Just Look: Knowing Peers with Image Representation*

Tomasz Kaczmarek† Kuntara Pukthuanthong‡

August 12, 2024

Abstract

What does an industry *look* like? We present a novel approach to assess firm similarity by analyzing four million visuals. Leveraging machine learning, we identify images representing companies' operations, forming Image Firm Similarities (IFS). IFS mirrors investor-defined peer groups and performs competitively against SIC, GICS, NAICS, and text-based similarity, akin to the brain's visual processing superiority. This outperformance appears in pair trading, diversification, and industry momentum strategies. The effectiveness of IFS is attributed to dynamic reclassification and high investor agreement within industries, leading to significant demand and supply effects on stock prices. IFS excels in industries with growth and intangibility.

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

*Code and data are available upon request. We thank Fred Bereskin, Kate Holland, Dhagash Mehta, Mike O'Doherty, and seminar participants at 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

†Tomasz is from the Department of Investment and Financial Markets, Institute of Finance, Poznan University of Economics and Business, Poland. Email: 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. For Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

‡Kuntara is from Trulaske College of Business at the University of Missouri, Columbia, MO, USA. Email: pukthuanthongk@missouri.edu

1 Introduction

What does an industry *look* like? How does image-based industry classification diverge from conventional methods? In an era where visual data plays a crucial role in shaping perceptions and decisions, understanding firm similarity through images offers a fresh perspective. Our study introduces Image-based Firm Similarity (IFS), a novel approach to measuring firm similarity by analyzing four million images. We leverage machine learning techniques to identify graphical objects that best represent companies' operations.

Capital market research frequently requires grouping companies into homogeneous categories based on firm similarities. These industry groups refine the analysis, select control firms, benchmark performance, and provide descriptive statistics. Traditional methods such as the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), and the Global Industry Classification Standard (GICS) are often employed.[1] However, they have significant limitations, particularly in adapting to the rapidly evolving nature of modern businesses.

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 mirrors investor-defined peer groups and, much like the brain's superior visual processing, performs competitively against traditional classification systems such as SIC, GICS, NAICS, and text-based similarity. This competitive performance is evident in pair trading, diversification benefits, and industry momentum strategies. The effectiveness of IFS is attributed to its dynamic reclassification capability and the high level of investor agreement within industries, resulting in significant aggregated demand and supply effects on stock prices. In particular, IFS excels in sectors characterized by 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).

[1]Standard Industry Classification (SIC) codes have been used since 1939. However, they are being replaced with North American Industry Classification System (NAICS) numbers. Simultaneously, the Global Industry Classifications Standard (GICS) system—established jointly by Standard & Poor's (S&P) and Morgan Stanley Capital International (MSCI)—is gaining widespread acceptance. This is especially true among financial practitioners and shown by Bhojraj, Lee, and Oler (2003) to outperform industries grouped by SIC and NAICS. As a result of these modifications, prominent data providers—such as S&P Compustat—now carry all three sets of industry codes. In addition, financial researchers have explored their answer to the industry classification problem, for example, Fama and French's FF algorithm, the text-based industry group from 10-K reports (Hoberg and Phillips, 2016; Lewellen, 2012), the analyst-based industries (Kaustia and Rantala, 2021; Ramnath, 2002).

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.[2] 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) have shown that the parietal cortex processes numeric values, particularly in the left hemisphere. Interestingly, evidence suggests a preference for symbolic notation over other numeric representations, indicating a bias toward more visual representations even within numerical processing (Cohen Kadosh, Cohen Kadosh, Kaas, Henik, & Goebel, 2007; Piazza & Izard, 2009). This preference for visual data extends beyond numbers; images capture attention more effectively and convey complex narratives, emotions, and details more efficiently than text alone (Obaid & Pukthuanthong, 2022).

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.[3] The distinctions between the 45 and 73 classes lie in the granularity of detail they capture, with more classes providing finer distinctions among firms.[4] 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 fluctuations.

We primarily use Google Images for image collection due to its extensive repository of firm-level photos.[5] Although company annual reports could be a source, they typically contain a limited number of high-quality

[2]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.

[3]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

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

[5]While news is another option, they often have limitations, such as a focus on prominent corporations and a predominance of logos over meaningful images. We will explain this in detail in the following footnote.

images, often supplemented by graphs and tables instead of visuals. Google's extensive repository of images aggregates visual data from various market sources, providing an objective and holistic view of a company's product offerings. This market-sourced data is more representative and unbiased compared to company-produced reports. By leveraging our brain's natural ability to process images efficiently, we can build a clearer and more precise understanding of firms' products, enhancing the effectiveness of image-based firm similarity measures.[6]

We introduce a novel algorithm for capturing entity similarities based on diverse image representations, particularly for companies depicted through varied visuals.[7] 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.
- 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

[6] We crawl photos from shareholders' reports and find approximately 13,508 photos for 937 firms per year from 2009 to 2021. The total number of photos is 175,604. We have 1,626,175 photos from Google, or 125,090 photos for 3,458 firms per year. The relatively small number of pictures from firms' annual reports makes applying any machine learning approach impractical. Consequently, we were unable to process photos from shareholders' reports.

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

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). 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, diversifi-

cation, 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. 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.

We also apply the industry momentum proposed by Moskowitz and Grinblatt (1999), short-term 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. 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.

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'

[8]This enhanced agility can be attributed to several key factors. First, IFS rapidly adapts to changes in companies' activities, strategies, and market positions through visual cues. Using advanced machine learning and image processing technologies, IFS analyzes visual content to capture real-time changes in product offerings, branding, and consumer engagement strategies. Second, although we download photos annually and update the model updated biennially, visual data still provides a more immediate reflection of a company's current state. Visual data bypassing the delays associated with updating textual descriptions or waiting for new financial reports. Anyone can upload photos to Google regularly throughout the year and they are available immediately. In contrast, the text and numbers from annual reports are reported with lags and are not available in real time. Finally, the human brain's ability to process images faster than text is reflected in the swift categorization of industry dynamics by IFS based on visual similarities. This enables a more responsive classification system that taps into unstructured data,

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.

While our project aims to develop a new industry categorization metric, we acknowledge certain limitations. First, IFS may not perform well for businesses that are not easily represented by images, such as consulting firms or those dealing with abstract technologies. It is more effective for industries that manufacture tangible products or services. Second, IFS cannot achieve high granularity in traditional classification systems such as SIC codes. While SIC codes can provide 300 to 400 clusters at the 3- and 4-digit levels, IFS becomes unreliable when reaching 100 clusters as the number of firms with defined similarity measure is too small.[10]. This limitation may restrict its ability to offer fine-grained industry classifications. Third, IFS can only be applied to recent years due to its reliance on image availability, restricting its use for historical analysis and long-term trend studies, which may be crucial for certain research applications.

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

offering a richer understanding of industry affiliations and evolutions. IFS contrasts with traditional systems such as SIC, NAICS, and GICS, which rely primarily on textual descriptions and numerical data that can lag behind real-world changes.

[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.

[10]The visual data used in IFS may not capture the nuanced differences required for more granular classifications, leading to potential misclassifications and reduced effectiveness. We define similarities for 1,084-1,109 companies per year, making it irrational to group into more than 100 classes

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 two approaches: transitive and intransitive.

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 et al. (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 suffer from 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 purely 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 compared to 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 that indicate the degree of 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 in this study for the pair-trading application.

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 average of 75 images after cleaning per firm in each 3 year rolling period (see Table 1). Each image is a three-dimensional matrix (224x224x3),[12] which 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 word count of minimum 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 many times a year, allowing IFS to quickly capture changes in market trends and product innovations through

[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.

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

visual cues, making it more responsive to real-time changes compared to 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.

Limitations: Despite its strengths, IFS has limitations. It may not work well for businesses that are not easily represented by images, such as consulting firms or those dealing with abstract technologies. Furthermore, IFS defines peers for around 1,100 firms per year, which in practice does not allow the granularity to exceed 100 clusters, which is achievable for other classification systems. It also relies on the availability of recent images, restricting its use for historical analyses. In comparison, other intransitive approaches can be successfully applied to a broader range of businesses, including those with intangible products, and often provide higher granularity in classifications. They can also incorporate historical data, making them more suitable for long-term analyses.

Investor Agreement and Adaptability: IFS shows high investor agreement within industry categorizations, leading to significant influences on stock prices. This makes it particularly advantageous for investment applications such as pair trading, diversification, and momentum strategies. IFS's visual-based categorization aligns closely with how investors naturally group companies, effectively capturing innate human responses. Although other intransitive approaches also provide valuable information, they may not capture the same level of investor consensus as IFS, as we show in Section 8.

In summary, while IFS offers a powerful and intuitive approach to industry classification through visual data, other intransitive methods provide valuable complementary strengths. The choice between these approaches depends on the specific research or investment application, the nature of the industries studied, and the availability of different types of data. IFS excels in dynamic adaptability and rich data representation, making it particularly effective for capturing the dynamic nature of modern industries. However, other intransitive approaches offer higher granularity and broader applicability, making them suitable for a more comprehensive range of business and long-term analyses.

3 Methodologies and Data

Our study presents a refined approach to industry classification, leveraging the visual data inherent in company product images. This methodology unfolds through stages, initially identifying similarities between companies based on image analysis, then creating industry clusters, and concluding with periodic updates based on new data. This approach provides a dynamic and nuanced understanding of industry relationships, offering a contemporary lens through which to examine firm connectedness in the digital age.

