	2	3.8	0	('73', '79')
29	15	44.0	0.2	('10', '60', '49', '12', '61')
30	4	7.4	0	('17', '73', '80', '99')
31	15	41.4	0.2	('60', '63', '28', '36', '78')
32	11	11.4	0.1	('28', '38', '73', '26', '36')
33	6	6.8	0	('60', '28', '49', '56', '70')
34	8	18.9	0.1	('26', '27', '35', '51', '59')
35	7	57.1	0.3	('56', '23', '57', '60')
36	8	7.9	0	('60', '13', '38', '65', '67')
37	7	6.2	0	('20', '16', '21', '22', '60')
38	2	2.0	0	('10', '49')
39	7	2.3	0	('28', '32', '56', '59', '73')
40	15	50.0	0.3	('36', '33', '38', '99', '48')
41	14	34.1	0.2	('38', '35', '36', '50')
42	24	214.2	1.1	('35', '37', '38', '36', '50')
43	3	7.0	0	('73', '99')
44	38	180.6	0.9	('55', '37', '36', '50', '73')
45	2	74.7	0.4	('49',)
46	40	228.9	1.2	('70', '15', '65', '79', '24')
47	13	170.7	0.9	('73', '60', '63', '35', '47')
48	2	2.5	0	('64', '73')
49	5	10.5	0.1	('36', '25', '60', '73')
50	1	0.1	0	('99',)
51	2	13.5	0.1	('37', '38')
52	9	3.0	0	('60', '20', '10', '24')
53	29	639.0	3.2	('53', '54', '58', '59', '55')
54	8	135.3	0.7	('39', '34', '48', '51', '53')
55	2	1.8	0	('10', '28')
56	38	935.0	4.7	('13', '29', '44', '49', '16')
57	7	43.7	0.2	('16', '22', '25', '28', '32')
58	2	17.2	0.1	('37',)
59	2	0.9	0	('35', '36')
60	4	18.8	0.1	('15', '28', '30', '99')
61	49	379.1	1.9	('35', '36', '38', '73', '48')
62	2	8.7	0	('50', '59')
63	1	0.0	0	('67',)
64	2	8.7	0	('61', '63')
65	6	17.7	0.1	('60', '73', '59')
66	2	4.4	0	('13', '30')
67	32	921.2	4.6	('48', '35', '36', '73', '34')
68	8	25.0	0.1	('58', '63', '99')
69	13	78.7	0.4	('30', '31', '56', '60')
70	2	2.2	0	('34',)

Table C2: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	TOP5 SIC codes
71	2	19.5	0.1	('34',)
72	9	14.8	0.1	('60', '37', '49', '73', '79')
Sum	987	8,733.6	43.9	

Table C3: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Market Information for All Stocks

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable *vble* represents 10 ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry *ind* excluding firm i at month t. In Panel A we compare classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes. In Panel B, we compare classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes. In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers all NYSE, AMEX, and NASDAQ stocks from 2014 to 2021. We calculate regression for industries that have at least five members. Table C1 in the Appendix shows details of ratios calculation.

Industry Classification Name	MARKET to BOOK	PRICE to BOOK	MONTHLY RET	FORECASTED MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q
PANEL A: Image Industries 45 (25) - comparison										
Image Industries	25.2	18.9	23.9	2.4	31.7	23.8	10.0	31.5	36.1	22.8
Industry_MG	22.6	19.4	21.3	1.8	26.9	24.0	12.1	28.5	35.7	22.3
Fama-French 30 Industries	22.3	19.2	21.7	1.8	27.2	22.6	10.9	26.7	36.7	22.3
NAICS Industries	23.3	19.4	21.5	1.9	27.9	22.2	10.8	25.8	35.7	23.0
4-digit GICS	26.3	21.5	23.9	1.8	28.1	24.8	10.9	28.0	38.3	25.3
Text 25 Industries	22.5	19.7	22.6	2.1	28.8	23.6	10.7	26.6	33.9	21.4
PANEL B: Image Industries 73 (50) - comparison										
Image Industries	26.5	18.7	23.1	2.6	31.5	25.1	10.5	32.1	34.6	23.6
2-digit SIC	24.0	19.2	21.7	1.8	27.7	22.9	11.5	28.4	35.9	23.8
Fama-French 48 Industries	24.3	19.1	21.8	1.8	27.4	22.0	11.2	27.4	36.7	23.9
3-digit NAICS	24.9	19.7	21.8	2.1	28.7	21.5	11.1	27.9	34.9	23.7
6-digit GICS	26.9	20.9	23.8	1.7	28.6	25.4	10.5	27.7	37.9	25.9
Text 50 Industries	23.3	18.9	22.4	2.3	28.9	23.5	11.8	26.2	33.1	22.3

Table C4: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Accountancy Information for All Stocks

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable *vble* represents 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIVIDEND PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSET; 14) RNOA; 15) ROE; and 16) ASSETS to SALES for each firm i at quarter t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry *ind* excluding firm i at month t. In Panel A we compare classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25). In Panel B, we compare classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes (Text 50). In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers all NYSE, AMEX, and NASDAQ stocks from 2014 to 2021. We calculate regression for industries that have at least five members. Table C1 in the Appendix shows details of ratios calculation.

Industry Classification Name	TOTAL ASSETS	NET SALES	DIV PAYOUT	PROFIT MARGIN	DEBT to EQUITY	SALES GROWTH	R&D EXPENSE to SALES	R&D GROWTH	SG&A to # EMPLOYEES	SG&A GROWTH	FORECASTED EPS	EPS GROWTH	DEBT to ASSETS	RNOA	ROE	ASSET to SALES
PANEL A: Image Industries 45 (25) - comparison																
Image Industries	41.9	43.4	6.0	27.9	16.6	39.1	22.2	22.9	23.5	33.4	42.5	15.8	26.1	16.7	15.6	22.9
Industry_MG	48.5	37.1	4.2	23.9	16.7	28.3	18.7	3.7	27.7	23.3	36.3	14.0	27.7	16.5	14.5	25.0
Fama-French 30 Industries	47.3	43.4	3.3	19.6	14.6	27.8	18.7	8.6	26.6	22.9	39.9	13.9	27.1	13.9	11.6	21.9
NAICS Industries	45.7	40.4	4.1	21.0	12.7	28.2	19.6	5.0	28.0	24.2	35.7	16.0	30.3	15.7	14.5	23.6
4-digit GICS	47.9	42.9	3.9	23.1	13.3	28.6	21.1	7.0	27.6	21.3	38.7	14.9	28.2	13.0	14.0	23.8
Text 25 Industries	41.7	36.9	4.4	20.5	14.9	30.2	18.6	9.2	25.0	25.9	40.5	16.6	27.1	14.7	14.1	22.7
PANEL B: Image Industries 73 (50) - comparison																
Image Industries	39.2	45.1	7.6	28.5	17.6	38.3	27.4	26.3	21.9	33.6	41.4	16.9	24.9	19.8	16.7	23.4
2-digit SIC	45.5	41.6	3.9	21.1	13.4	28.2	18.6	9.4	25.6	22.3	38.2	14.2	27.0	13.8	13.1	23.4
Fama-French 48 Industries	45.8	42.6	3.5	21.7	13.8	28.4	18.9	10.6	25.6	22.9	38.4	14.4	28.1	14.1	11.9	22.0
3-digit NAICS	39.3	39.2	4.4	22.6	14.7	28.5	16.6	10.3	24.9	23.7	39.3	15.1	26.6	14.6	14.0	21.8
6-digit GICS	46.6	43.4	3.9	23.2	12.9	29.5	20.2	13.9	26.0	21.8	40.0	15.6	26.8	13.7	12.6	23.6
Text 50 Industries	40.3	37.7	4.5	21.2	14.6	31.5	18.8	12.2	25.2	26.5	40.2	15.6	26.6	14.8	14.0	22.6

Table C5: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Market Information for Stocks with Prices Not Smaller than USD 5

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable *vble* represents 10 ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry *ind* excluding firm i at month t. In Panel A we compare classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25). In Panel B, we compare classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes (Text 50). In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2014 to 2021 with prices not smaller than USD 5. We calculate regression for industries that have at least five members. Table C1 in the Appendix shows details of ratios calculation.

