



CapTech[®]

Others Talk, We Listen.

Image Recognition Services: **Searching for Value Amid Hype**

Contributing Researchers:

Jack Cox, Fellow; Chris Heinz, Senior Consultant; and Kevin Vaughan, Manager

Independent research conducted by CapTech Ventures, Inc.

Information is based on best available resources. CapTech[®] is a registered trademark for CapTech Ventures, Inc. All other trademarks are the property of their respective companies.

For more information, go to www.captechconsulting.com.

View infographic at: <https://www.captechconsulting.com/blogs/help-or-hype-an-unbiased-image-recognition-services-vendor-assessment>

Table of Contents:

3	Executive Summary
3	Functions Studied
4	Approach to Study
4	Major Takeaways
5	Test Results and Observations
5	Product Identification
8	Custom Item Recognition
9	Item Categorization
10	Logo Recognition
12	Facial Detection
14	Facial Recognition
15	Mood Analysis
17	Text Recognition (OCR)
18	Adult Content Detection
19	Vendor Summary
22	More About Costs
23	Conclusion
24	Appendix

Executive Summary

With interest in artificial intelligence (AI) services on the rise, image recognition services promise to drive new capabilities and efficiencies while transforming customer service across diverse industries. Marketers of these services have already pitched them for a wide range of imaginative uses:

- As people enter an upscale retail store, image recognition services could immediately identify high-value customers and gauge their moods.
- The state highway department could quickly review images of thousands of miles of roads to identify potholes and prioritize repairs.
- Manufacturers might assess photographs of the current inventory and rapidly determine what needs to be re-ordered.
- Law enforcement agencies could measure crowd sizes and identify people believed to present security risks.
- Customers of an e-commerce company could take pictures of a product and submit them to the retailer, whose app could locate similar products and allow the customers to purchase them immediately.

Given these and other dramatic possibilities, *image recognition*, a branch of artificial intelligence and machine learning, is rapidly gaining the attention of businesses and government agencies. But are the image recognition services that are commercially available today capable of delivering on their promises?

To cut through the hype and understand the utility these services truly offer, CapTech tested the six leading image recognition services: Amazon Rekognition, Microsoft Azure Computer Vision, Clarifai, CloudSight, Google Cloud Vision, and IBM Watson. This paper summarizes the results of the tests.

We presented each service with the same set of approximately 4,800 images, distorting many by blurring, overexposing or underexposing, positioning the images at odd angles, and otherwise recreating real-world conditions. We evaluated them across performance and the services' confidence in nine distinct areas of function including adult content detection, facial detection, facial recognition, mood analysis, text recognition, logo recognition, branded product identification, item classification, and item recognition. We also evaluated overall correctness and average response time of the services.

We found that no one service has a clear lead across all tested functions. With no one-size-fits-all solution available on the market today, we recommend that organizations looking to adopt image recognition be prepared to use multiple vendors to accomplish their strategic goals and objectives. In addition, because of the rapid pace of change in this industry, we recommend that integration of existing systems with image recognition services be architected to provide maximum flexibility so that organizations can switch vendors as needed and adapt to the rapidly changing image recognition landscape.

Functions Studied

CapTech assessed nine common image recognition functions:

- **Branded product identification:** identifying the presence of branded items within an image; for example, a Coca-Cola can or a Levi's jeans label
- **Custom item recognition:** identifying the presence or absence of items unique to a company or other organization; for example, a pack of cigarettes or a roller coaster
- **Item categorization** (also referred to as label detection): identifying classes of items such as dogs or flowers within an image
- **Logo recognition:** identifying the presence of a brand logo in an image
- **Face detection:** reporting on the position of faces within an image
- **Facial recognition:** identifying people whose faces appear in an image
- **Mood analysis:** determining the mood of a person in an image
- **Text recognition** [also referred to as Optical Character Recognition (OCR)]: reading and converting digital text, whether handwritten or typed, to machine-readable text
- **Adult content detection:** determining if an image contains sexually graphic content

Cost Modeling

In addition to assessing the accuracy of the leading image recognition services, we determined the price of each service across a common set of functions. Google, Amazon, and Azure were similarly priced. IBM Watson was twice as costly as those three, and CloudSight was 30 times as costly. For additional details, please see the More About Costs section below (page 23).

Approach to the Study

We gathered a set of images that contained content suitable for each of the nine functions we sought to test. In addition to using original clean images, we resized, filtered, and distorted images to recreate conditions that the systems will encounter in the real world. The filters included overexposure, underexposure, and motion blurs. We also distorted images by rotating or skewing them. We used the same images to test each service.