The core of our method begins with calculating similarities between firms, distilled from comparing their respective image sets. This process involves computing the similarity measure between individual images from the sets and then aggregating these measures to determine the overall similarity between the firms. The similarity between two images, a_i from firm A and b_j from firm B, is defined as:

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

where $\text{distance}(a_i, b_j)$ is the distance between images a_i and b_j.

The overall similarity between two firms, A and B, is then quantified by summing the similarities between their respective image sets:

$$S(A, B) = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \text{similarity}(a_i, b_j), \tag{2}$$

where n_A and n_B are the number of images representing firms A and B, respectively.

This similarity score $S(A, B)$ captures the aggregate measure of visual similarity between the image sets of firms A and B, refined through a dimensionality reduction and pairing process to ensure a focus on substantive visual correlations.

After establishing firm similarities, we use a clustering procedure to delineate industries, grounded in the premise that firms with higher mutual similarities suggest a shared industry space. The clustering can be formalized as:

$$C = \text{cluster}(S(A, B)), \tag{3}$$

where C represents the resulting set of industry clusters derived from applying a clustering algorithm to the matrix of similarity scores $S(A, B)$ across all firm pairs. This methodological step is pivotal, as it translates the complex web of pairwise firm similarities into a coherent structure of industries, each defined by the collective visual narrative of its constituent firms.

The dynamic aspect of our methodology allows for the reevaluation and update of these clusters as new images emerge, ensuring that the industry classification remains reflective of the latest market developments. This process underscores the adaptability and relevance of our approach in capturing the evolving landscape of industry affiliations.

In the next part of this section, we discuss the data and technology we use to define similarities between companies. Furthermore, we demonstrate how we construct Image Industries and, finally, explain the methodology to update firms assigned to each industry that sets a backbone for the out-of-sample testing.

3.1 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.[13,14]

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 A.1.

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 A1 in the Appendix A.2 demonstrates a detailed calculation methodology for each ratio.[15]

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 an-

[13]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, dedicated to image. It uses a combination of information from PageRank extended with input from the image to create a ranking of images per query (Brin & Page, 1998; Jing & Baluja, 2008).

[14]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, 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.

[15]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.

alysts 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.[16]

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 3.2. Second, we move to identify yearly updated firms' similarities based on their image representation from all 3-year rolling window periods. Our procedure of peer definition is demonstrated in Section 3.3. 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.[17] 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 3.4.

3.2 Extracting content from image

Defining image industries is meticulously designed to pinpoint objects emblematic of each industry, enabling the grouping of companies based on visual similarities in their associated images. This identification is crucial in distinguishing industries by aligning companies with shared visual characteristics, facilitating a more intuitive and visually coherent industry classification system.

The fundamental step in this procedure involves the comparison of images between pairs of companies to uncover similarities. Figure 1 demonstrates that the comparison is grounded in the extraction of numerical representations from each image, a task accomplished through the application of the VGG19 model (Simonyan & Zisserman, 2014). The VGG19, a convolutional neural network (CNN) model pre-trained on the ImageNet dataset, is renowned for its effectiveness in image recognition tasks. Its architecture, designed for deep image processing, consists of 19 layers, including 16 convolutional layers, three fully connected layers, five max-pooling layers, and a softmax output layer to classify images into 1000 categories.

[16] 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 12.

[17] 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.

The process of feature extraction with VGG19 unfolds as follows:

- Input Processing: Each image is resized to match the VGG19's input dimensions of 224x224 pixels. The input layer adjusts the image to this uniform size to ensure consistency across the dataset.
- Convolutional Layers: The model employs multiple convolutional layers, each applying filters to the image to capture a range of features, from basic edges and textures in the initial layers to more complex patterns and objects in deeper layers. These layers are represented mathematically as:

$$F_{l+1} = \text{ReLU}\left(W_l * F_l + b_l\right) \tag{4}$$

  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.
- Feature Vector Extraction: After processing through the convolutional and pooling layers, the image's representation is flattened and passed through fully connected layers, culminating in a 4096-dimensional feature vector. This vector encapsulates the essence of the image, encoding the visual information necessary to identify similarities between images.

Through this intricate process, we harness the power of deep learning and the rich visual data encapsulated in company-related images to define firm similarities in a manner that transcends traditional classification approaches. This offers a novel perspective on the company grouping based on the visual domain.

3.3 Identification of Firms' Similarity

Building upon the foundation of image comparison through feature extraction via the VGG19 model, our methodology for defining image industries involves refining the comparison process to identify meaningful similarities between firms based on their photos.

Cosine Similarity

The cornerstone of our image comparison process is cosine similarity, a metric that is adept at capturing the orientation differences between high-dimensional vectors. This measure is crucial for assessing the degree of similarity between images, providing a range that signifies complete similarity to total dissimilarity. The

formula for cosine similarity between two vectors, a and b, is given by:

$$\text{cosine_similarity}(a, b) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}, \tag{5}$$

where a and b present the vectors of image i of firm A and image j of firm B. a and b are feature vectors, each representing the visual characteristics of an image associated with a firm including various visual attributes such as color, texture, shape, and other patterns recognized by the CNN.[18]

However, applying cosine similarity directly to the 4096-dimensional feature vectors extracted by the VGG19 model introduces challenges, primarily due to the "curse of dimensionality."[19] This phenomenon, where space becomes sparse in high dimensions, can obscure meaningful distinctions between images. To mitigate this, we implement Principal Component Analysis (PCA) to reduce the dimensionality of our feature vectors, thereby retaining the most significant aspects of the data while eliminating redundancy. PCA achieves this by transforming the original data into a set of linearly uncorrelated variables (principal components), with the transformation defined as:

$$Z = XP, \tag{6}$$

where X represents the matrix of original feature vectors with the dimension of number of images $\times$ 4096, P denotes the matrix of principal components, and Z is the transformed data in reduced dimensions.

We apply PCA to the matrix X to reduce dimensionality while preserving at least 70% of the variability. The exact number of principal components depends on the specific characteristics of the dataset to meet the 70% variability preservation criterion.

Upon reducing the dimensionality, we obtain Z. We then use these reduced-dimensional representations to calculate the cosine similarity between each pair of photos representing company A and B and convert estimated similarities to distances.[20]

After this step, we have a matrix demonstrating distances between all possible combinations of images from company A and B, where the closest images have a distance close to zero and the most distant reach values close to one. According to Equation 2, the overall similarity between two sets of photos requires

[18]The images are processed using a Convolutional Neural Network (CNN), specifically the VGG19 model pre-trained on the ImageNet dataset. The VGG19 model is known for its effectiveness in image recognition tasks. Each image is resized to 224x224 pixels to match the input dimensions required by VGG19. The model applies multiple convolutional layers to capture a range of features, from basic edges and textures to more complex patterns and objects. After processing through the convolutional and pooling layers, the image's representation is flattened and passed through fully connected layers, resulting in a 4096-dimensional feature vector.

[19]The number of 4096 is specific to the architecture of the VGG19 model. In VGG19, the last few layers include two fully connected layers, each with 4096 neurons. The output of the final 4096-neuron layer is used as the feature vector for each image.

[20]The cosine similarity defined in Equation 5 measures similarities between two objects. The cosine distance measure is calculated as 1 - cosine similarity.

summing the similarities between their respective image sets. However, simply aggregating similarities across all possible image combinations can dilute the impact of the most similar photos. This is because when comparing two sets of photos, only the closest pairs demonstrate a meaningful connection between the firms, while other combinations may represent noise.

To concentrate on the closest matches of images, we implement two solutions. First, we apply a threshold to the cosine distance of 0.4, filtering out less relevant matches. This ensures that only the most similar images contribute significantly to the overall similarity score.[21] Second, we combine images from the two sets of photos representing companies A and B into pairs in a way that minimizes the sum of all pairs' distances. We use the Linear Sum Assignment Problem (LSAP) to solve this optimization problem, ensuring that the overall similarity measure reflects the strongest visual correlations between the firms.

Linear Sum Assignment Problem

The Linear Sum Assignment Problem (LSAP) is a combinatorial optimization approach that aims to find the optimal one-to-one matching between two sets, minimizing the total cost.

The assignment problem can be formulated as:

$$\min_{x_{i,j}} \sum_{i=1}^{n} \sum_{j=1}^{m} Distance_{ij}\, x_{ij}, \tag{7}$$

subject to the constraints:

$$\sum_{j} x_{i,j} \leq 1 \quad \forall i \tag{8}$$

$$\sum_{i} x_{i,j} \leq 1 \quad \forall j \tag{9}$$

$$x_{i,j} \in \{0, 1\} \tag{10}$$

where $Distance_{ij}$ is the distance between the i photo of firm A and the j photo from firm B, and x_{ij} is a binary variable indicating whether these photos are matched (Crouse, 2016).[22] By solving this

[21] The threshold ensures that only images with distances below this value (i.e., more similar) are included in the calculation of similarity between two sets of images, while those above it are excluded. We use the 0.4 threshold in the study and verify that the results with thresholds of 0.35 and 0.45 are comparable.

[22] $x_{i,j}$ 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.

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.

The total similarity between two sets of photos can be decomposed into expected similarity and excess similarity. Expected similarity is driven by common photos that frequently appear in sets of many companies, such as iPhones, 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 accurately decompose the total similarity into expected and excess components, we estimate the expected similarity by regressing the total similarity on the number of image pairs between any two firms. This approach allows us to differentiate significant similarities from those that could occur by chance, ensuring that the identified similarities are robust and truly reflective of the companies' visual correlations.

Weighted Least Squares (WLS) Regression

To account for the heteroscedasticity inherent in our data, we employ the Weighted Least Squares (WLS) regression model. Unlike ordinary least squares regression, WLS assigns weights to each observation based on the inverse of their variances, ensuring that observations with greater variability have less influence on the estimated parameters.