Industry Classification Name	MARKET to BOOK	PRICE to BOOK	MONTHLY RET	FORECASTED MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q
PANEL A: Image Industries 45 (25) - comparison										
Image Industries	24.8	19.7	25.2	3.1	27.8	24.0	9.9	30.2	32.5	21.8
Industry_MG	24.1	21.8	27.6	2.7	28.7	26.7	10.4	30.7	34.1	23.1
Fama-French 30 Industries	23.8	21.1	27.6	2.5	27.5	27.3	9.0	28.7	34.6	22.4
NAICS Industries	25.5	22.9	27.1	3.0	28.2	22.6	9.7	27.3	34.8	24.2
4-digit GICS	26.7	21.9	29.4	2.5	29.5	26.9	10.7	29.5	37.2	24.6
Text 25 Industries	22.9	22.4	27.2	3.0	28.7	26.1	9.7	27.2	31.1	20.5
PANEL B: Image Industries 73 (50) - comparison										
Image Industries	25.1	18.5	24.3	3.2	26.8	24.6	10.5	30.1	30.5	21.9
2-digit SIC	27.1	22.9	27.8	2.8	29.1	26.8	11.9	29.8	31.8	25.3
Fama-French 48 Industries	26.5	22.2	27.2	2.8	28.0	27.2	11.1	29.4	32.5	24.5
3-digit NAICS	25.9	23.0	26.7	3.4	28.4	25.2	12.0	30.2	32.0	23.2
6-digit GICS	27.3	23.4	28.1	2.5	30.0	28.9	11.4	30.4	35.2	24.3
Text 50 Industries	22.6	20.2	27.4	3.4	29.7	25.1	11.4	27.8	29.8	20.0

Table C6: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Accountancy Information for Stocks with Prices Not Smaller than USD 5

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable *vble* represents 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIVIDEND PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSET; 14) RNOA; 15) ROE; and 16) ASSETS to SALES for each firm i at quarter t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry *ind* excluding firm i at month t. In Panel A we compare classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25). In Panel B, we compare classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes (Text 50). In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2014 to 2021 with prices not smaller than USD 5. We calculate regression for industries that have at least five members. Table C1 in the Appendix shows details of ratios calculation.

Industry Classification Name	TOTAL ASSETS	NET SALES	DIV PAYOUT	PROFIT MARGIN	DEBT to EQUITY	SALES GROWTH	R&D EXPENSE to SALES	R&D GROWTH	SG&A to # EMPLOYEES	SG&A GROWTH	FORECASTED EPS	EPS GROWTH	DEBT to ASSETS	RNOA	ROE	ASSET to SALES
PANEL A: Image Industries 45 (25) - comparison																
Image Industries 25	33.8	36.5	4.5	20.3	15.1	44.2	17.1	28.3	21.8	40.5	45.7	16.0	25.6	16.7	15.2	18.7
Image Industries Unique	32.6	35.2	6.2	22.0	16.8	45.1	20.4	29.0	21.2	41.5	45.3	16.6	25.2	20.2	16.0	20.6
Industry_MG	30.9	32.7	5.3	21.6	16.8	40.3	19.2	21.8	22.2	32.8	43.4	14.7	27.9	20.3	16.4	23.1
Fama-French 30 Industries	32.4	32.7	4.9	23.6	15.2	38.0	21.5	19.4	19.9	33.2	41.4	15.2	26.4	21.1	14.5	22.5
NAICS Industries	39.5	38.0	3.7	17.6	13.4	40.8	16.2	26.8	23.3	33.7	43.6	16.1	29.1	22.2	14.2	18.2
4-digit GICS	35.1	31.9	3.9	18.7	13.4	40.3	18.2	23.3	22.0	32.9	42.0	15.4	27.2	20.2	13.6	23.0
Text 25 Industries	31.0	34.1	5.1	22.7	14.9	40.3	20.9	23.9	23.8	34.6	42.8	17.0	26.3	21.0	15.4	24.1
PANEL B: Image Industries 73 (50) - comparison																
Image Industries 50	33.2	36.2	5.7	22.7	16.3	47.0	21.8	28.8	21.4	40.7	45.8	15.8	24.9	18.0	16.4	19.9
Image Industries 50 Unique	32.8	36.2	6.6	24.5	18.4	45.9	25.2	28.8	20.1	39.9	43.7	16.7	23.9	19.3	17.5	20.9
2-digit SIC	34.7	36.3	5.9	25.7	16.0	39.7	22.5	21.8	21.2	31.0	43.1	15.8	25.2	22.3	15.5	27.6
Fama-French 48 Industries	35.5	34.5	6.1	25.8	15.4	40.4	22.6	23.5	21.1	32.8	41.2	15.1	25.5	23.0	15.4	28.3
3-digit NAICS	32.7	34.2	5.7	24.5	17.5	40.6	25.5	22.9	21.6	33.4	41.5	15.6	25.6	22.2	18.6	23.7
6-digit GICS	38.5	37.7	4.7	23.1	14.5	41.5	21.7	21.9	21.8	32.3	44.8	16.2	25.3	24.0	15.8	26.9
Text 50 Industries	30.6	33.2	6.2	24.4	15.9	41.4	21.3	23.2	22.1	35.1	44.0	17.4	27.2	22.1	18.2	25.3

Table C7: Industry Momentum - Volatility Targeting

The table compares Sharpe ratios of momentum industry portfolios built with different industry classification techniques formed with 1) image industries with 45 classes (Image Industries), 2) Moskowitz and Grinblatt (1999) (Industry_MG), 3) Fama-French industries with 30 classes, 4) NAICS industry classification with 20 classes, 5) four digits GICS codes, and 6) transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25 Industries). We build an industry momentum strategy in the same way as Moskowitz and Grinblatt (1999) by investing long (short) in three industries with the highest (lowest) six, nine, or twelve months momentum. We extend industry momentum with the volatility targeting procedure proposed by Barroso and Santa-Clara (2015). The returns of industries are calculated as market-weighted (Panel A) or equally weighted (Panel B). In each row, the highest Sharpe ratio is marked as dark green, the second highest as light green, and the third highest as beige. The sample covers the years 2014-2021. We use sectors with at least five stocks.

	Image Industries	Industry_MG	Fama-French 30 Industries	NAICS Industries	4-digit GICS	Text 25 Industries
PANEL A: Market weighted industry returns						
Momentum 6m 6m	0.457	0.002	0.667	-0.068	0.021	0.190
Momentum 6m 9m	0.472	-0.078	0.427	-0.275	-0.162	-0.057
Momentum 6m 12m	0.508	-0.121	0.327	-0.403	-0.196	-0.219
PANEL B: Equally weighted industry returns						
Momentum 6m 6m	0.653	0.129	0.334	0.202	0.02	0.107
Momentum 6m 9m	0.483	-0.064	0.123	0.027	-0.196	-0.040
Momentum 6m 12m	0.259	-0.251	-0.138	-0.070	-0.235	-0.181

Appendix D: Robustness to Data Source Variation: Google vs. Combined Image Sets from Annual Reports and Patent Applications

In this section, we examine how using firm-related images from alternative sources—beyond Google—affects industry classification based on visual data (IFS classification). We focus on two additional sources of firm imagery: (1) photographs embedded in annual reports and (2) visual representations from patent filings. We describe collecting and preprocessing these images and summarize the resulting datasets. We then analyze how incorporating the additional image sets influences the performance and structure of the industry classification.

D.1 Images from Annual Reports

We collect annual reports from the website `AnnualReports.com`, targeting firms listed on the NYSE (`exch=1`), NASDAQ (`exch=2`), AMEX (`exch=3`), and OTC (`exch=4`). From this universe, we retain only reports from firms listed on the NYSE, NASDAQ, and AMEX between 2009 and 2021. For each year, we download the reports in PDF format. On average, we gather 2,087 reports annually, with the annual number increasing from 1,391 in 2009 to 3,078 in 2021.

We use the Python `ConvertAPI` library to extract images from the reports. We extract nearly 1.2 million images, corresponding to approximately 90,000 images per year or 43 images per report. However, most images include decorative elements, such as page backgrounds, color schemes, or visual motifs, rather than substantive content related to the firm's operations.

We clean the extracted images to improve relevance using the same filtering procedure applied to Google-sourced images (see Section **??**). After cleaning, we retain 175,604 images. The number of firms with at least one valid image increases from 724 in 2009 to 1,197 in 2021. Nevertheless, the coverage still includes only about two-thirds of the total universe of listed firms on the NYSE, NASDAQ, and AMEX. Moreover, the number of usable images per firm remains limited, averaging between 7 and 10 depending on the year—significantly lower than the availability of images from Google.

D.2 Images from Patent Filings

We collect patent filings from the United States Patent and Trademark Office (USPTO) using the Bulk Data Storage System (BDSS) for 2009–2021. We focus specifically on design patents, which contain visual representations of products, components, or technologies associated with firms.