Figure 1 - Original Mood Analysis

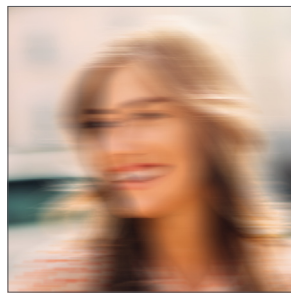


Figure 2 - Motion Blurred

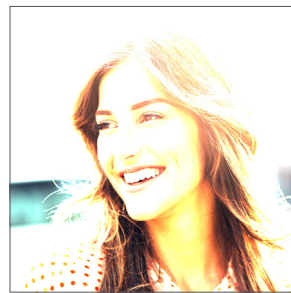


Figure 3 - Overexposed



Figure 4- Rotated Image

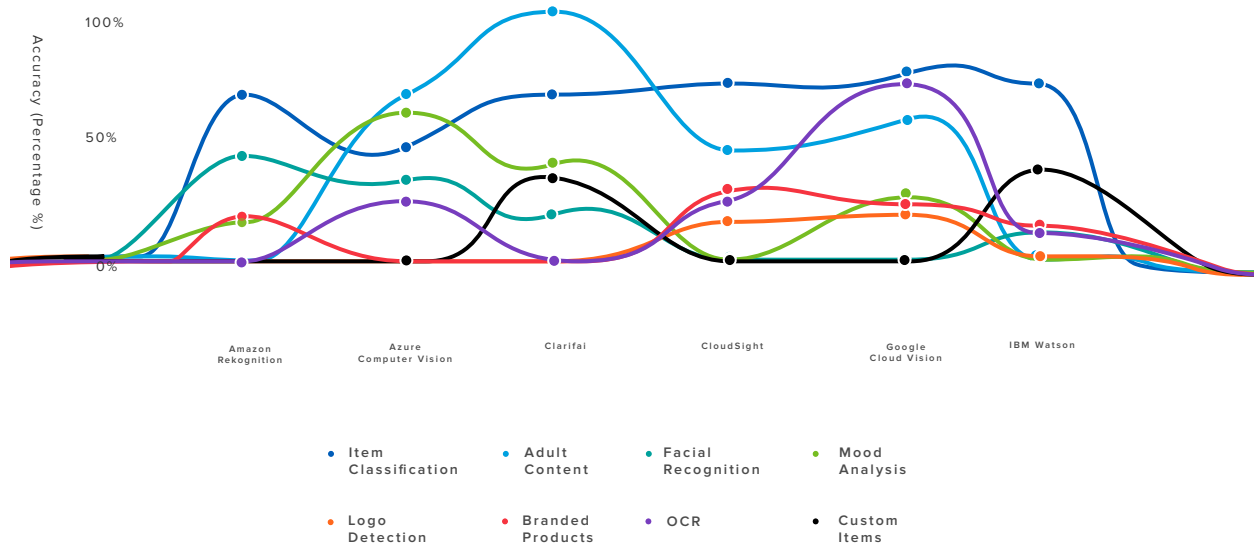
Major Takeaways

Based on the results of the tests, we concluded that it is unlikely that any one service by itself will meet all the image recognition needs of a business or government agency. Each service excelled in some functions but came up short in others. Which services are best for your organization will depend on your specific needs. We recommend that organizations:

- Plan on using more than one service. We suggest architecting your system with an orchestration layer between the system consuming the services and the services themselves. This layer should route the request to the best service for that type of request and isolate the consuming app from variations in service protocols.
- Plan on architecting the systems so that it is easy to change vendors. The orchestration layer should be designed to enable this.
- Be prepared to handle fuzzy and noisy responses (i.e., incorrect answers). Your systems will need to apply some intelligence of their own to correctly interpret and respond to answers that include probabilities of correctness.
- Develop a use model to estimate costs and determine which vendor(s) will provide the lowest cost for the functions needed based on your unique business case.
- Consider training your own image recognition engine to support your unique business case. An engine trained on a specific type of image will significantly outperform general-purpose engines.

Overall Accuracy Assessment

We found that no one service has a clear lead across all tested functions. With no one-size-fits-all solution available on the market today, we recommend that organizations looking to adopt image recognition be prepared to use multiple vendors to accomplish their strategic goals and objectives.



Test Results and Observations

For each of the nine functions tested, we compiled detailed results as well as observations about vendors and their capabilities. Not all vendors offer all nine functions. The summaries below include brief discussions of the methodology used in evaluating functionality.

Branded Product Identification

This function involves the identification of specific popular brands within an image. A company may want to use this to ensure that images of competitors' brands are not shown in images displayed on the company's systems. It could also be used to provide contextual knowledge about the person who submitted an image; for example, it might note that, in the image, the branded product appears on a dinner table set for Thanksgiving dinner.