The WLS regression model is given by:

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

where Y is the vector of total similarities with dimension $n \times 1$. X is the design matrix with dimension $n \times p$, where each entry represents the maximum number of pairs of images from two sets of images describing

companies A and B. β represents the regression coefficients with dimension $p \times 1$. Σ is the diagonal matrix of weights with dimension $n \times n$. In our regression p is 1 because we have a single explanatory variable.

After estimating the regression coefficients $\hat{\beta}$, we calculate the expected similarity scores $\hat{Y} = X\hat{\beta}$. The difference between total similarity Y and $\hat{Y}$ represents the excess similarity. This excess similarity can be decomposed into spurious and statistically significant components, where the latter indicates a meaningful visual correlation between the firms that exceeds what could be expected by chance.

Significant Excess Similarities

To identify significant excess similarity, we test the hypothesis that the excess similarity is zero ($H0\colon Y - \hat{Y} = 0$), which suggests that any observed similarity is purely random. This hypothesis is tested with a one-sided test at a 99% confidence level.

In practice, we calculate the excess similarity as the difference between the observed similarity Y and the 99% upper confidence interval of the expected similarity $\hat{Y}$.

The significant excess similarity is then calculated as:

$$\text{Excess Similarity} = Y - \hat{Y}_{99\%}. \tag{12}$$

This approach ensures that only similarities significantly higher than what could be expected by chance are considered as excess similarity, thereby confirming robust visual correlations between the firms.

Statistically Significant Similarity Over Time

Our approach evaluates similarity scores for each year, incorporating mechanisms to minimize Type I and Type II errors related to the hypothesis testing.

Type I error:

A company is considered similar to another in year t if they are similar in at least two out of three years $t-2$, $t-1$, t. This ensures that a single-year similarity is not mistaken for a significant trend. We establish a nonrandom connection if the null hypothesis is rejected in at least two of the three last years.

Type II error:

To ensure stability and reduce sensitivity to short-term fluctuations, we implement a rewriting period in our update process. Specifically, we require a pair of companies to maintain similarity from period t to the period $t+1$ once we establish a nonrandom similarity.

Similarity Matrix and Updates:

Using the significant similarities identified, we create a sparse similarity matrix for each year, showing connections between all pairs of companies. Only about 1% of all possible connections are deemed statistically significant.

Let us illustrate this with an example. Consider the similarity evaluation between companies A and B using their photos from 2009-2013. To establish a similarity in a given year (denoted as period t), we require photos from the previous two years ($t-2$ and $t-1$) in addition to the current year (t), allowing us to compute similarities for periods 2011, 2012, and 2013. Let's suppose that the similarities for 2011 and 2013 are determined to be nonrandom, while that of 2012 is random. In this scenario, two of the three similarities between 2011-2013 are deemed nonrandom. Consequently, we consider companies A and B as peers in 2013. To mitigate type II errors and ensure robustness, we extend this peer relationship to the following year, 2014. This approach necessitates five years to define peer groups, prompting us to utilize photos from 2009 to 2013 to establish the earliest similarities in our dataset. The peer groups defined for 2013 are subjected to an out-of-sample testing approach in the subsequent year, 2014.

This method, including the rewriting period, provides a stable classification system that does not overreact to short-term changes while still capturing significant long-term shifts in company similarities. It balances the need for accuracy with the need for stability in classifications.

3.4 Image Industry Definition

The definition of image firm similarities (IFS) is a critical phase in our methodology, designed to cluster firms into industries based on visual similarities. Given the unique characteristics of our similarity matrix, which is highly sparse and composed only of statistically significant similarities, we adopt a two-step approach to address the challenges inherent in clustering with such data.

Scaling Similarity Scores

The first step involves a nuanced scaling procedure for similarity scores. Traditional clustering methodologies often scale all similarities to fall within a 0-1 range. However, given that our similarities are already filtered for statistical significance, a standard scaling would obscure the distinction between firms with the lowest statistically significant similarity and those without statistical similarity. To avoid this issue, we adopt a more tailored scaling approach, where all significant similarities are scaled to range from 0.5 to 1, while pairs deemed not statistically similar are assigned a zero similarity.[23] This scaling preserves the granularity of our

[23]There are a few reasons why we choose this range. First, this approach maintains a clear separation between statistically significant and nonsignificant similarities. Setting the minimum significant similarity to 0.5 creates a distinct gap between significant (0.5) and nonsignificant (0) similarities. Second, scaling from 0.5 to 1 allows for more nuanced differentiation among statistically significant similarities. This preserves the relative strengths of relationships between firms, which could be lost if all similarities were scaled from 0 to 1. Third, by assigning all significant similarities a value of at least 0.5, we emphasize

similarity scores, ensuring that the clustering process reflects genuine, statistically validated similarities.[24]

Clustering Methodology

The second step addresses the challenge posed by our sparse similarity matrix. Clustering involves identifying and assigning similar instances to clusters or groups of similar cases. Traditional clustering methods, such as k-neighbors or spectral clustering, falter in the face of such sparsity (Géron, 2019). Inspired by Hoberg and Phillips (2016), we implement a dedicated clustering methodology. Initially, each company is placed in a singular cluster. We then iteratively measure the distance between clusters, merging the two with the highest similarity at each step. This process continues until we achieve the desired number of clusters. Additionally, after each merging step, we conduct a single reclassification to ensure each company is assigned to its most appropriate cluster, acknowledging that cluster mergers may displace some firms from their optimal grouping.

Dynamic Classification Update

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.

3.5 Out-of-sample setting

To affirm the economic relevance of the IFS and the efficacy of its associated applications, we conduct a rigorous examination in an out-of-the-sample (OOS) setting. Central to our OOS approach is the precise timestamping of all photos utilized, a feature ensured by the Google search algorithm employed for image downloads. This process allows for the collection of photos within specified upload periods, such as querying images of Noble Energy Inc. from 2016 with the criteria "Noble Energy Inc. products after 2016-01-01 before

these relationships in subsequent analyzes. This can be particularly important in clustering algorithms, where higher similarity scores have greater influence. Lastly, setting nonsignificant similarities to zero effectively removes noise from the data, allowing clustering algorithms to focus solely on meaningful relationships.

[24] If a similarity score is 0.20, it would be considered not statistically significant. As per our approach, all non-significant similarities are assigned a value of zero.

2016-12-31," ensuring that the photos retrieved could not have been uploaded later than 2016 to protect against any look-ahead bias. However, it should be noted that images taken before 2016 and uploaded within the year do not compromise our analysis due to the absence of a look-ahead bias.

Each image within our database is tagged with the year it was made available on the internet. As detailed in Section 3.3, the similarity between two companies for year t is determined based on images from years $t-2$, $t-1$, and t. Moreover, this assessment indirectly incorporates images from years $t-3$ and $t-4$ to mitigate Type I and Type II errors. Therefore, we employ a classification that is delayed by one period to conduct our economic tests. For instance, when evaluating our classification against market or financial data from 2014 to 2015, we utilize firm similarities and IFS classification for the year 2013, which was formulated based on images from 2009 to 2013. Consequently, uploaded data from 2014 to 2015 was not used to test IFS performance. This OOS procedure is methodically repeated across four cycles: we predict IFS for 2016 to 2017 using photos from 2011 to 2015, 2018 to 2019 with images from 2013 to 2017, and 2020 to 2021 with pictures from 2015 to 2019. This comprehensive OOS framework supports the results presented in the subsequent sections, ensuring a robust validation of our findings free of predictive bias.

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 and across industries and are visualized in Figures 2, 3, 4, and 5.

For Exxon Mobil Corp, the 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.

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 2. 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 build the Image Industries with a large set of photos. Table 1 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 certainly beyond the scope of this research.[25] Nevertheless, we wish to indicate several observable characteristics even after looking at only dozens of images.

Figures 6 and 7 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, where one class collects visually similar objects across different industries, rather than identical ones within a single industry. If the clustering procedure only aggregated identical objects, the classifications would be overly narrow and exclude products with slight variations in shape or appearance.

In contrast, Figures 8 and 9 highlight industries with lower inter-industry correlations, demonstrating greater object diversity within each industry. Each of the six industries depicted in these figures contains a wide variety of 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

[25] 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.

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.

Next, we shift our attention from the micro level to the bigger picture. Table 2 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.[26] 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, however, encompasses the number of industries with a minimum of five companies at the time of the first industry definition.[27] 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, where calculations require a minimum of five companies per industry.

Table 3 presents the number of industries and the number of companies for the classifications used as benchmarks in our tests. The image industries constrain the number of companies in other classifications to standardize the research sample.

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 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:

$$\mathrm{v}ble_{i,t} = \alpha + \beta vble_{ind,t} + \varepsilon_t, \tag{13}$$

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)

[26] Image Industries classify on 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.

[27] Table A2 demonstrates descriptive statistics of image industries defined in 2013. Industries marked in bold have at least five stocks.

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-side 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 A1 in the Appendix shows details of ratios calculation. Based on market information, we calculate $vble_{ind,t}$ monthly frequency for ten indicators. We estimate the quarterly industry averages for 16 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 4 and 5 present pairwise results comparisons derived from the *same* set of firms classified using Image Industries and other classification techniques. Table 4 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 Tables A3), 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, 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 Tables A5), the ranking of Image Industries remains consistent.

We compare ratios based solely on accounting information, as reported in Table 5. 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 outperform GICS in 63% of cases.

Considering all firms (see Table A4), 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 A6). 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 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 economic insights and sample applications of the proposed image-based classification.