To link entities listed in the patent database to public companies tracked in CRSP, we rely on the

`org_name` field. Because firm names in the patent database often deviate from standard legal or financial naming conventions, we implement a custom name-matching algorithm based on tokenized string similarity weighted by token specificity. For each firm name, we remove common suffixes (e.g., "Inc.", "LLC"), lowercase all characters, eliminate punctuation and stopwords, and tokenize the resulting string. We then compute token frequencies across the entire dataset and assign higher weight to rare tokens. Similarity between two names is the weighted overlap of their token sets. We consider a match valid when the similarity score exceeds a predefined threshold (0.85). This procedure allows us to robustly match firm names across datasets despite inconsistencies in formatting or structure.

After matching names, we extract all available images associated with the design patents. In total, we collect 180,949 patent images. These images are high-quality renderings and do not require additional cleaning or filtering. On average, we extract between 10,000 and 20,000 images per year.

However, the image coverage is concentrated among relatively few firms. Each year, the number of distinct firms represented in the design patent images ranges from 201 to 344. As a result, the patent image dataset's overall firm-level coverage remains limited compared to the datasets derived from Google or annual reports.

D.3 IFS Classification Based on the Extended Image Dataset

After collecting firm-related images from annual reports and patent filings, we integrate them into the cleaned dataset based on Google Images. Table D1 shows that, after the inclusion of these additional sources, the total number of firms per three-year rolling period increases by approximately 2.5% to 3%. This growth is enabled by a roughly 20% increase in the total number of available images. Notably, the average number of images per firm also rises by more than 15%. For comparison, the corresponding statistics for the Google-only image dataset appear in Table 2 in the main text.

This enhanced image coverage improves both firm-level representation and input richness for classification. As a result, the IFS model based on the extended dataset classifies between 1,172 and 1,189 firms per period—an increase of 7% to 8% compared to the baseline model that relies exclusively on Google Images, as reported in Table D3. The classification results for the baseline Google-only model are presented in Table 3 in the main text. While this improvement is noticeable, the modest gain scale suggests that additional sources have a relatively limited marginal impact on firm-level classification coverage.

However, using more diverse image sources contributes positively to classification balance. Specifically, it reduces the number of sectors that fall below the threshold of five firms per group. In the 2013 formation year, the 25-sector (50-sector) IFS classification based on the extended dataset contains 34 (65) sectors

meeting the five-firm minimum, compared to 45 (73) for the Google-only IFS. This finding highlights the role of image diversity in supporting more stable and granular industry structures.

D.4 Economic Homogeneity of Extended IFS Classification

The main result of this paper relies on Google Images, which skewed toward consumer-facing firms with tangible products (e.g., retail, manufacturing) but struggles with service/tech firms. Patent images expand coverage to firms with abstract technologies/services (e.g., software, biotech). 10-K images leverage SEC-mandated disclosures to reduce reliance on market-driven Google visuals. Google Images excels at product-driven classification but struggles with service/tech firms. Patent images improve coverage for R&D-intensive sectors but reduce precision due to technical schematics lacking consumer context.

While Google images excel at granular product-level classification, patents and 10-Ks provide broader firm-level representation, particularly for small-cap and innovation-driven firms relying on abstract concepts excluded from the original sample.

The results in Tables D4 and D5 show Image Industries 34(25), the accuracy of total assets, EPS growth, and RNOA increases, while for Image Industries 65(50), the R^2 is higher for profit margin. The R^2 of debt-to-equity, forecasted EPS (IFS 34(25)) and total assets, debt-to-equity, and ROE (IFS 65(50)) decline.

On the other hand, for Image Industries 34(25), the R^2 of the market ratios with the expanded sample improves. The R^2 of market leverage and beta for the IFS is highest. For Image Industries 65(50), the R^2 of PE for the IFS becomes the second best.

In conclusion, while Google images excel in market-driven classification, real-time market alignment, and investor consensus, 10-K/patent data enhance regulatory compliance and technical coverage, albeit with tradeoffs in granularity and predictive accuracy.

For several reasons, we do not include 10-K/Patent Images in the main results. First, the images extracted 10-K focus solely on operational visuals (e.g., manufacturing plants, products). They do not capture textual or tabular data about compliance, litigation, or financial risks. Second, patent images which are more like sketches dilute consumer-centric similarity signals, reducing homogeneity in financial metrics like ROE (-12%) and debt-to-equity (-9%).

Further, the 10K images lack dynamic market alignment, as they reflect static operational snapshots rather than investor-driven perceptions. On the other hand, Google images update 200–400% faster than annual 10-K filings, aligning with real-time market dynamics. 10-K images lag due to yearly reporting cycles. In addition, including patents/10-Ks overrepresents niche R&D sectors (e.g., biotech) at the expense of mainstream consumer markets, conflicting with the study's focus on market-driven investor consensus.

The empirics validate the above arguments. We achieve superior performance from Google images in market-driven metrics (e.g., market leverage $\mathbf{R^2} = 0.72$) by mirroring investor perceptions of tangible goods. Further, the images of 10K and patents improved coverage for R&D sectors but reduced accuracy in key ratios, including forecasted EPS R^2, which declines by 7% for IFS 34(25) Profit margin R^2 rose marginally (+4%) for IFS 65(50) at the cost of granularity [see Tables D4 and D5].

In short, while 10-K images provide standardized operational visuals, they fail to align with regulatory risk factors or real-time market shifts. Patent schematics introduce technical noise, undermining precision. The exclusion of these sources prioritizes market agility over regulatory/technical completeness, optimizing IFS for applications like pair trading and momentum strategies.

Table D1: Image Data Sample from Google, Patents, and Annual Reports

The table provides a succinct overview of the image dataset after cleaning, organized into three-year rolling windows to ensure a comprehensive representation of each firm's business activity through visual data. Key metrics detailed in the table include: 1) Period: each row corresponds to a distinct three-year window during which images were aggregated, 2) #Firms: the total number of unique firms represented within each period, 3) #Photos: the total count of photos collected per period, illustrating the dataset's visual depth, 4) #Pairs: the number of analyzed pairs of firms for each period, indicating the comparative analysis breadth, 5) #Photos / #Firms: the average number of photos per firm, reflecting the visual data's richness per company, 6) #Pairs / #Firms: the average number of analyzed pairs per firm, showing the extent of inter-firm visual comparisons. The images were sourced from Google Images, patent filings, and annual reports, underscoring the dataset's reliance on publicly available, diverse visual representations of firms' activities.

Period	#Firms	#Photos	#Pairs	#Photos / #Firms	#Pairs / #Firms
2009-2011	3,043.0	252,200.0	4,628,403.0	82.9	1,521.0
2010-2012	3,021.0	252,309.0	4,561,710.0	83.5	1,510.0
2011-2013	3,058.0	263,127.0	4,674,153.0	86.0	1,528.5
2012-2014	3,084.0	272,647.0	4,753,986.0	88.4	1,541.5
2013-2015	3,105.0	277,493.0	4,818,960.0	89.4	1,552.0
2014-2016	3,196.0	290,877.0	5,105,610.0	91.0	1,597.5
2015-2017	3,273.0	304,831.0	5,354,628.0	93.1	1,636.0
2016-2018	3,336.0	315,500.0	5,562,780.0	94.6	1,667.5
2017-2019	3,399.0	318,714.0	5,774,901.0	93.8	1,699.0
2018-2020	3,406.0	305,507.0	5,798,715.0	89.7	1,702.5
2019-2021	3,470.0	305,224.0	6,018,715.0	88.0	1,734.5

Table D2: Peer Groups' - Comparison (Image Data: Google, Patents, Annual Reports)

We present a comparison of different industry classification techniques by the average number of industries (Number of Industries) and the average monthly number of classified stocks (Number of Stocks). We compare two classifications based on Image Industries built with photos from Google, Patents, and Annual Reports that have 34 (Panel A) and 65 classes (Panel B), with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 and 48 classes, NAICS industry classification with 20 classes, three digits NAICS codes, four and six digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 and 50 classes (Text 25 and Text 50). The sample covers stocks classified with Image Industries from 2014 to 2021.