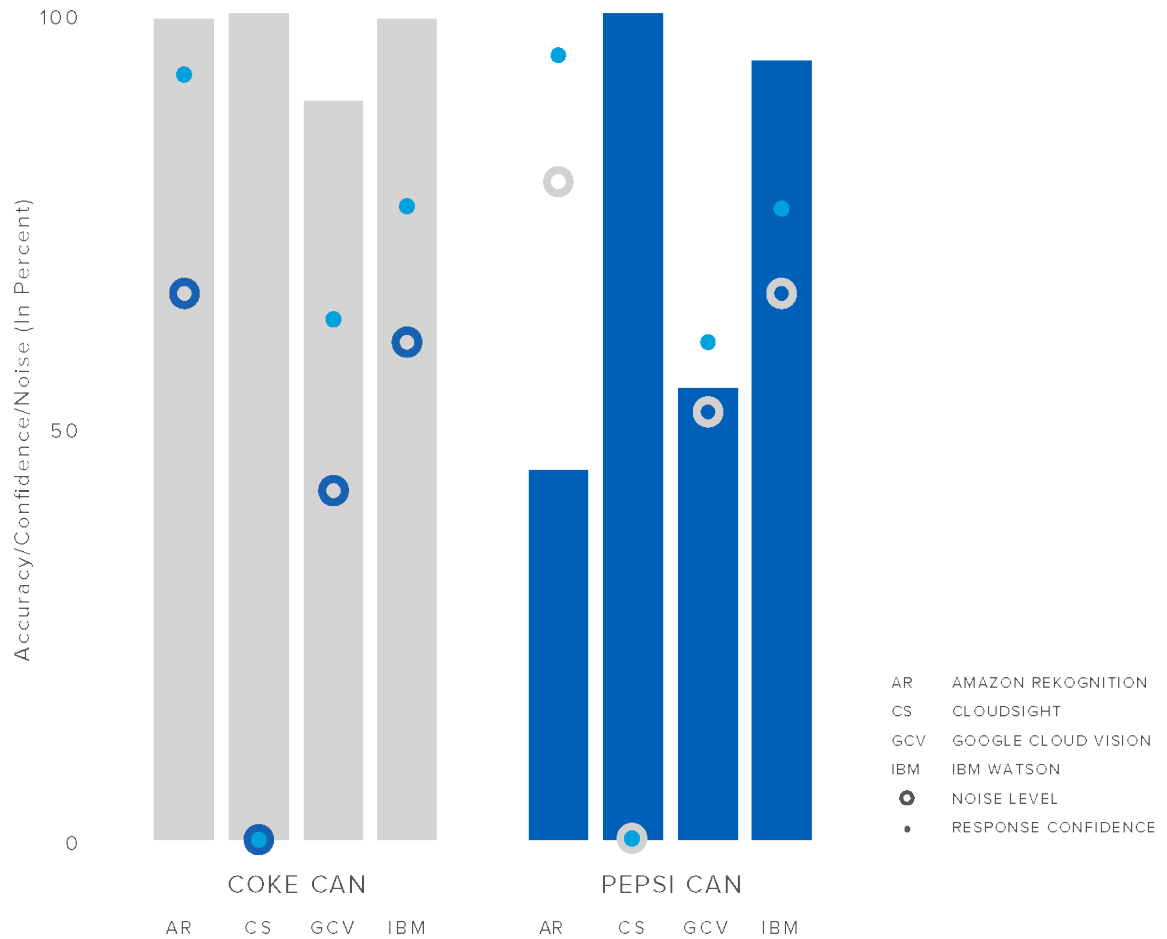
None of the services that offered this functionality performed well in our tests. All the services we evaluated, except CloudSight, expressed misplaced confidence in their own answers. All services tested provided many incorrect answers. CloudSight, the most accurate of the services, was incorrect in 60% of all test cases.

Another issue we identified involves training bias within data sets. Some services are so heavily trained to recognize images of Coca-Cola cans, for example, that they incorrectly identify other objects (such as fire hydrants and red barrels) as Coca-Cola cans. A related concern is that training is not systematic across branded products. The services we tested did well in identifying Coke cans but were less effective in identifying Pepsi cans.

Methodology:

We tested 464 image variations of single groups of branded products. The services' responses included free-text definitions of any identified brands and products. We assessed accuracy by matching the returned text to the combinations of any acceptable variation of the expected brand or company name as well as the type of product.

Coke vs Pepsi Image Recognition



In the classic battle between Coke and Pepsi, Coke came out on top. The noise levels of the services give an indication of the classification style response.

Branded Product Details

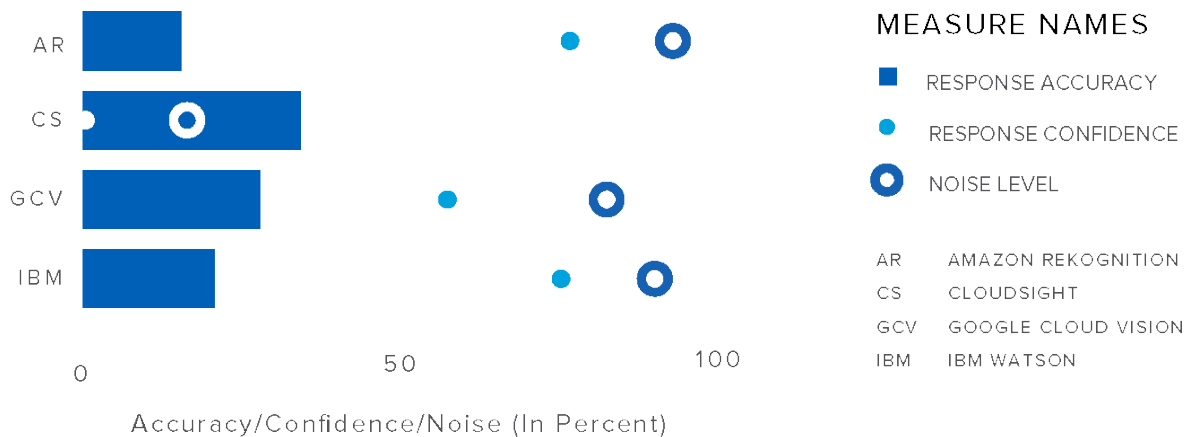
	AR	CS	GCV	IBM
CAN OF CAMPBELL'S CONDENSED..	0%	87%	66%	0%
COCA COLA CAN	99%	100%	96%	100%
DOMINION TRUCK	50%	34%	46%	45%
DUKE'S MAYO	13%	93%	40%	15%
JOHN DEERE	6%	57%	10%	0%
LEVI'S JEANS LABEL	5%	0%	0%	0%
NIKE SHOE	50%	72%	47%	30%
PAIR OF IPHONES (PHOTO)	4%	52%	0%	0%
PEPSI CAN	44%	99%	57%	96%
RYOBI DRILL	0%	65%	30%	41%
SAUER'S SPICE	0%	26%	10%	23%
SINGLE LACOSTE SHIRT	0%	76%	42%	42%
STACK OF PACKAGED FOODS	0%	1%	9%	0%
STANLEY FATMAX TOOL	0%	75%	12%	21%
TOYOTA EMBLEM	0%	100%	0%	0%
USG DUROCK	0%	72%	10%	0%

AR AMAZON REKOGNITION
 CS CLOUDSIGHT
 GCV GOOGLE CLOUD VISION
 IBM IBM WATSON

CloudSight generally outperformed the other services, particularly for well-known products or brands that featured some sort of labeling to help identify them. This is anticipated based on the mechanical turk nature of their responses and the associated slow response times.