5 Image Clustering: Economic Insights

In this section, we want to understand which industries and firm characteristics are better clustered by images compared to text. We use the text-based measure from (Hoberg & Phillips, 2016) as a benchmark, as it provides both transitive and intransitive measures and relies on product offerings, similar to the IFS. Besides, text and images are often presented together in corporate communications. We also examine the frequency of

firm reclassifications across different industry classification approaches, which impacts applications discussed in the next section.

5.1 Which industry do images cluster well?

To determine which industries are better clustered by images versus text, we choose firms that have HP and IFS classification and regress the difference in R^2 between Image-based Firm Similarity (IFS) and HP for each financial ratio on industry dummy variables, using the industries defined by Moskowitz and Grinblatt (1999). We analyze data from 2014 to 2021, excluding the "Other" industry as the reference category.

Table 6 presents the results of this analysis. Our findings suggest that IFS performs better in industries with tangible, visually distinctive products and complex manufacturing processes. Specifically, the industries where IFS excels are manufacturing, food, chemical, and fabricated metals. These industries typically have products easily recognizable in images and often involve complex operational processes that visual data can capture effectively.

The industries where text-based measures outperform include the financial industry, apparel, retail, and electrical equipment. These industries tend to be more service-oriented or have products that may not be as visually distinctive. They rely more on textual descriptions of services, brand identities, or technical specifications.

This analysis suggests that the effectiveness of image-based versus text-based classification methods depends on the nature of the industry and its products or services. Industries with high visual distinctiveness and operational complexity benefit more from image-based classification. At the same time, those with a stronger emphasis on services, branding, or technical details may be better represented by text-based measures. These findings highlight the complementary nature of image-based and text-based classification methods, indicating that each may be more effective for different companies or industries. This underscores the potential value of incorporating visual and textual data into industry classification systems to capture a more comprehensive view of firm similarities across diverse sectors.

5.2 Which characteristics of firms do images cluster well?

This section examines which firm characteristics the IFS can effectively cluster. For each ratio, we regress the difference in R^2 of IFS and R^2 of HP of the firms that have the data in both our and HP samples on the characteristics used in Green, Hand, and Zhang (2017). For each firm's characteristic, we take an average across time for each firm's characteristic as an independent variable in this regression.

Initially, we aim to conduct three-year rolling regressions to capture the time variation of accounting—and

market-based firm characteristics. However, this approach yields only twelve observations per characteristic, which is insufficient for deriving accurate slope coefficients. Consequently, we opt for a single regression using one observation and one R^2 value, utilizing the average of characteristics over the study period.

Table 7 shows that companies with more complex accounting practices (higher *absacc*) better short-term liquidity (higher *currat*) and more efficient cash use (higher *cashpr*) are more accurately grouped by image-based methods. Image classification better captures businesses with sales to price products or services (higher *sp*), suggesting that visual cues are more effective for premium or luxury goods. Companies with lower profitability relative to their stock price (lower cfp) are more accurately classified by images, indicating that visual data might better represent firms with growth potential or intangible assets not reflected in current cash flows.

In general, IFS excels for companies in traditional industries or those with easily visualized products. In addition, it is particularly effective for firms with complex operations, strong liquidity positions, and those focused on premium products or services.

The method may capture aspects of businesses that are not easily conveyed through financial metrics alone, such as product quality, brand image, or market positioning. This analysis highlights the complementary nature of image-based and text-based classification methods, suggesting that each may be more effective for different companies or industries. It underscores the potential value of incorporating visual data into industry classification systems, especially for companies where traditional financial metrics may not fully capture their market position or growth potential.

6 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 assets, 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 2 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.[28] 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.

6.1 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

[28] 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.

apparent similarities between companies, potentially uncovering more profitable pair trading opportunities.

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.[29] 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 8, 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.[30] 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.

6.2 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

[29] 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.

[30] 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.

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.[31]

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 9 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 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 lowest, respectively. Although we often emphasize that reasonable industry classifications should provide distinct industries with low correlation across sectors, our findings highlight that the ultimate goal of portfolio optimization is to maximize risk-adjusted returns. In this regard, IFS demonstrates its value by enabling the construction of portfolios that achieve this goal, even if not through the expected mechanism of minimizing the standard deviation.

[31] 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.

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

6.3 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) construct 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 one-month return and holds them for six, nine, or 12 months.[32]

Table 10 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

[32]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).

portfolios. Table A7 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 11 demonstrates the simulation results, where Sharpe ratios of Image Industry 45 are higher than those of any random portfolio.

In conclusion, IFS likely outperforms other methods because it captures unique visual cues that signify economic links and operational similarities between firms, generating stronger momentum profits. Unlike textual descriptions and quantitative information, images can provide a more immediate and intuitive understanding of a firm's products, services, and operational environments. These visual data can highlight similarities and differences that might not be evident through text and numbers alone, leading to more accurate groupings of firms with comparable economic activities.

One of the critical insights from Hoberg and Phillips (2018) is that shocks often drive momentum profits to less visible peers within the industry. Image-based classifications seem to excel in identifying these less visible economic links by capturing subtle visual details that might be overlooked in traditional or text-based classifications. These details can include specific production processes, product designs, and other visual elements that signify a firm's operational and strategic positioning. By incorporating these nuanced visual elements, image-based classifications can better identify and group firms likely to experience similar economic shocks, thus improving the predictability and profitability of industry momentum strategies.

Moreover, Hoberg and Phillips (2018) highlight the robustness of their TNIC approach compared to traditional SIC codes, which are often criticized for their broad and sometimes outdated classifications. Similarly, IFS provides a level of precision and contemporaneity that is difficult to achieve with static textual classifications, as Table 12 shows that image-based classifications are the most dynamic compared to other classifications. Images are rich in information and can be continuously updated to reflect the current state of firms and industries. This dynamic nature of image-based data ensures that the classifications remain relevant and accurate over time, leading to more robust momentum strategies.

7 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 12 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.

8 Underlying mechanism

The industry categorization significantly influences stock trading, contingent upon how market players discern their peers. An optimal industry classification should manifest a pronounced consensus (or lack thereof) concerning stocks within the same (or different) industries. The more cohesive the agreement within one sector, the more pronounced the influence of aggregated demand and supply on stock prices, amplifying the

efficacy of any market-based industry categorization application. This principle underpins the application of industry categorization based on market pricing (Diermeier, Ibbotson, & Siegel, 1984; Ibbotson, Diermeier, & Siegel, 1984; Merton, 1973).

Section 6.1 illustrates that image-based similarities facilitate the construction of highly profitable pair trading strategies. Sections 6.2 and 6.3 demonstrate that the image industry classification surpasses all other schemes in delivering the most substantial portfolio diversification benefits and profitability from the industry momentum strategy. This Section delves into the underlying theories and presents empirical evidence supporting these assertions.

To illustrate, when a company releases any information, the market responds by adjusting its stock price and the stock price of other peer firms. However, investors employ diverse methods to determine each company's comparables, and their investment choices are not uniform. Consider two different investors. The first one thinks that firm 1 has peers in companies 2 and 3. The second one has a different opinion and believes that firm 1's competitors are companies 2 and 4. Now, firm 1 discloses positive information that might result in a favorable position for all companies in company 1's industry (or company 1's peers). Therefore, the first investor can acquire shares in firms 1 and companies 2 and 3 that are comparable. The positive news may lure the second investor; however, their investment selections will differ. The second investor may purchase shares in companies 1, 2, and 4. The demand for shares in companies 1, 2, 3, and 4 will vary because the first and second investors would evaluate their peers differently. Specifically, the demand for shares of businesses 1 and 2 will arise from the combined purchase choices of the first and second investors. However, the demand for shares of companies 3 and 4 will be lower than those of firms 1 and 2 due to the first and second investor's purchasing decisions, respectively.

We can extend this inference and say that different views about peers by two investors create two separate industry classifications. One classification has a first group consisting of stocks 1, 2, and 3 and the other with stocks 1, 2, and 4. However, our investors do not limit their peers' perceptions only to those four stocks; they have a view on peers for many firms. To generalize, the views of Investor A on peers define classification X, while the views of Investor B define classification Y. In addition, the perception of these investors is not unique. n investors share the same opinion on industry classification as investor A (we can define it as group A_n), and m investors share a view of investor B (group B_m).

Now, let us look into firms classified into any of the two industries of classification X. Investors A_n have a homogeneous demand-supply for shares within the same sector but a heterogeneous one for claims classified into two distinct sectors. This results in a higher agreement (disagreement) about stocks ranked in the same (different) industry. Furthermore, the higher the aggregated demand/supply from investors A_n, the higher the market agreement (disagreement) when stocks are classified into the same (different) industry.

This phenomenon is a foundation of the rationale for any industry classification application based on market prices (Diermeier et al., 1984; Ibbotson et al., 1984; Merton, 1973).

Comparing the quality of classifications X and Y should be based on their abilities to capture aggregate demand/supply from investors A_n and B_m. Suppose the perception of A_n and B_m is more congruent. In that case, the classification will significantly impact stock prices and generate good returns from the market-based price application—as shown in Sections 6.1, 6.2 and 6.3. Referring to the relationships we described earlier, the aggregated demand/supply related to classification based on the image is high. The following Section attempts to provide the reasons for this.

8.1 Cognitive psychology and the concept of similarity

According to the theory of cognitive psychology of similarity, investors A_n and B_m disagree on firm connections because we cannot consider similarity a universal concept.[33] Therefore, it requires a frame of reference. Goodman (1972) argues that the similarity comparison process systematically fixes respects. This concept states that two things share an arbitrary number of predicates and differ from each other in a random number of ways. Therefore, searching for similarity requires non-arbitrary assumptions to constrain the predicates.