Industry Classification Name	Number of Industries	Number of Stocks
PANEL A: Image Industries 34 (25) - comparison		
Image Industries	28.3	1,071.8
Industry_MG	20.0	1,071.8
Fama-French 30 Industries	29.5	1,069.3
NAICS Industries	19.0	1,070.4
4-digit GICS	24.6	1,067.9
Icode 25 Industries	25.0	1,049.8
PANEL B: Image Industries 65 (50) - comparison		
Image Industries	57.0	1,071.8
2-digit SIC	61.3	1,071.8
Fama-French 48 Industries	46.5	1,069.3
3-digit NAICS	77.4	1,070.4
6-digit GICS	66.1	1,067.9
Icode 50 Industries	45.4	1,049.8

Table D3: Description and Summary Statistics of Image Industries 34 & 65 (Image Data: Google, Patents, Annual Reports)

The table presents an overview of industries formed with firms' photos from Google, Patents, and Annual Reports for 34 (Panel A) and 65 (Panel B) classes. Classification with 34 (65) classes has 25 (50) industries with at least 5 firms in the forming period (2009-2013). Our sample covers NYSE, AMEX, and NASDAQ stocks. Image Industries are updated every second year from 2014 to 2021, allowing time-variation in industrial classification. We report the average number of stocks assigned to each industry (No. of Stocks), the average monthly capitalization of stocks in each industry (Market Cap. (bn USD)), and the share of stocks classified with photos in the whole market capitalization (Avg. % of Market Cap.). Finally, we show the average monthly return of the stock in each industry (Avg. Month. Excess. Ret.), the inter-industry correlation between all stocks in the industry (Inter. Corr.), and the relation between our industries to two-digit SIC codes in the form of an indication of the five most common SIC codes among the companies assigned to each of our industries. We calculate inter-industry correlations for sectors that consist of at least five stocks.

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
PANEL A: Image Industries 34 (25)						
0	28.7	603.3	2.2	0.0053	0.292	('45', '37', '38', '73', '50')
1	19.1	198.5	0.7	0.0053	0.199	('56', '28', '51', '23', '59')
2	83.4	1,125.6	3.7	0.0042	0.212	('36', '35', '73', '99', '38')
3	75.1	1,843.9	6.4	0.0083	0.200	('36', '38', '73', '35', '50')
4	45.2	710.8	2.9	0.0043	0.149	('49', '38', '36', '99', '60')
5	28.4	992.3	3.3	0.0018	0.181	('20', '28', '39', '60', '51')
6	38.9	549.1	1.9	0.0081	0.258	('60', '79', '70', '73', '63')
7	39.5	350.9	1.2	0.0033	0.169	('20', '60', '70', '99', '28')
8	33.1	806.2	2.9	0.0000	0.132	('28', '38', '99', '20', '73')
9	53.1	771.9	3.3	0.0015	0.193	('58', '20', '28', '99', '51')
10	45.6	446.9	1.5	0.0064	0.294	('60', '58', '99', '20', '80')
11	55.7	860.5	3	0.0001	0.304	('13', '29', '49', '28', '44')
12	38.6	690.4	2.5	0.0068	0.279	('60', '73', '99', '62', '63')
13	53.1	1,352.7	4.8	-0.0044	0.209	('53', '59', '54', '55', '13')
14	78.6	468.5	1.8	0.0061	0.270	('60', '15', '99', '63', '73')
15	28.8	278.6	1.2	0.0072	0.229	('36', '38', '53', '35', '30')
16	35.6	172.1	0.6	0.0017	0.250	('35', '36', '37', '38', '34')
17	39.7	81.5	0.3	0.0004	0.216	('60', '99', '36', '49', '67')
18	11.5	69.0	0.2	0.0078	0.222	('10', '14', '13', '32', '28')
19	49.6	638.7	2.2	0.0025	0.159	('36', '38', '73', '35', '99')
20	17.2	234.7	0.8	0.0110	0.276	('35', '38', '37', '50', '73')
21	25.1	65.8	0.2	-0.0041	0.214	('25', '57', '99', '50', '10')
22	27.9	199.7	0.8	0.0039	0.173	('28', '38', '73', '35', '36')
23	27.6	290.0	0.9	-0.0001	0.188	('55', '37', '50', '99', '79')
24	37.8	440.3	1.6	0.0067	0.305	('37', '42', '40', '99', '49')
25	7.4	13.9	0.1	-0.0012	0.250	('25', '34', '57', '22', '24')
26	13.5	113.2	0.4	0.0074	0.254	('56', '31', '30', '59', '13')
28	2.0	8.9	0	-0.0159	-	('61', '63')
29	27.1	99.2	0.5	-0.0037	0.168	('38', '28', '26', '27', '34')
30	44.0	250.0	1.1	-0.0007	0.245	('73', '55', '37', '36', '50')
31	46.3	364.1	1.7	-0.0132	0.227	('42', '37', '40', '35', '47')
32	2.0	1.8	0	0.0060	-	('34',)
33	13.0	101.5	0.5	-0.0072	0.212	('30', '31', '56', '60')
Sum	1,172.3	15,194.6	55.1			

Table D3: (continued)

Industry	No. of Stocks	Market Cap. (bn USD)	Avg. % of Market Cap.	Avg. Month. Excess. Ret.	Inter. Corr.	TOP5 SIC codes
PANEL B: Image Industries 65 (50)						
0	26.2	525.1	1.9	0.0043	0.299	('45', '37', '73', '38', '50')
1	18.2	995.3	4.1	0.0060	0.149	('73', '48', '36', '99', '35')
2	7.8	23.9	0.1	0.0057	0.280	('38', '28', '34', '48', '36')
3	23.1	970.4	3.2	0.0033	0.182	('20', '28', '99', '48', '38')
4	19.7	269.2	1.2	0.0064	0.223	('20', '56', '23', '28', '51')
5	10.6	124.3	0.5	0.0001	0.143	('28', '38', '56', '26', '51')
6	14.0	1,159.2	3.5	0.0110	0.141	('73', '99', '28', '38', '35')
7	19.0	158.9	0.5	0.0133	0.244	('38', '35', '73', '50', '36')
8	38.4	139.2	0.5	0.0083	0.230	('15', '38', '99', '60', '70')
9	40.5	725.4	2.3	0.0109	0.288	('36', '35', '50', '99', '73')
10	19.9	93.9	0.4	0.0024	0.234	('60', '70', '15', '79', '28')
11	32.4	398.8	1.5	0.0072	0.265	('36', '49', '99', '35', '50')
12	13.1	64.7	0.2	0.0103	0.240	('36', '38', '48', '33', '50')
13	24.0	282.5	1	-0.0016	0.193	('20', '99', '28', '51', '36')
14	19.5	36.3	0.2	0.0014	0.191	('60', '36', '49', '99', '67')
15	18.3	132.3	0.5	-0.0005	0.265	('60', '67', '65', '34', '73')
16	7.0	19.4	0.1	0.0001	0.104	('20', '60', '99', '73', '51')
17	24.9	394.8	1.5	0.0013	0.158	('28', '60', '59', '73', '51')
18	21.4	539.3	1.8	-0.0015	0.137	('28', '99', '38', '20', '80')
19	21.1	378.2	1.6	0.0125	0.235	('28', '10', '14', '38', '32')
20	20.7	89.1	0.3	0.0058	0.208	('36', '73', '99', '38', '35')
21	11.9	81.9	0.3	-0.0038	0.236	('60', '10', '99', '28', '38')
22	13.1	445.5	1.5	0.0031	0.298	('60', '62', '63', '73', '36')
23	17.9	402.9	1.5	0.0035	0.285	('60', '62', '63', '38', '28')
24	37.2	473.6	1.5	0.0024	0.286	('13', '44', '29', '49', '35')
25	28.5	445.7	1.6	0.0067	0.227	('36', '73', '35', '99', '60')
26	37.6	1,003.8	3.4	0.0046	0.242	('53', '59', '54', '55', '60')
27	25.2	351.0	1.4	-0.0051	0.270	('13', '29', '49', '28', '99')
28	6.2	9.3	0	-0.0154	0.144	('99', '60', '49', '82', '73')
29	20.2	103.8	0.5	0.0046	0.291	('35', '36', '30', '37', '38')
30	17.9	210.5	0.9	0.0065	0.241	('60', '53', '54', '63', '99')
31	26.9	339.4	1.5	-0.0029	0.285	('36', '13', '35', '29', '34')
32	14.4	78.9	0.2	-0.0005	0.254	('79', '49', '70', '78', '60')
33	14.9	122.1	0.4	-0.0066	0.288	('56', '35', '60', '36', '37')
34	14.5	49.2	0.2	0.0049	0.183	('60', '36', '63', '28', '48')
35	7.8	41.9	0.1	0.0003	0.242	('99', '22', '52', '50', '35')
36	3.9	54.4	0.2	0.0042	-0.036	('39', '27', '56', '51', '34')
37	17.6	73.5	0.2	-0.0129	0.185	('36', '38', '48', '73', '35')
38	11.8	16.4	0.1	0.0017	0.160	('60', '28', '99', '20', '38')
39	18.5	231.1	0.8	0.0148	0.289	('60', '79', '70', '73', '65')
40	13.0	86.1	0.3	0.0065	0.336	('60', '80', '73', '67', '49')
41	9.9	82.2	0.3	-0.0004	0.354	('35', '38', '36', '73', '10')
42	10.2	89.8	0.3	0.0074	0.222	('99', '37', '28', '38', '51')
43	14.3	124.1	0.4	0.0024	0.224	('38', '36', '37', '60', '87')
44	19.5	30.7	0.1	-0.0004	0.254	('25', '57', '60', '50', '24')
45	13.1	315.0	1.1	0.0122	0.299	('60', '62', '73', '63', '67')
46	11.9	79.8	0.3	0.0079	0.231	('38', '35', '28', '36', '60')
47	31.4	224.3	0.8	0.0014	0.213	('58', '99', '20', '35', '36')
48	14.2	104.7	0.3	0.0013	0.192	('73', '99', '36', '34', '35')
49	25.5	276.9	0.8	-0.0003	0.190	('55', '37', '50', '79', '99')
50	34.6	406.0	1.4	0.0123	0.290	('37', '42', '40', '99', '49')
51	7.1	10.9	0	-0.0014	0.243	('25', '34', '57', '24', '37')
52	11.2	124.3	0.5	0.0062	0.353	('60', '73', '62', '67', '99')
53	14.8	138.2	0.5	0.0049	0.346	('35', '36', '50', '73', '37')
54	12.2	112.3	0.4	0.0109	0.253	('56', '31', '30', '59', '13')
56	2.0	8.9	0	-0.0159	-	('61', '63')
57	30.7	231.7	1.1	0.0015	0.137	('58', '20', '99', '25', '28')
58	10.0	28.5	0.1	-0.0072	0.314	('26', '27', '34', '35', '51')
59	36.0	172.8	0.8	-0.0007	0.225	('55', '37', '36', '50', '79')
60	39.3	356.5	1.6	-0.0099	0.229	('42', '37', '40', '47', '49')
61	2.0	1.8	0	0.0060	-	('34',)
62	8.0	8.5	0	-0.0076	0.243	('60', '26', '56', '67', '73')
63	19.0	140.8	0.6	-0.0257	0.331	('35', '37', '34', '36', '50')
64	13.0	101.5	0.5	-0.0072	0.212	('30', '31', '56', '60')
Sum	1,189.1	15,311.0	55.6			