Branded Product Identification

Branded Product Identification tests measured the ability for the service to determine the specific brands and types of products found in the image. To score well, the type of product must be identified either directly or through classification terms, and the brand must be recognized.



Custom Item Recognition

Custom item recognition tests the ability of the services to be trained to detect items unique to a business case and then detect those items when they appeared in new images. This test was done by training (or seeding) the various systems with three to five images for each of a variety of items such as a pair of shoes or a roller coaster. We then determined whether the systems could identify the same pair of shoes or the same roller coaster when it appeared in new images.

This capability allows organizations to provide contextual recognition unique to their products or locations. For example, a retailer that has set up an in-store display could use custom item recognition to identify the display when it appears in images. Similarly, a mobile app could allow the customer to take a picture of an item, and the app would then find similar items offered so the customer could make a purchase immediately.

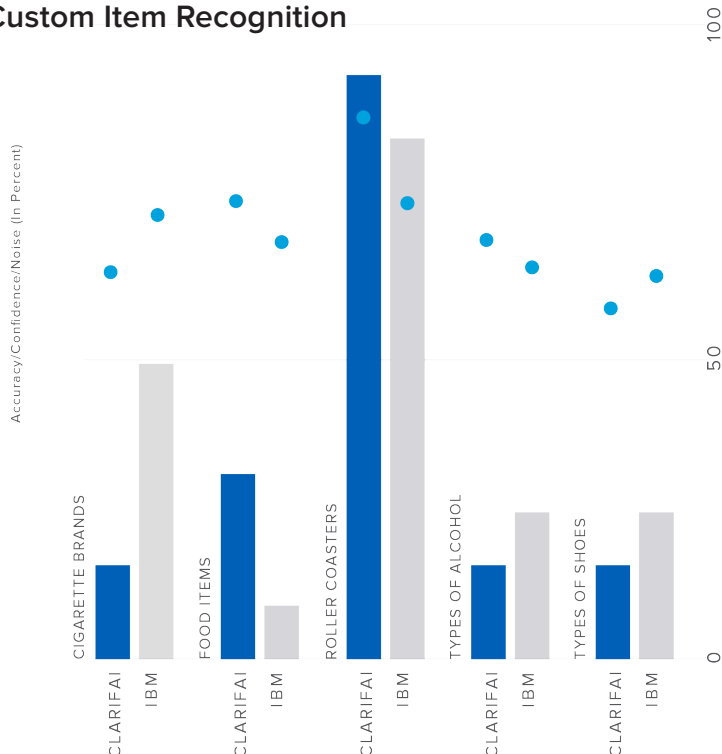
Only two vendors – Clairifai and IBM Watson – provided this function. Generally, both performed better when items were visually distinctive. IBM Watson did a better job of identifying items with similar packaging; for example, clear bottles of alcohol with only minor label differences. Clarifai was better at identifying items that were dramatically different from one another.

Accuracy correlated highly with the quality of the training set. A large training set with photos taken from diverse angles was more accurate than a small set with limited angles. As with any type of machine learning, the more data provided to the training data set, the more accurate the answers.

Methodology:

We created a training set of 145 curated training images of 46 items. The test set included 1,769 image variations of individual and grouped items. The images were of the quality and composition of photos quickly taken by phone. The services' responses included ratings of the similarity of test images to the images in the training set. We evaluated accuracy by matching the most confident responses to expected items.

Custom Item Recognition



Item Categorization

Item categorization is the process of looking at an image and identifying the items the system sees within the image; for example, a dog sitting in flowers. This functionality is often the go-to feature for demonstration purposes. This function would enable a state highway department to assess images of roads and highways and identify potholes. The same functionality would enable a media organization with a large photo library to gain a deeper understanding of the assets in the library.



Ideally, item categorization should provide not only the general category of the item (dog or flower) but the relevant subcategory as well (Labrador Retriever or daisy).