8.2 Respects of firms similarity

As similarities can be ambiguous, representing market industries has no clear solution. The solution depends on the relative perception of similarity as a concept that links peers into groups and on the desired application. From this perspective, each industry classification can only be one approximation of how different investors define similar companies. Meanwhile, the best industry classification is the one that is the most useful in a particular application. Referring to our earlier relationship, if a demand/supply related to industry X is higher than that of Y, then industry X is better than Y for applications that benefit from it. Furthermore, if all existing industry schemes (e.g., NAICS, SIC, GICS, or our image industry classification) only approximate classifications defined by the market participants, we should consider the aspects that drive market participants to define industries.

Not surprisingly, alternative aspects of firm similarity may have a dominant role in determining investment behavior. Hoberg and Phillips (2016) show that a significant market participant compares firms' product offerings to identify peers. We agree that considering products to measure firms' similarity is intuitive. After

[33] According to Tversky's featural theory of similarity (e.g., Gati and Tversky (1984) and Tversky (1977)), Murphy and Medin (1985) noted that "the relative weighting of a feature (as well as the relative importance of common and distinctive features) varies with the stimulus context and task, so there is no unique answer to the question on how similar is one object to another" (p. 296). Even more bluntly, a philosopher Goodman (1972, p. 437), who called similarity "invidious, insidious, a pretender, an imposter, a quack." Goodman says that the similarity of A to B is an ill-defined, meaningless notion unless one can say "in what respects" A is similar to B.

all, a company's cognitive process is knowing what products are offered to customers. However, determining similarities is not universal; evaluating the products' similarities may differ depending on the cognitive sources. Hoberg and Phillips (2016) argue that a valuable source of such information is the description of the company activity defined by the management and is available in the business description section of annual 10K reports. We agree that reading this description may be the basis for some investors to learn about the product offering. However, we argue that these descriptions are not a precise and common source of information to define the companies' product offerings. First, text descriptions in the 10K report usually have a broader context than a company's product offering. It also relates to, e.g., company history, size, competitive advantages, or suppliers; it is a mix of information describing a company. Therefore, drawing universal conclusions about product offerings based solely on this description is tricky. Second, a subjective management description often underlines only the company's advantages. Third, many unprofessional market participants are unaware of a section like that. After all, a business description of, e.g., Apple, Ford, or Coca-Cola, is not a common source of information about their product offerings.

We also argue that illustrations of offered products build a more familiar, objective, and precise picture of the companies' product offerings than management descriptions. First, photos and movies dominate human cognitive processes, related to the widespread use of image-based communication based on television, the internet, and mobile devices (Branthwaite, 2002; Dewan, 2015). For instance, most investors are likely familiar with numerous photos representing products from companies like Apple, Ford, and Coca-Cola. Second, various sources of images, such as Google, provide a comprehensive representation of companies' product offerings. Unlike 10-K reports, which firms produce, images from Google are market-sourced and thus suitable for asset pricing applications, offering an objective and holistic view of a company's products. Third, scientists discovered that our brain has excellent abilities regarding image processing. We process images up to 60,000 times faster than text (Potter et al., 2014). In an age where an overload of information is delivered to us, using the natural abilities of our mind increases our efficiency by building a clear image of companies' product offerings. Consequently, all these features make the image an important respect of firm similarity and explain the mechanism influencing the high application value of our image industry classification.

Finally, we demonstrate the empirical evidence supporting our theory about images' importance in determining firm similarities. The R^2 of FORECASTED MONTHLY RETURN from Table 4 is very high relative to other classifications with similar granularity. This suggests that image industries agree with investors' expectations of future stock returns, meaning there is a high demand/supply from investors who share common expectations about firms' peers aggregated with the image. This relation is the foundation for the benefits of image-based portfolio diversification evaluated in Section 6.2.

The dispersion of analysts' EPS estimates across stocks captures the rate of investors' agreement. The lower the dispersion of analysts' estimates, the higher the investor agreement about stocks in the industry. Figure 10 visualizes that Image Industries 45 has the highest agreement rate for stocks classified into a single sector. This is consistent with results observed for pair trading, industry diversification benefits, and industry momentum, where Image Industries 45 delivers the highest Sharpe ratios.

9 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

Crouse, D. F. (2016). On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, *52*(4), 1679–1696. doi:10.1109/TAES.2016.140952

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

Diermeier, J. J., Ibbotson, R. G., & Siegel, L. B. (1984). The Supply of Capital Market Returns. *Financial Analysts Journal*, *40*(2), 74–80. doi:10.2469/faj.v40.n2.74

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.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. doi:10.1016/0010-0285(84)90013-6

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Incorporated

Goodman, N. (1972). Seven Strictures on Similarity. In *Problems and projects*. Bobs-Merril.

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.

Ibbotson, R. G., Diermeier, J. J., & Siegel, L. B. (1984). The Demand for Capital Market Returns: A New Equilibrium Theory. *Financial Analysts Journal*, *40*(1), 22–33. doi:10.2469/faj.v40.n1.22

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

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.

Merton, R. C. (1973). An Intertemporal Capital Asset Pricing Model. *Econometrica*, *41*(5), 867–887

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

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. doi:10.1037/0033-295X.92.3.289

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

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352. doi:10.1037/0033-295X.84.4.327

Table 1: 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 2: 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 2: (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 2: (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 3: 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
Text 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 4: 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 A1 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 5: 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 A1 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 6: 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 A1 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 7: 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 A1 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

Table 8: 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 A1 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 9: 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 10: 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 11: 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 12: 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

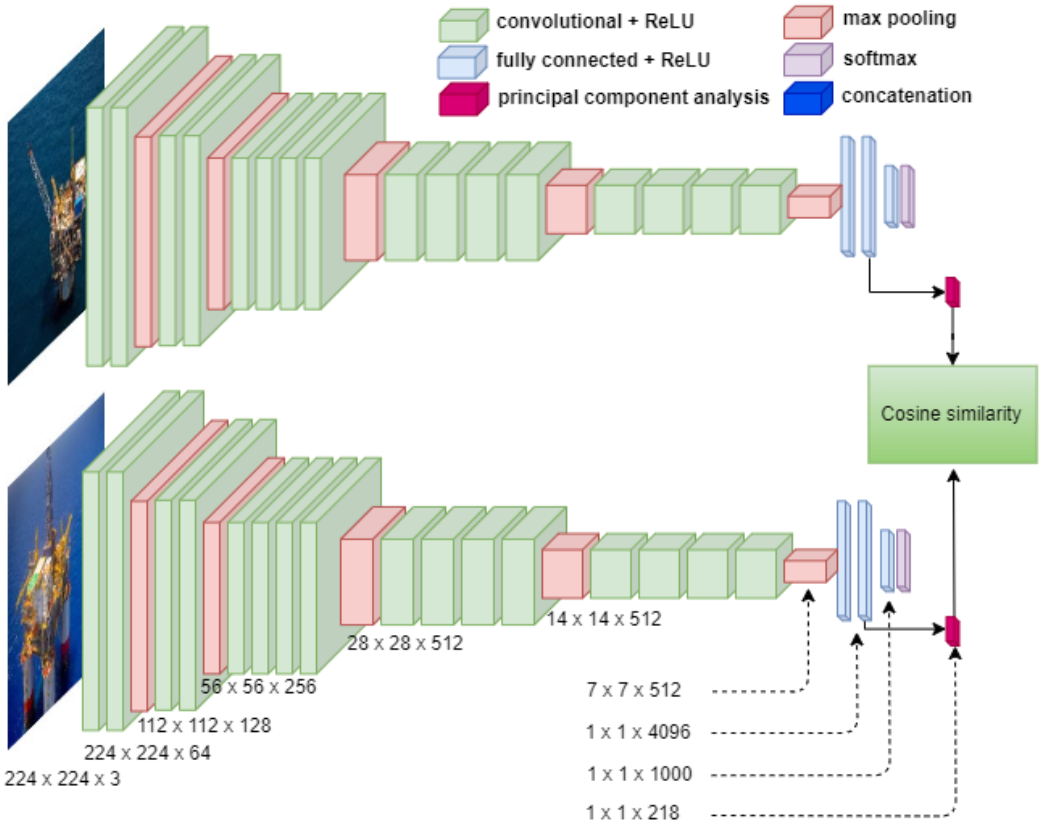


Figure 1: Image similarity

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.