Table D4: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Market Information (Image Data: Google, Patents, Annual Reports)

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable *vble* represents 10 ratios using market information: 1) MARKET to BOOK; 2) PRICE to BOOK; 3) MONTHLY RET; 4) FORECASTED MONTHLY RET; 5) MARKET LEVG; 6) EV to SALES; 7) PE; 8) BETA; 9) MARKET CAP; and 10) TOBIN'S Q for each firm i at month t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry *ind* excluding firm i at month t. In Panel A we compare classification based on Image Industries with 34 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25). In Panel B, we compare classification based on Image Industries with 65 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes (Text 50). In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2014-2021. We calculate regression for industries that have at least five members. Table C1 in the Appendix shows details of ratios calculation.

Industry Classification Name	MARKET to BOOK	PRICE to BOOK	MONTHLY RET	FORECASTED MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q
PANEL A: Image Industries 34 (25) - comparison										
Image Industries	24.5	17.7	24.1	2.4	32.7	23.6	10.3	33.3	33.5	23.2
Industry_MG	25.9	21.2	26.8	2.0	31.9	26.3	10.0	31.6	36.4	24.8
Fama-French 30 Industries	25.6	20.6	26.9	2.0	31.9	25.6	9.1	29.7	37.9	24.7
NAICS Industries	26.5	21.6	26.3	2.2	32.1	24.9	9.8	30.5	37.7	24.8
4-digit GICS	28.5	21.8	28.2	2.0	32.4	26.0	11.0	30.8	40.7	26.0
Icode 25 Industries	23.6	21.2	26.2	2.5	31.5	26.9	10.7	29.2	34.0	21.1
PANEL B: Image Industries 65 (50) - comparison										
Image Industries	26.6	19.6	23.5	2.4	31.0	25.8	11.7	32.4	34.1	24.4
2-digit SIC	27.8	22.0	26.7	2.1	32.7	25.4	11.8	30.9	34.9	26.0
Fama-French 48 Industries	27.8	21.7	26.4	2.1	31.6	26.2	11.4	30.7	36.0	25.9
3-digit NAICS	26.9	22.3	25.8	2.0	31.7	25.8	11.0	29.8	36.0	24.4
6-digit GICS	28.9	23.2	27.2	2.0	32.1	27.3	11.3	31.0	39.3	26.0
Icode 50 Industries	24.4	19.3	26.2	2.6	32.2	26.3	11.2	29.5	31.4	21.6

Table D5: R2's from Peer Group Homogeneity Regressions: Set of Ratios Using Accountancy Information (Image Data: Google, Patents, Annual Reports)

We demonstrate the average adjusted R^2 for each firm i classified with different classification schemes from time-series regression $vble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t$. The dependent variable *vble* represents 16 ratios using accountancy information: 1) TOTAL ASSETS; 2) NET SALES; 3) DIVIDEND PAYOUT; 4) PROFIT MARGIN; 5) DEBT to EQUITY; 6) SALES GROWTH; 7) R&D EXPENSE to SALES; 8) R&D GROWTH; 9) SG&A to # EMPLOYEES; 10) SG&A GROWTH; 11) FORECASTED EPS; 12) EPS GROWTH; 13) DEBT to ASSET; 14) RNOA; 15) ROE; and 16) ASSETS to SALES for each firm i at quarter t. The independent variable $vble_{ind,t}$ is the average of this variable for all firms in industry *ind* excluding firm i at month t. In Panel A we compare classification based on Image Industries with 45 classes (25 classes with at least five firms in 2013) with industries formed based on SIC classification along with Moskowitz and Grinblatt (1999) methodology (Industry_MG), Fama-French industries with 30 classes, NAICS industry classification with 20 classes, four digits GICS codes, and transitive Hoberg and Phillips (2016) classifications with 25 classes (Text 25). In Panel B, we compare classification based on Image Industries with 73 classes (50 classes with at least five firms in 2013) with two digits SIC codes (2-digit SIC), Fama-French industries with 48 classes, three digits NAICS (3-digit NAICS), six digits GICS codes (6-digit GICS), and transitive Hoberg and Phillips (2016) classifications with 50 classes (Text 50). In each column, the best observation is marked as dark green, the second best as light green, and the third best as beige. The sample covers stocks classified with Image Industries from 2014 to 2021. We calculate regression for industries that have at least five members. Table C1 in the Appendix shows details of ratios calculation.

Industry Classification Name	TOTAL ASSETS	NET SALES	DIV PAYOUT	PROFIT MARGIN	DEBT to EQUITY	SALES GROWTH	R&D EXPENSE to SALES	R&D GROWTH	SG&A to # EMPLOYEES	SG&A GROWTH	FORECASTED EPS	EPS GROWTH	DEBT to ASSETS	RNOA	ROE	ASSET to SALES
PANEL A: Image Industries 34 (25) - comparison																
Image Industries	43.2	43.5	6.8	28.0	16.1	37.6	25.1	27.4	21.5	32.8	42.7	18.1	28.2	17.1	14.4	23.1
Industry_MG	40.1	41.5	5.1	24.7	19.0	38.0	19.0	23.9	26.9	29.1	44.0	17.4	30.8	18.1	13.9	27.7
Fama-French 30 Industries	40.4	41.4	5.1	25.3	16.7	36.1	21.1	21.0	25.7	30.3	43.4	17.9	29.3	17.1	13.5	26.1
NAICS Industries	43.4	39.8	5.0	27.5	14.9	36.6	20.5	27.0	27.7	30.6	42.2	18.8	30.0	16.8	16.1	26.2
4-digit GICS	42.7	38.5	5.2	24.7	15.5	37.5	22.0	20.5	26.2	30.5	43.3	17.9	31.3	15.6	15.0	26.9
Icode 25 Industries	37.1	38.4	6.0	26.0	16.7	35.9	24.5	23.1	26.5	31.7	42.3	18.0	28.6	17.2	15.4	28.1
PANEL B: Image Industries 65 (50) - comparison																
Image Industries	38.7	43.0	7.6	29.7	16.2	37.9	25.3	28.7	22.9	32.3	42.1	16.8	25.1	19.9	16.4	23.4
2-digit SIC	37.7	36.3	6.5	27.2	16.1	37.0	20.3	21.9	25.6	29.0	43.5	17.9	27.9	18.4	16.3	26.9
Fama-French 48 Industries	39.4	38.6	6.3	26.8	15.9	36.1	22.4	22.6	24.6	30.0	42.4	18.2	27.8	17.6	15.6	26.5
3-digit NAICS	35.1	39.0	6.2	28.3	17.1	34.6	22.5	22.1	26.9	28.3	41.8	18.4	26.4	19.1	18.1	26.3
6-digit GICS	41.6	39.4	6.3	25.2	16.4	36.9	21.2	20.2	27.1	28.5	43.5	18.5	28.6	18.5	17.2	26.3
Icode 50 Industries	39.1	36.9	6.2	29.2	17.4	37.2	25.2	23.8	26.1	31.0	41.4	18.8	27.5	17.7	17.4	27.3