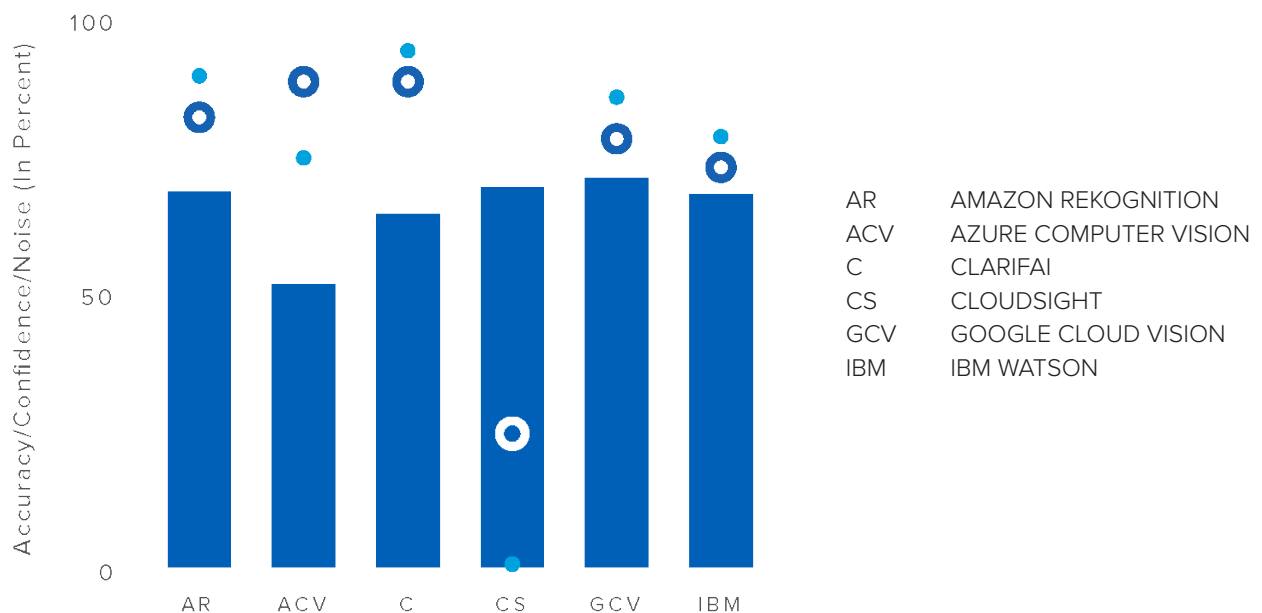
All six vendors provided an item categorization function. We graded them based on the specificity of their answers; for example, a service that identified the breed of dog or the type of flower received a higher score than a service that identified only a dog or a flower.

On average, performance of this function was the strongest of all functions tested. Azure and Clarifai delivered the weakest results in this category. The services performed better on certain types of items; for example, the services were more accurate in identifying stringed instruments than in identifying animals.

Methodology:

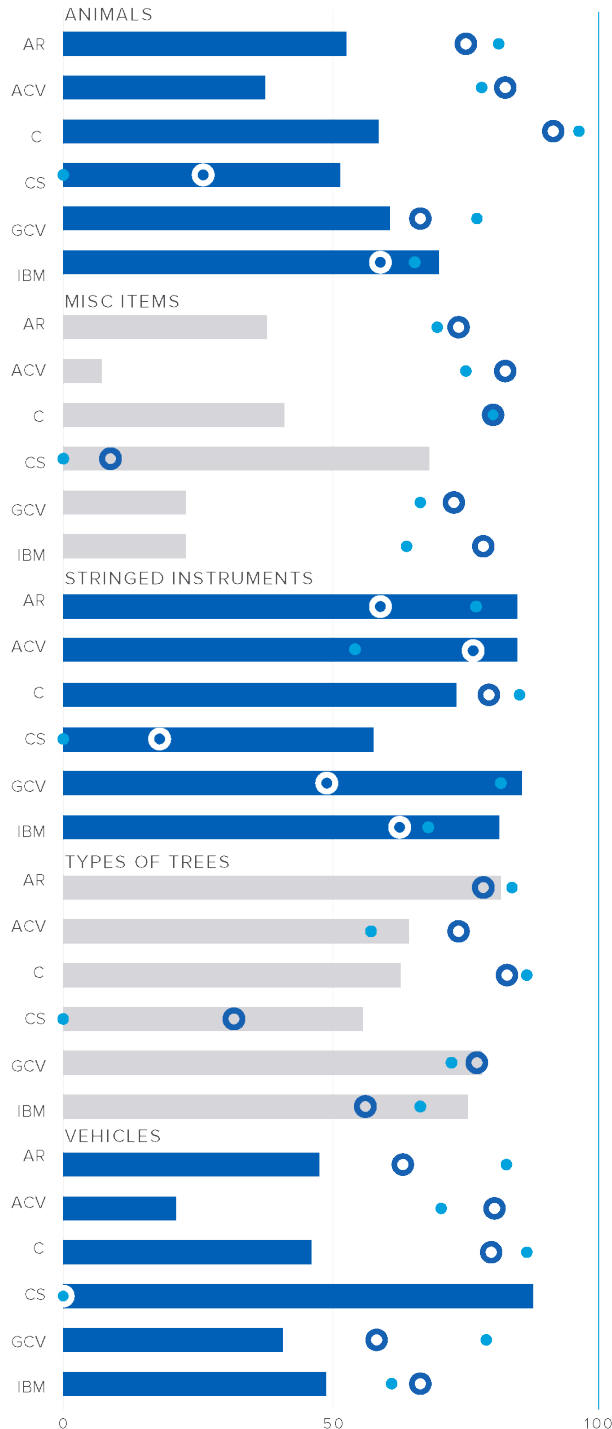
We tested 609 image variations of singled or grouped generic items. Service responses included classification descriptors for each of the items identified in an image. We assessed accuracy by noting the most specific classification (e.g., mammal, dog, golden retriever) returned for each expected item, correlating specificity with accuracy.

Item Categorization



Item Categorization Groups

Classification tests are broken down by general conceptual domains. When evaluating classification services for viability, the particular domain of interest should be used in the comparison as each service is stronger in some domains than others.