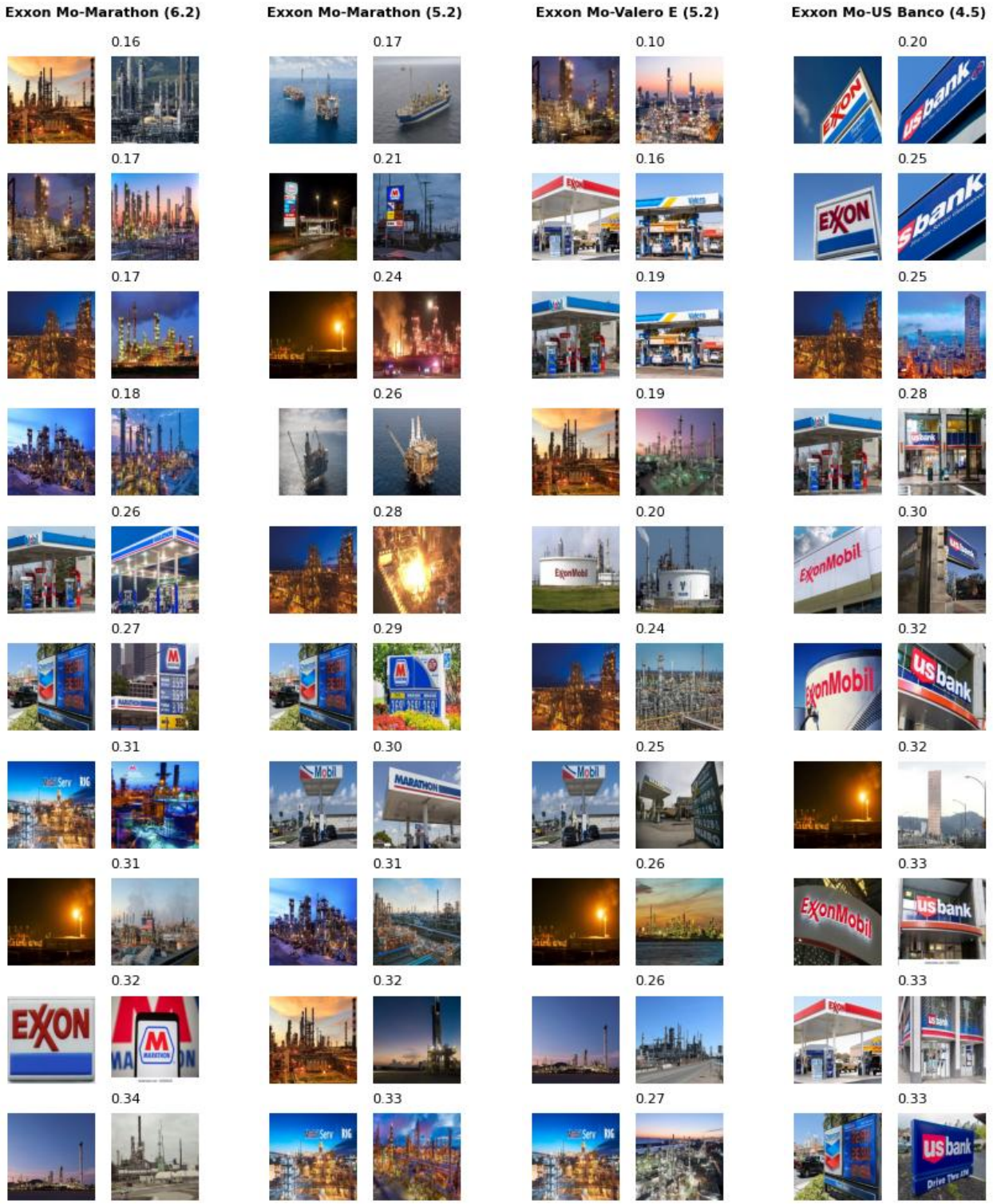


Figure 2: Photos representing four peers of Exxon Mobil Corp with the highest similarity score

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.

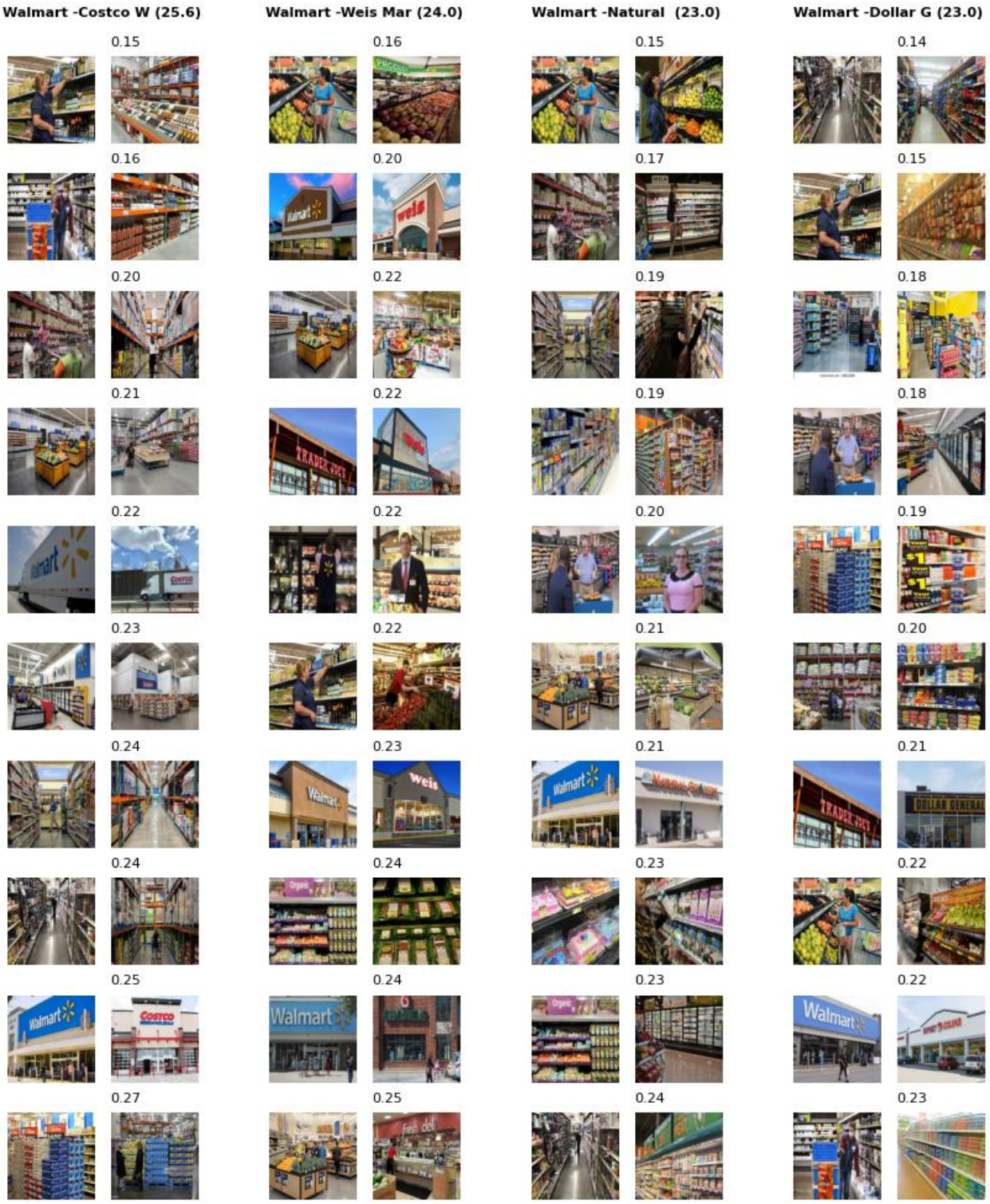


Figure 3: Photos representing four peers of Walmart Inc with the highest similarity score

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.

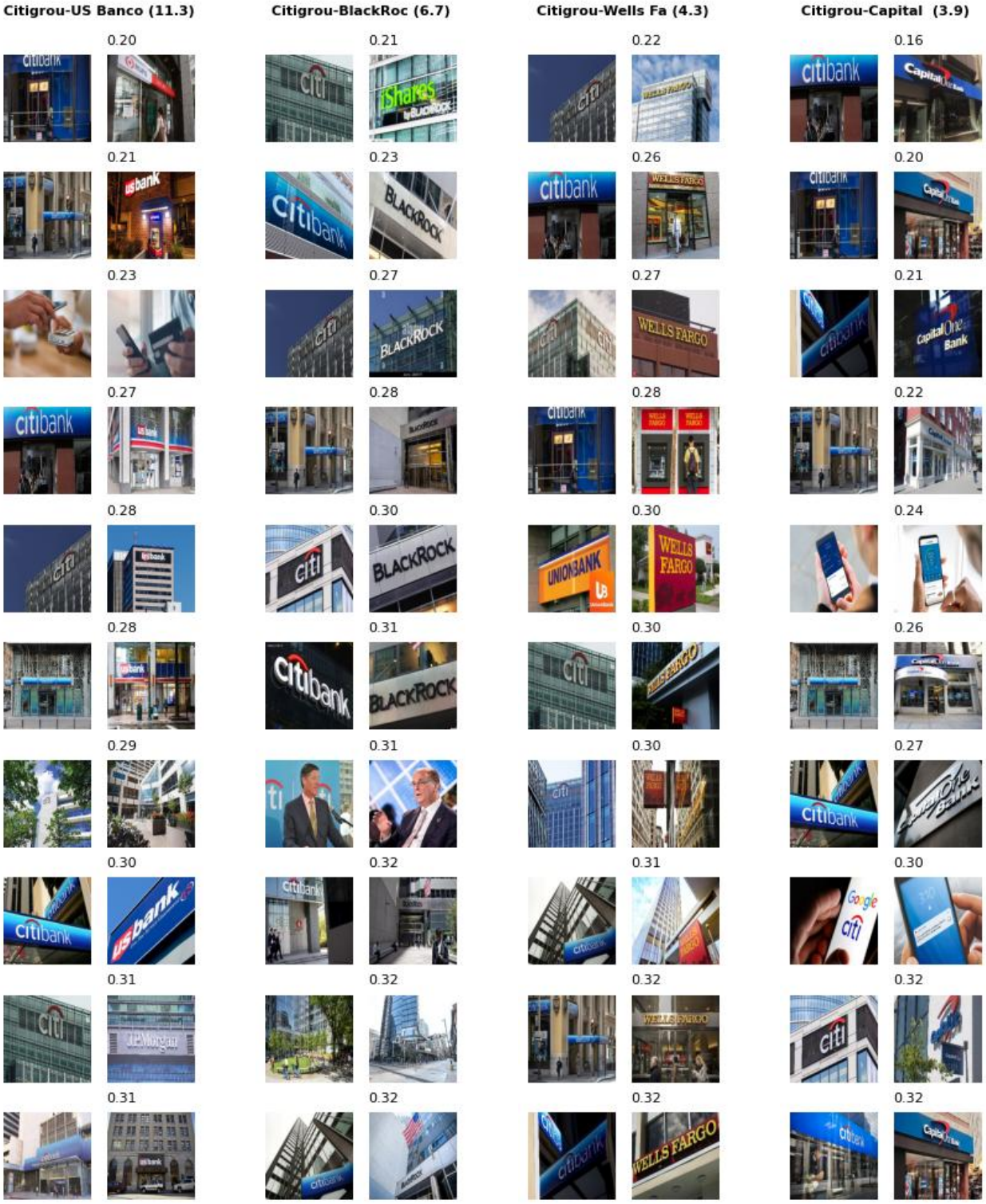


Figure 4: Photos representing four peers of Citigroup Inc with the highest similarity score

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.