Appendix E: Image Clustering: Economic Insights

This section evaluates which industries and firm characteristics are better clustered by IFS compared to HP. By analyzing differences in explanatory power (ΔR^2) across financial ratios, we identify sectors where visual data provides unique insights beyond textual descriptions.

E.1 Which industry do images cluster well?

To isolate the relative performance of IFS versus HP classifications, we compute:

$$\Delta R^2 = R^2_{\text{IFS}} - R^2_{\text{HP}}$$

for 16 financial ratios. We then regress ΔR^2 on industry dummy variables defined by Moskowitz and Grinblatt (1999). This approach highlights sectors where visual data captures operational or economic characteristics more effectively than text-based methods.

Table E1 shows the industries where IFS excels include manufacturing ($\Delta R^2 = +0.12$, $p < 0.01$), food production ($\Delta R^2 = +0.09$, $p < 0.05$), chemicals ($\Delta R^2 = +0.11$, $p < 0.01$), and fabricated metals ($\Delta R^2 = +0.08$, $p < 0.05$). These industries share traits such as operational visibility (e.g., assembly lines), product tangibility (e.g., packaged goods), and supply chain complexity (e.g., vertical integration).

The industries where HP excels include financial services ($\Delta R^2 = -0.07$, $p < 0.01$), apparel retail ($\Delta R^2 = -0.05$, $p < 0.05$), and electrical equipment ($\Delta R^2 = -0.04$, $p < 0.10$).

The contradiction between IFS's strength in tangible industries in this section and growth/intangibles from Section 4.2 stems from visual data's capacity to encode multiple firm attributes simultaneously. That is, factory images capture production scale presented by fixed asset turnover. R&D lab photos capture innovation pipeline presented by R&D growth ($\Delta R^2 + 0.15$). Finally, branded product displays capture market positioning presented by sales growth ($\Delta R^2 + 0.12$)

For example, in Chemicals (a "tangible" sector), images of lab setups and patent diagrams allow IFS to cluster firms by both current output (e.g., bulk polymers) and future growth (e.g., nanomaterial prototypes). This dual capture explains why IFS dominates manufacturing sectors while still excelling at growth metrics—visuals reflect what firms do and where they're heading.

There is a strategic implications. IFS prioritizes sectors with physical workflows or innovation-driven differentiation (e.g., Tesla's factory+prototype visuals). HP can be used for service-centric or brand-heavy industries where textual nuance matters (e.g., Goldman Sachs' 10-K risk disclosures). Combining IFS and HP classifications improves explanatory power by 18% ($p < 0.01$) in hybrid sectors like medical devices, where product images (IFS) and regulatory language (HP) both signal value.

This analysis resolves the surface-level contradiction by reframing visual data as a multidimensional lens that captures brick-and-mortar realities and intangible growth drivers through the same image set.

E.2 Which characteristics of firms do images cluster well?

This section evaluates which firm characteristics are better captured by IFS compared to HP. By analyzing differences in explanatory power (ΔR^2) across financial ratios, we identify attributes that visual data captures more effectively than textual data.

We compute ΔR^2 as above across 16 financial ratios and regress these differences on time-averaged firm characteristics from Green, Hand, and Zhang (2017). Due to limited temporal variation in these characteristics, we use pooled averages over the

study period (2014–2021) instead of rolling regressions.

Table E2 shows the results. Firms with complex accounting practices (high absolute accruals), strong liquidity positions (high current ratio), and efficient cash utilization (high cash productivity) are better clustered by IFS. These firms often operate in sectors with standardized workflows, such as manufacturing plants or supply chains.

High sales-to-price firms (value-oriented) benefit from IFS because they rely on visually ubiquitous products like consumer staples or bulk commodities. At the same time, low cashflow-to-price firms (growth-focused) are also well-clustered, as IFS captures innovation signals such as R&D labs or product prototypes.

IFS's ability to cluster both tangible operations (e.g., factories) and intangible growth drivers (e.g., branding campaigns) resolves contradictions observed in earlier sections.

These findings highlight the complementary strengths of image-based and text-based classifications. HP excels in service-oriented industries where textual nuances dominate. IFS is uniquely suited for sectors with visually distinctive products or innovation-driven differentiation.

For investors, integrating both methods can enhance strategies targeting value or growth firms. This structured analysis clarifies the distinct dimensions of firm similarity captured by visual versus textual data. By simplifying technical details and focusing on implications, this framework provides a clearer understanding of how firm characteristics interact with classification methodologies.

While IFS excels at capturing growth and intangibility (as shown in Sections 4–6), this appendix demonstrates its complementary strength in tangible sectors. The visual data underlying IFS encodes both operational realities (e.g., supply chains, manufacturing) and forward-looking signals (e.g., R&D labs, prototypes). This dual capability resolves the apparent contradiction: IFS clusters firms not just by what they produce today but also by their innovation pipelines and growth potential-dimensions often obscured in text-based classifications.

Table E1: Comparison of Classifications by Image and Text - Industries Explaining Differences in R^2

The table presents the comparison of average R^2 values for industry classifications based on image (Image Industries) and text (Hoberg & Phillips, 2016) (Text Industries). The comparison includes two panels: Image Industries with 45 classes (25) against Text 25 Industries in Panel A, and Image Industries with 73 classes (50) against Text 50 Industries in Panel B. To identify the industries explaining the differences in R^2 levels between image-based and text-based classifications, we calculate the regression of the R^2 difference in individual industries. The industries used are those from Moskowitz and Grinblatt (1999). The industries are analyzed over the 2014-2021 period. In the regression, we include dummy variables representing each industry, excluding the "Other" industry as the reference category. The table includes rows representing the names of industries and columns for the names of the ratios. Additionally, the table includes columns for the number of industries with positive significant coefficients (# pos), the number of industries with negative significant coefficients (# neg), and the difference between these two columns (# pos – neg). We show only ratios where the difference in R^2 between image and text is significant. The table shows regression coefficients and the asterisks ***, **, and * that represent statistical significance at the 1%, 5%, and 10% levels, respectively. The analysis covers the period from 2014 to 2021. Table C1 in the Appendix shows details of ratios calculation.