Logo Recognition

Logo recognition is the ability of the service to identify a brand logo in an image. A business might use this function in a rewards system, awarding points to customers who post images of branded items on social media. The image recognition service automatically validates the existence of the logo.

Only two of the services that we evaluated provided this function: CloudSight and Google Cloud Vision. Given CloudSight's costs and response times, Google Cloud Vision appears to be the more viable offering.

For both services, accuracy was mediocre in instances in which logos did not include text. However, it is worth pointing out that few logos that appear on consumer products do not bear a company name.

When logos were included with text, the two services were highly accurate. We believe this occurs primarily for two reasons:

- The services use some level of optical character recognition (OCR) to read the text and then use that information to augment general shape and color matching.
- Adding text provides corners as well as variations in shape and color, all of which offer visual cues as to brand.

Methodology:

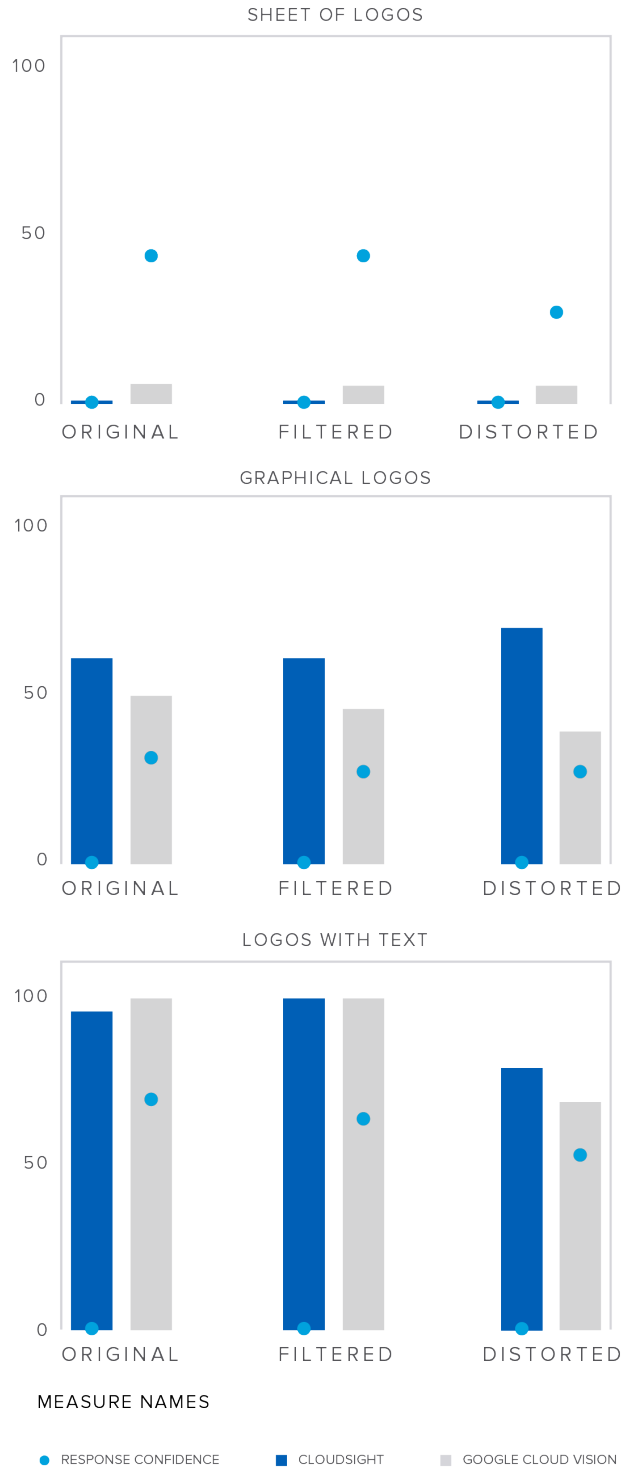
We tested 696 image variations of clean brand logos on a plain background as well as 87 image variations containing multiple logos on a plain background. The services' responses included free-text descriptions of any identified brands. We evaluated accuracy by matching the returned text to any acceptable variation of the expected brand or company name (e.g., Coca-Cola, Coke, Coca Cola).

AR AMAZON REKOGNITION
 ACV AZURE COMPUTER VISION
 C CLARIFAI
 CS CLOUDSIGHT
 GCV GOOGLE CLOUD VISION
 IBM IBM WATSON

Logo Recognition

Logo recognition tests measured the ability for the service to determine the company indicated by the logo image. Test images included only the logo brand itself.

Both services evaluated were generally able to determine the proper company if the logo contained text and were not distorted, but success dropped off dramatically for graphic-only logos, while a sheet of logos were missed nearly entirely.



Facial Detection

This involves the detection, as opposed to the identification, of faces within an image. Face detection is probably the best-understood and best-performing of all the functions we tested. One important use of face detection, for businesses as well as government agencies, involves counting people via camera. CapTech has used this functionality within a mobile app to help the visually impaired take selfies.



The reliability of this function is heavily influenced by image size; the larger the image, the greater the likelihood of successful detection.

For businesses considering adding this function, it is worth pointing out that modern mobile devices perform face detection on-device, so service-based face detection by itself is not particularly compelling.



Methodology:

We tested 90 image variations, ranging from a camouflaged face to a group of 13 individuals of varying age, race, and focal depth. The services' responses included bounding boxes around the detected faces. We evaluated accuracy by matching the center of the returned boxes to expected face positions, within a range of tolerance. IBM Watson was the only service that detected the camouflaged face.