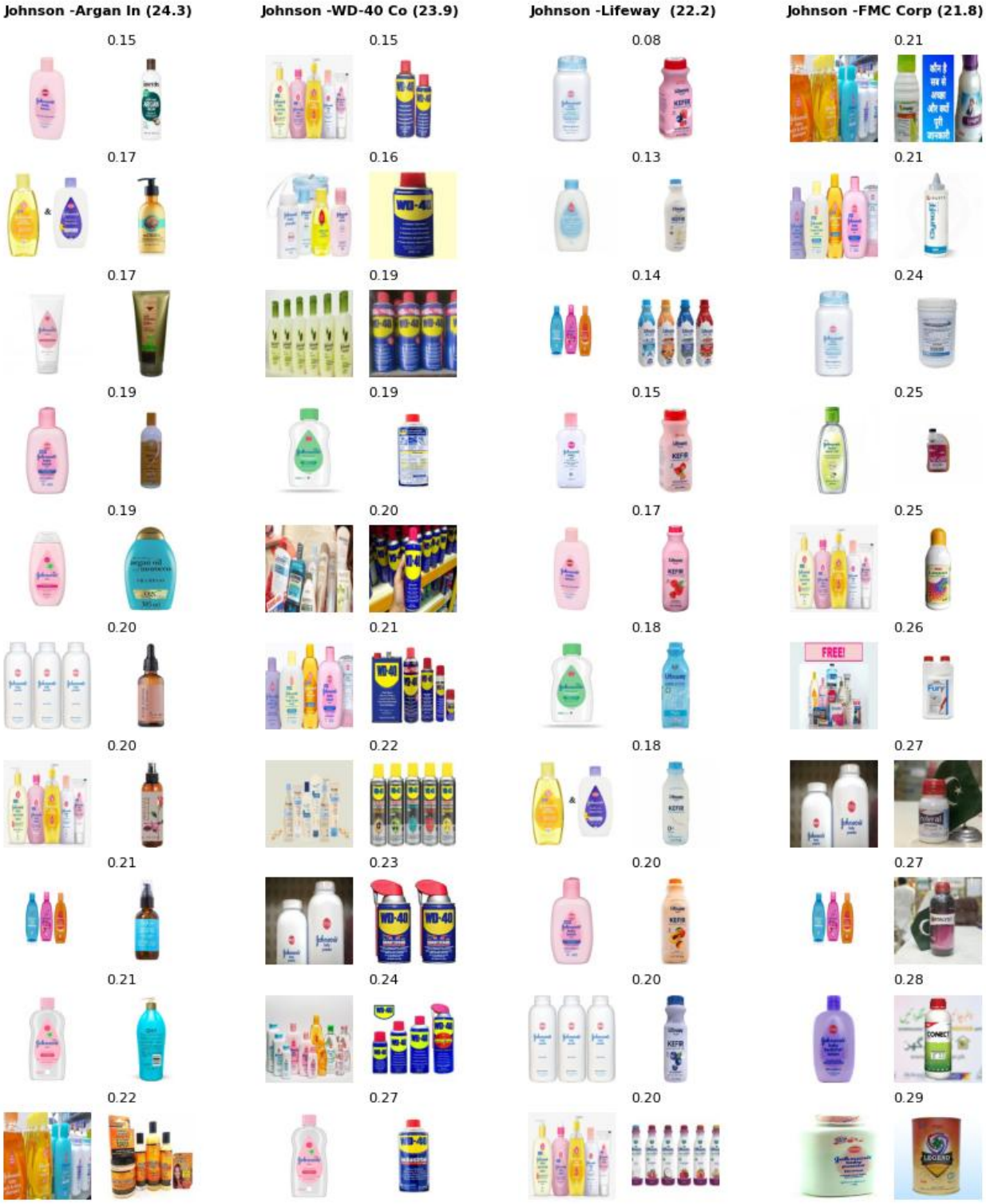


Figure 5: Photos representing four peers of Johnson & Johnson with the highest similarity score

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.

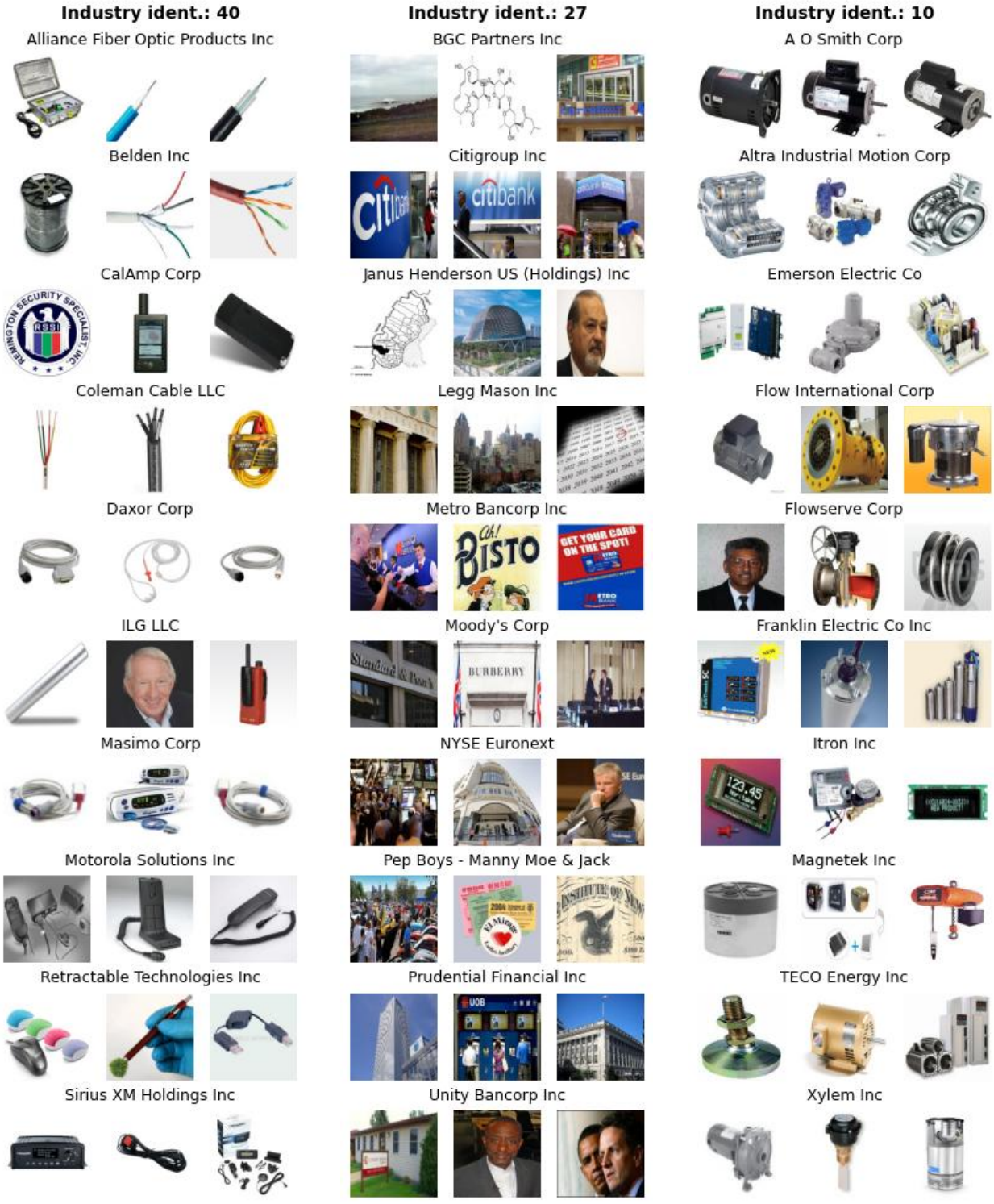


Figure 6: Photos representing industries with highest inter-industry correlations (1)

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.