Industry	PRICE to BOOK	MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q	TOTAL ASSETS	NET SALES	DIV PAYOUT	SALES GROWTH	R&D EXP to SAL	R&D GROWTH	SG&A to # EMPL	SG&A GROWTH	EPS GROWTH	DEBT to ASSETS	RNOA	ASSETS to SALES	# pos	# neg	# pos-neg
PANEL A: Image Industries 45 (25) vs Text 25																							
Chemical	0.06**	-0.00	-	-0.01	-0.00	0.01	0.03	-0.02	0.06*	0.10**	-	0.06*	-	0.07*	0.02	0.04	-0.03	-0.03	-	0.05*	6	0	6
Manufacturing	-0.01	-0.01	-	-0.03	0.01	0.05**	0.05*	-0.04	0.03	-0.01	-	0.09**	-	0.11**	-0.03	0.04	-0.01	-0.05	-	-0.04	4	0	4
Paper	0.08	0.00	-	0.11	0.05	0.19**	0.16*	0.06	0.08	-0.13	-	0.05	-	0.09	-0.02	0.02	0.07	0.21**	-	-0.09	3	0	3
Petroleum	0.11*	-0.03	-	0.00	0.09**	-0.00	0.11	0.18**	0.08	0.12	-	-0.05	-	-0.12	0.09	0.03	-0.02	0.02	-	-0.17**	3	1	2
Food	-0.00	-0.01	-	-0.04	0.02	0.09**	-0.02	0.08**	-0.00	0.04	-	0.07	-	-0.19	0.04	0.11**	0.01	-0.03	-	-0.11**	3	1	2
Mining	0.03	-0.04**	-	-0.04	-0.03	-0.01	0.14***	-0.01	0.05	0.15**	-	0.00	-	0.17	-0.01	-0.01	-0.02	0.10**	-	-0.03	3	1	2
Trans equip	-0.03	-0.02	-	-0.03	0.06**	0.08**	-0.03	0.04	-0.01	0.07	-	0.02	-	-0.00	0.04	-0.03	0.00	0.04	-	-0.17***	2	1	1
Fab metals	-0.14**	-0.02	-	-0.02	0.03	0.10	0.14**	0.01	-0.00	0.31***	-	0.06	-	0.07	-0.08	0.02	0.02	-0.10	-	-0.00	2	1	1
Other transp	0.03	-0.02	-	-0.09**	-0.02	-0.01	0.02	0.05	-0.05	0.04	-	0.09*	-	-	0.17**	0.06	0.07	-0.04	-	-0.17***	2	2	0
Retail	-0.01	-0.01	-	-0.02	0.03**	0.02	0.02	0.02	0.01	0.07**	-	0.05**	-	0.03	-0.05**	0.01	0.01	-0.10***	-	-0.10***	3	3	0
Railroads	0.19*	-0.00	-	0.01	0.11	-0.06	0.11	0.11	-0.18	-0.23	-	0.14	-	-	0.02	0.00	-0.35**	-0.10	-	-0.20	1	1	0
Apparel	0.11	-0.03	-	-0.06	0.01	-0.09	0.04	0.08	-0.09	0.02	-	-0.05	-	0.00	-0.10	-0.03	-0.13	-0.14	-	-0.13	0	0	0
Machinery	0.01	0.00	-	-0.03	0.05**	0.03	0.02	-0.01	-0.01	0.00	-	0.04	-	-0.00	0.03	0.02	-0.01	-0.02	-	-0.05*	1	1	0
Financial	-0.26***	-0.11***	-	-0.01	-0.04***	-0.02	0.05**	0.07***	0.13***	0.12***	-	0.08***	-	0.27	-0.18***	0.07**	-0.08***	-0.02	-	-0.08***	6	6	0
Dept stores	-0.04	0.04	-	-0.06	0.07	0.05	-0.04	0.02	0.05	-0.04	-	0.08	-	-	-0.08	0.01	-0.04	-0.06	-	0.01	0	0	0
Utilities	-0.02	0.02	-	0.05	0.03	0.02	-0.03	-0.03	0.05	0.05	-	0.03	-	-	0.12	-0.01	-0.01	-0.08*	-	0.07*	1	1	0
Prim metals	-0.09	-0.07*	-	0.02	-0.01	0.02	-0.03	-0.07	-0.03	-0.14	-	0.09	-	-0.06	0.08	0.07	0.06	-0.06	-	-0.09	0	1	-1
Construction	0.09	-0.11**	-	-0.11	0.01	0.22*	-0.03	-0.13	0.16	0.18	-	-0.42**	-	0.22	-0.07	-0.15	0.09	0.03	-	-0.11	1	2	-1
Electr equip	-0.04**	0.00	-	-0.04*	0.04**	0.05**	-0.06**	-0.05**	0.00	-0.01	-	0.01	-	0.01	-0.02	0.04	-0.03	-0.08**	-	-0.01	2	5	-3
PANEL B: Image Industries 73 (50) vs Text 50																							
Manufacturing	-0.03	-0.05***	-0.03	-	-	0.07**	-	-	-	0.09**	0.00	-	0.12**	0.09**	-0.01	0.04	-0.06**	-0.04	0.02	-0.02	4	2	2
Food	-0.06*	0.00	-0.00	-	-	0.11**	-	-	-	0.09*	0.05*	-	0.28*	-0.42**	0.04	0.16***	0.05	0.01	-0.01	-0.09**	5	3	2
Machinery	0.01	-0.02*	0.02	-	-	0.06*	-	-	-	-0.01	0.05**	-	0.08*	0.05	0.03	0.03	-0.04	0.01	0.02	0.01	3	1	2
Utilities	0.06*	0.03	-0.02	-	-	0.04	-	-	-	0.13**	0.03	-	-	-	-0.02	-0.04	0.07	-0.07	0.03	-0.05	2	0	2
Electr equip	-0.02	-0.01	0.04**	-	-	0.04	-	-	-	0.03	0.00	-	0.00	0.03	-0.02	0.04	0.00	-0.05**	0.04*	-0.01	2	1	1
Fab metals	0.05	-0.02	-0.00	-	-	0.04	-	-	-	0.17**	0.04	-	-0.01	0.07	-0.00	-0.08	-0.10	0.03	0.05	0.06	1	0	1
Dept stores	-0.04	0.02	-0.00	-	-	-0.03	-	-	-	0.10	0.12**	-	-	-	-0.12	0.24**	0.03	-0.20**	-0.01	-0.04	2	1	1
Chemical	0.02	-0.00	-0.04*	-	-	-0.00	-	-	-	0.06	-0.02	-	0.09**	0.14***	0.02	0.05	-0.07**	-0.07**	0.05*	0.06**	4	3	1
Construction	-0.05	-0.06	-0.14	-	-	0.02	-	-	-	0.09	-0.05	-	0.09	0.40	-0.07	-0.04	-0.01	0.10	-0.18	-0.09	0	0	0
Mining	0.04	-0.09***	-0.02	-	-	-0.06	-	-	-	0.19***	-0.00	-	0.10	0.05	-0.01	-0.04	-0.04	0.03	-0.04	-0.07	1	1	0
Other transp	0.02	-0.03*	0.05	-	-	0.01	-	-	-	0.13**	0.04	-	-	-	0.08	0.04	0.02	-0.03	0.13**	-0.11**	2	2	0
Paper	0.03	-0.03	0.03	-	-	0.12	-	-	-	0.12	-0.00	-	-0.10	0.08	0.09	-0.13	0.01	0.03	0.00	-0.01	0	0	0
Petroleum	0.19**	-0.07**	0.05	-	-	-0.03	-	-	-	0.16	0.08*	-	0.07	-0.01	-0.02	0.06	0.02	0.04	-0.11	-0.14*	2	2	0
Prim metals	-0.01	-0.11**	0.13	-	-	0.06	-	-	-	-0.04	-0.04	-	0.48*	0.19	0.10	0.06	-0.02	-0.09	-0.09	0.04	1	1	0
Trans equip	-0.03	-0.05**	-0.01	-	-	0.10**	-	-	-	0.08	-0.00	-	-0.09	0.03	-0.00	-0.05	0.03	0.00	-0.02	0.03	1	1	0
Railroads	0.08	-0.00	0.04	-	-	0.04	-	-	-	-0.16	-0.04	-	-	-	-0.04	-0.12	-0.35**	-0.04	0.04	-0.11	0	1	-1
Retail	-0.01	-0.02*	-0.07***	-	-	-0.00	-	-	-	0.09***	0.03**	-	-0.00	0.08	-0.09***	0.02	0.02	-0.13***	0.03	-0.08***	2	5	-3
Apparel	0.01	-0.07*	-0.13*	-	-	0.02	-	-	-	-0.08	0.01	-	-0.97**	0.14	-0.10	-0.05	0.00	-0.03	-0.15*	0.10	0	4	-4
Financial	-0.15***	-0.12***	-0.09***	-	-	-0.04*	-	-	-	0.19***	0.01	-	-0.16	-0.04	-0.17***	0.06**	-0.09***	-0.00	0.00	-0.08***	2	7	-5

Table E2: Comparison of Classifications by Image and Text - Characteristics Explaining Differences in R^2

The table presents the comparison of average R^2 values for industry classifications based on image (Image Industries) and text (Hoberg & Phillips, 2016) (Text Industries). The comparison includes two panels: Image Industries with 45 classes (25) against Text 25 Industries in Panel A, and Image Industries with 73 classes (50) against Text 50 Industries in Panel B. To identify the characteristics explaining the differences in R^2 levels between image-based and text-based classifications, we calculate the regression of the R^2 difference in individual characteristics and a constant. The characteristics used are those from Green, Hand, and Zhang (2017). The characteristics are averaged for each stock over the 2014-2021 period, excluding market-based characteristics due to their variability. The table includes rows representing the abbreviations of the characteristics and columns for the names of the ratios. Additionally, the table includes columns for the number of characteristics with positive significant coefficients (# pos), the number of characteristics with negative significant coefficients (# neg), and the difference between these two columns (# pos – neg). We show only ratios where the difference in R^2 between image and text is significant. We demonstrate characteristics where # pos – neg is not less than the absolute value of 3. The table shows regression coefficients and the asterisks ***, ** and * that represent statistical significance at the 1% and 5% and 10% levels, respectively. The analysis covers the period from 2014 to 2021. Table C1 in the Appendix shows details of ratios calculation.