Facial Detection - Size Impact

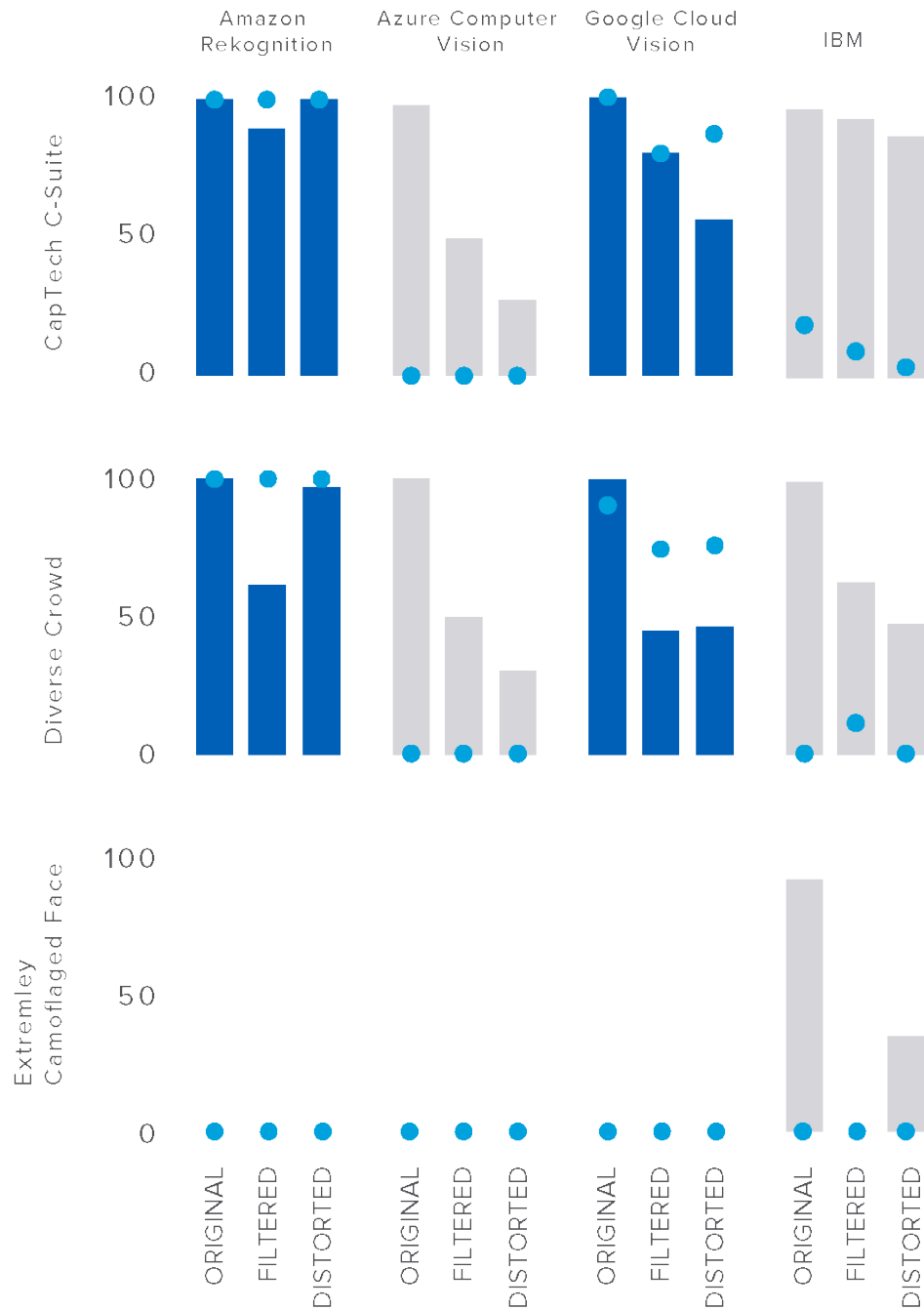
Facial detection is significantly impacted by image resolution. When considering a gradient of smaller image resolutions, IBM Watson detects faces better overall.



Facial Detection

Facial detection tests measure how accurately services are able to pinpoint the location of one or more faces in an image.

Amazon Rekognition has solid facial detection, and handily beats the other evaluated services, with the exception of IBM Watson which was able to successfully identify the camouflaged face that no other service was able to identify. However, Watson's accuracy comes at the expense of a significantly slower response when more faces are in the image.



Facial Recognition

This functionality goes beyond face detection and identifies people in photos. Organizations might use facial recognition to identify VIP customers in a retail store or to identify people who are believed to present security risks.

Facial recognition is a two-part process. First, a training set of images is loaded, providing images of the people who will be searched for. Second, each image is tested, with the service reporting those people from the training set whose faces appear in test images.

Only Amazon Rekognition and Azure provided true facial recognition. With Watson and Clarifai, we had to use a combination of face detection and custom item recognition. This is not a particularly effective approach, and it is time-consuming because it takes more calls to the service to get results.

The services were often stymied by distortions to the images. An example of this involved simple inversion of the image. The services recognized actor Brad Pitt in the upright image, but did not recognize him when the image was upside-down. The exception to this was CloudSight, which could identify celebrities despite heavy distortions and filters. The probable reason for CloudSight's performance is not better algorithms, but rather its apparent reliance on human labor.