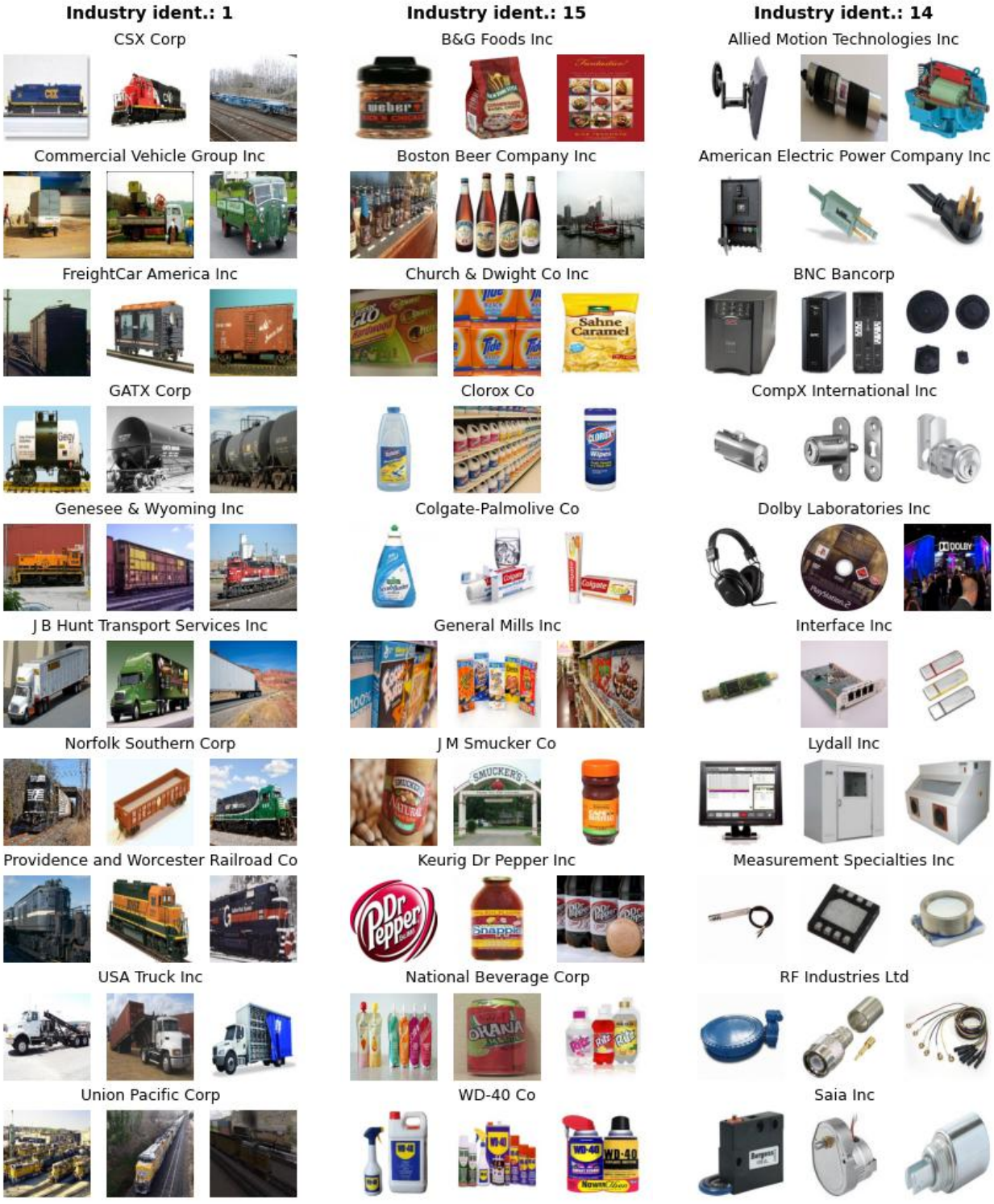


Figure 7: Photos representing industries with highest inter-industry correlations (1)

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.

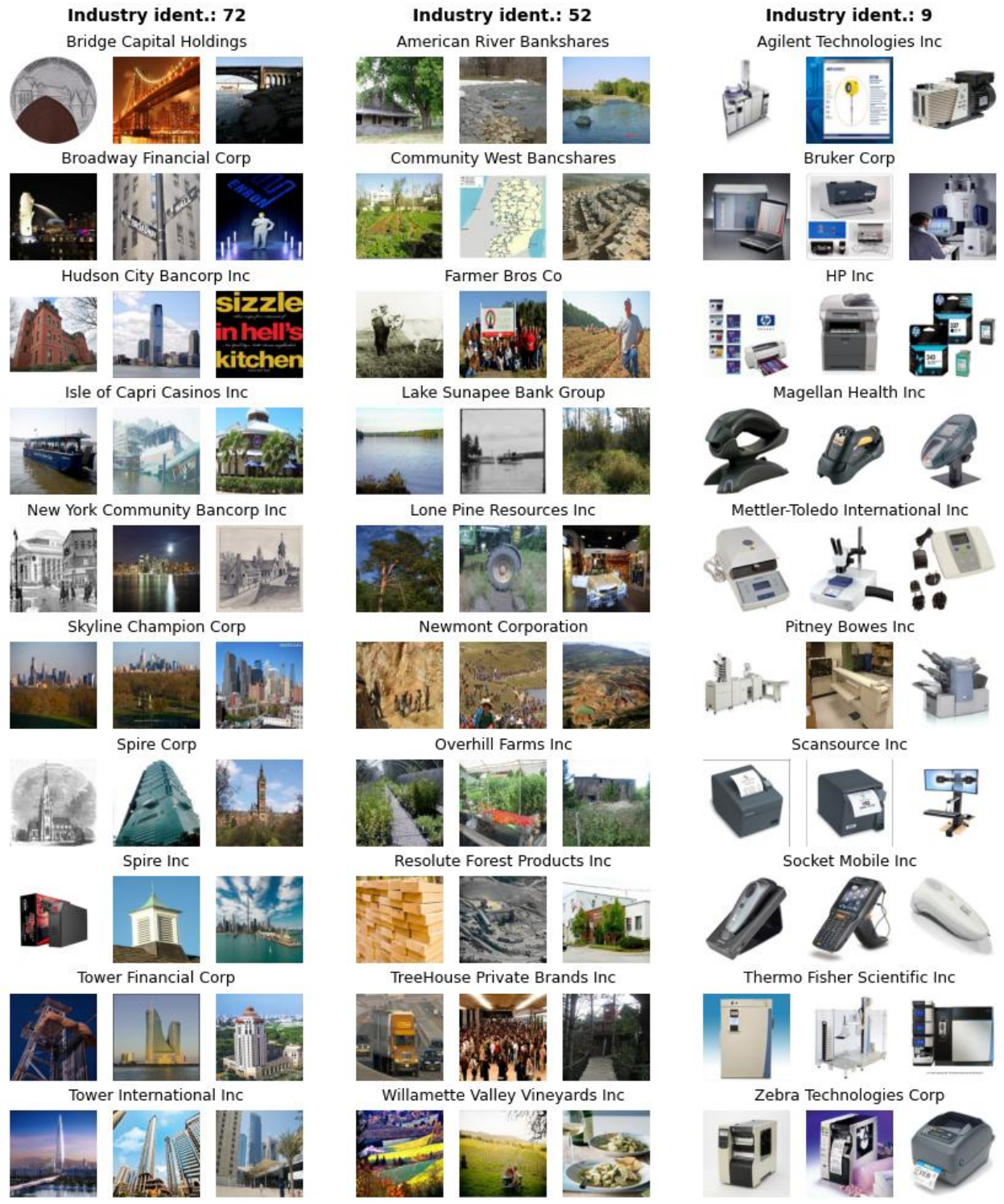


Figure 8: Photos representing industries with the lowest inter-industry correlations (1)

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 9: 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.

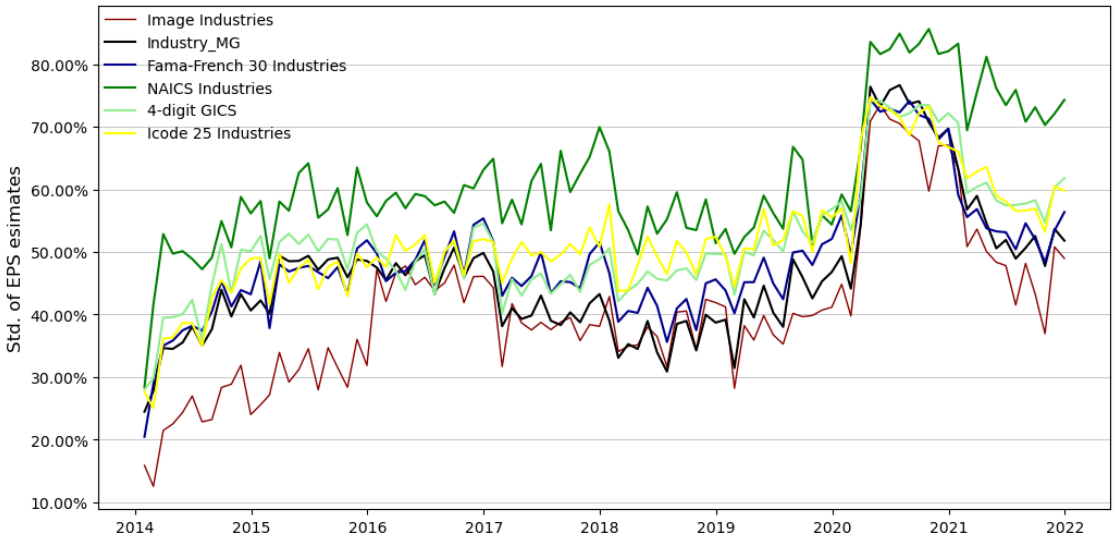


Figure 10: Investor's agreement

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.1 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 that are not related to the firm's business operations. This step is pivotal in the data cleaning process and accounts for the largest 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 images per year. This refined dataset forms the basis for our subsequent analysis, ensuring that the visual data accurately reflects the business activities and characteristics of the firms.

A.2 Tables and Figures

Table A1: 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 A1: (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 7			
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 A1: (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 A2: 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 A2: (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 A2: (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 A3: 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 A1 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 A4: 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 A1 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 A5: 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 A1 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 A6: 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 A1 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 A7: 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