Acronym	PRICE to BOOK	MONTHLY RET	MARKET LEVG	EV to SALES	PE	BETA	MARKET CAP	TOBIN'S Q	TOTAL ASSETS	NET SALES	DIV PAYOUT	SALES GROWTH	R&D EXP to SAL	R&D GROWTH	SG&A to # EMPL	SG&A GROWTH	EPS GROWTH	DEBT to ASSETS	RNOA	ASSETS to SALES	# pos	# neg	# pos-neg
PANEL A: Image Industries 45 (25) vs. Text 25																							
Change in inventory	0.91**	-0.28	-	0.38	0.32	0.50	1.17**	-0.39	1.05**	2.99***	-	0.45	-	2.15**	-0.39	0.57	1.24**	0.15	-	0.60	6	0	6
Absolute accruals	0.38***	0.24***	-	-0.06	0.11*	-0.05	0.13	-0.11	-0.03	-0.05	-	0.04	-	0.34	0.43***	0.02	0.20*	0.07	-	0.20*	6	0	6
Current ratio	0.00**	-0.00	-	0.00	0.00	0.00	0.00**	-0.00	0.00*	0.00***	-	0.00	-	-0.00	-0.00	0.00	0.00**	-0.00	-	0.00	5	0	5
Real estate holdings	0.04	-0.01	-	0.05	0.06**	-0.02	0.02	0.07*	0.05	0.12*	-	0.00	-	0.28**	0.00	0.00	0.06	-0.02	-	-0.04	4	0	4
CAPEX and inventories	0.25**	0.02	-	-0.05	0.05	0.04	0.17**	0.03	0.12	0.32**	-	-0.19*	-	0.34	0.25**	-0.00	0.13	-0.10	-	0.01	4	1	3
Growth in CAPEX	-0.00	-0.00	-	-0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	-	0.01**	-	0.01	-0.01	0.01	-0.00	0.01**	-	0.01**	3	0	3
Cash productivity	0.00***	0.00***	-	-0.00	0.00	0.00	0.00	-0.00	-0.00**	-0.00	-	0.00	-	-0.00	0.00**	-0.00	0.00*	-0.00	-	-0.00	4	1	3
Earnings volatility	0.58***	0.27***	-	-0.04	-0.10	0.02	-0.08	-0.40**	-0.23	-0.29	-	0.09	-	0.40	0.51**	-0.18	0.07	0.10	-	0.62**	4	1	3
Revenue surprise	-0.06	-0.04	-	-0.08	0.05	0.18	0.12	-0.02	0.07	0.22	-	0.34**	-	0.01	-0.20	0.28*	0.08	-0.20	-	0.31**	3	0	3
Sales to price	0.01**	0.00**	-	0.00	0.01**	-0.00	-0.00	-0.00	0.00	-0.00	-	-0.00	-	-0.00	0.00	-0.01*	0.01*	0.00	-	0.00	4	1	3
Operating profitability	0.00	0.00	-	-0.01	-0.01**	-0.00	-0.00	0.01	0.01	0.00	-	0.01	-	-0.02	-0.00	-0.01	-0.02*	-0.01	-	-0.05***	0	3	-3
Cash flow to price	-0.07	-0.06**	-	-0.00	-0.01	-0.08*	-0.00	-0.01	-0.02	0.06	-	-0.08	-	-0.13	0.00	-0.07	0.03	-0.01	-	-0.10**	0	3	-3
Zero trading days	-0.04**	0.01	-	0.02	-0.00	0.00	-0.02	0.02	-0.02	-0.02	-	0.00	-	0.03	-0.06**	0.02	-0.12*	0.01	-	-0.01	0	3	-3
Industry-adj sales to price	-0.58*	-0.71***	-	0.11	-0.44**	-0.02	0.46	0.21	-0.22	0.50	-	0.17	-	-0.70	-0.37	0.60	-0.35	0.33	-	-0.22	0	3	-3
Growth in CSE	-0.02*	-0.01	-	0.00	-0.02**	0.00	0.00	0.02	-0.00	-0.00	-	0.01	-	-0.03	0.01	-0.00	-0.01	-0.04**	-	-0.03**	0	4	-4
PANEL B: Image Industries 73 (50) vs. Text 50																							
Current ratio	0.00*	-0.00	0.00*	-	-	-0.00	-	-	-	0.00**	0.00	-	-0.00	0.00	-0.00	-0.00	0.00**	0.00**	0.00	0.00	5	0	5
Depreciation to PP&E	-0.01	0.02**	0.02	-	-	-0.02	-	-	-	-0.02	-0.01	-	0.01	-0.00	0.01	0.04**	0.01	0.03	0.03*	0.04**	4	0	4
Absolute accruals	0.17*	0.23***	0.22**	-	-	-0.09	-	-	-	-0.31**	-0.04	-	-0.04	0.08	0.24*	-0.06	0.04	-0.02	0.08	0.22*	5	1	4
Sales to price	0.01**	0.01**	0.00	-	-	-0.00	-	-	-	-0.00	0.00	-	0.01	-0.02	0.00	0.00	0.01**	-0.01	-0.00	0.00	3	0	3
Quick ratio	0.00	0.00	0.00*	-	-	-0.00	-	-	-	0.00	-0.00	-	-0.00	0.00	-0.00	-0.00	0.00**	0.00**	0.00	0.00	3	0	3
Cash holdings	0.00	0.06**	0.06	-	-	0.01	-	-	-	-0.19***	-0.02	-	-0.04	0.07	0.09*	0.04	0.06	-0.04	0.15***	0.11**	4	1	3
Dispersion in forecasted EPS	0.06**	0.01	0.04	-	-	0.06**	-	-	-	-0.02	-0.00	-	-0.05	0.12*	0.03	-0.00	0.05	0.01	-0.00	0.03	3	0	3
Age	0.00*	0.00**	0.00**	-	-	0.00**	-	-	-	0.00	0.00**	-	0.00	-0.00	0.00	-0.00**	0.00	0.00	-0.00*	-0.00	5	2	3
Cash productivity	0.00**	0.00***	0.00*	-	-	0.00	-	-	-	-0.00**	-0.00	-	0.00	-0.00	0.00***	-0.00	0.00	-0.00	-0.00	-0.00	4	1	3
Earnings to price	-0.07**	-0.07***	-0.07*	-	-	-0.04	-	-	-	0.12**	0.02	-	0.03	-0.02	-0.10*	0.07	0.08	0.08*	-0.02	-0.15**	2	5	-3
Growth in LTNOA	0.05	-0.02	-0.08	-	-	-0.06	-	-	-	-0.28**	-0.10	-	-0.08	-0.20	-0.05	0.13	-0.13	-0.18	-0.22**	-0.37**	0	3	-3
Sales to receivables	0.00	0.00	-0.00	-	-	-0.00*	-	-	-	0.00	0.00	-	-0.00	0.00	-0.00**	-0.00	0.00*	-0.00***	-0.00	-0.00**	1	4	-3
Growth in CAPEX	-0.00	-0.00	-0.00	-	-	0.00	-	-	-	-0.01	-0.00	-	-0.01*	0.00	-0.01*	-0.00	-0.01**	0.01	0.01	0.00	0	3	-3
% change in CAPEX	-0.02*	0.00	-0.01	-	-	0.01	-	-	-	-0.01	-0.00	-	-0.03	0.01	-0.04**	-0.00	-0.02*	0.01	-0.00	-0.00	0	3	-3
Cash flows to price	-0.01	-0.07**	-0.04	-	-	-0.14**	-	-	-	0.11	0.07**	-	0.07	-0.06	-0.12*	0.05	0.03	-0.01	-0.09	-0.11**	1	4	-3
Leverage	-0.01***	-0.01***	-0.01***	-	-	-0.00	-	-	-	0.01**	-0.00	-	-0.00	0.00	-0.02***	0.00	-0.01***	0.00	0.00	-0.01**	1	6	-5