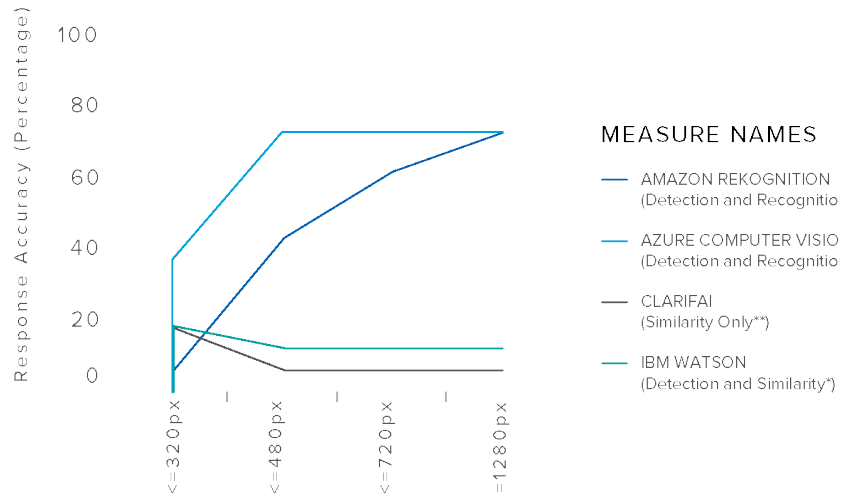


Methodology:

We created a training set of 63 images of 21 people from different perspectives. Our test set included 58 image variations containing subsets of those 21 people. The services' responses included bounding boxes around the detected faces. We evaluated accuracy by matching the center of the returned boxes to expected face positions, within a range of tolerance, with the correct associated matching face.

Facial Recognition - Size Impact

Facial recognition is significantly impacted by image resolution. When considering a gradient of smaller image resolutions, Azure recognizes faces better overall.

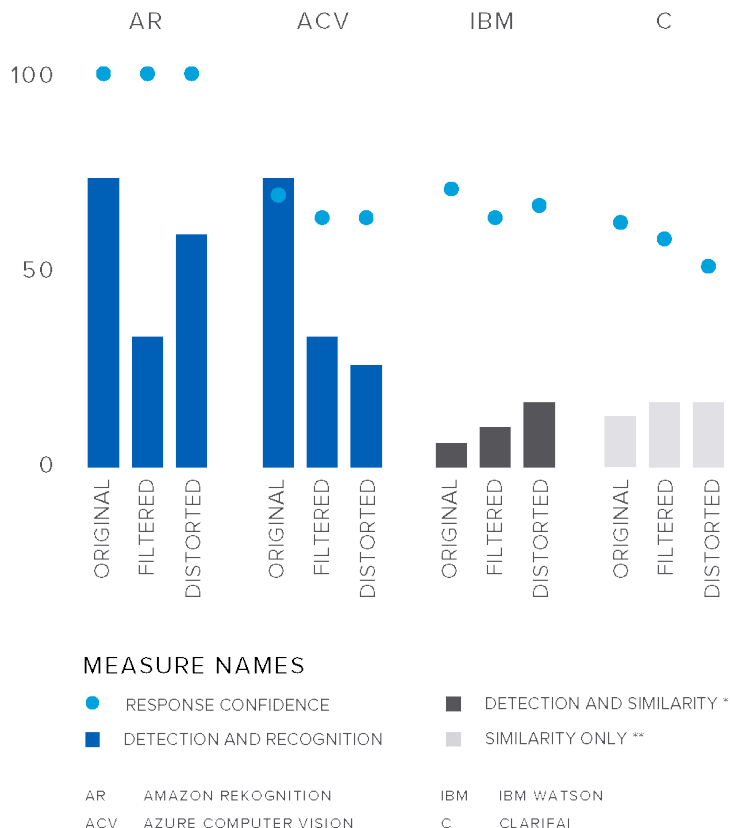


Facial Recognition

Facial recognition tests measure how accurately services are able to pinpoint the location of one or more faces in an image and recognize the person that face belongs to based on a set of training images.

* IBM Watson does detection (position), but neither Watson or Clarifai support facial recognition directly. They were evaluated using image similarity to the source images.

Both of the two services supporting full detection and recognition capabilities suffer under filtered image conditions (blurry, bad exposure). Amazon Rekognition performs significantly better with image distortions, however, which is likely to be a common use case for anything but security-focused bio recognition.



Mood Analysis

Mood analysis goes a step beyond facial recognition and seeks to identify the mood of the person in an image. We tested five moods: angry, happy, sad, scared, and surprised.

This functionality is marketed for a wide range of applications. For example, a fast food restaurant might judge patrons' moods as they order at a kiosk, providing a service akin to continuous usability testing. Mood analysis could help rate television programs more effectively. It could add a security layer for shopping malls, airports, sports arenas, and other public venues where detecting malicious intent is of value. Mood analysis services could be linked to wearable devices, helping autistic people, for example, discern others' emotions.



Only Azure and Google supported this function directly. We attempted to get Amazon and Clarifai to detect moods through their general classification features. (Clarifai offers a domain for "general" recognition. Amazon Rekognition provides a similar feature.) Clarifai returned reasonably good results.

Generally, we found that the more filtered or distorted an image was, the more trouble the services had with it. Rotating or inverting images tended to reduce accuracy greatly. We also found that happy was the mood most likely to be identified.

For businesses interested only in mood, there are application programming interfaces (APIs) specific to mood analysis that may be more effective than the general purpose services we tested.

Methodology:

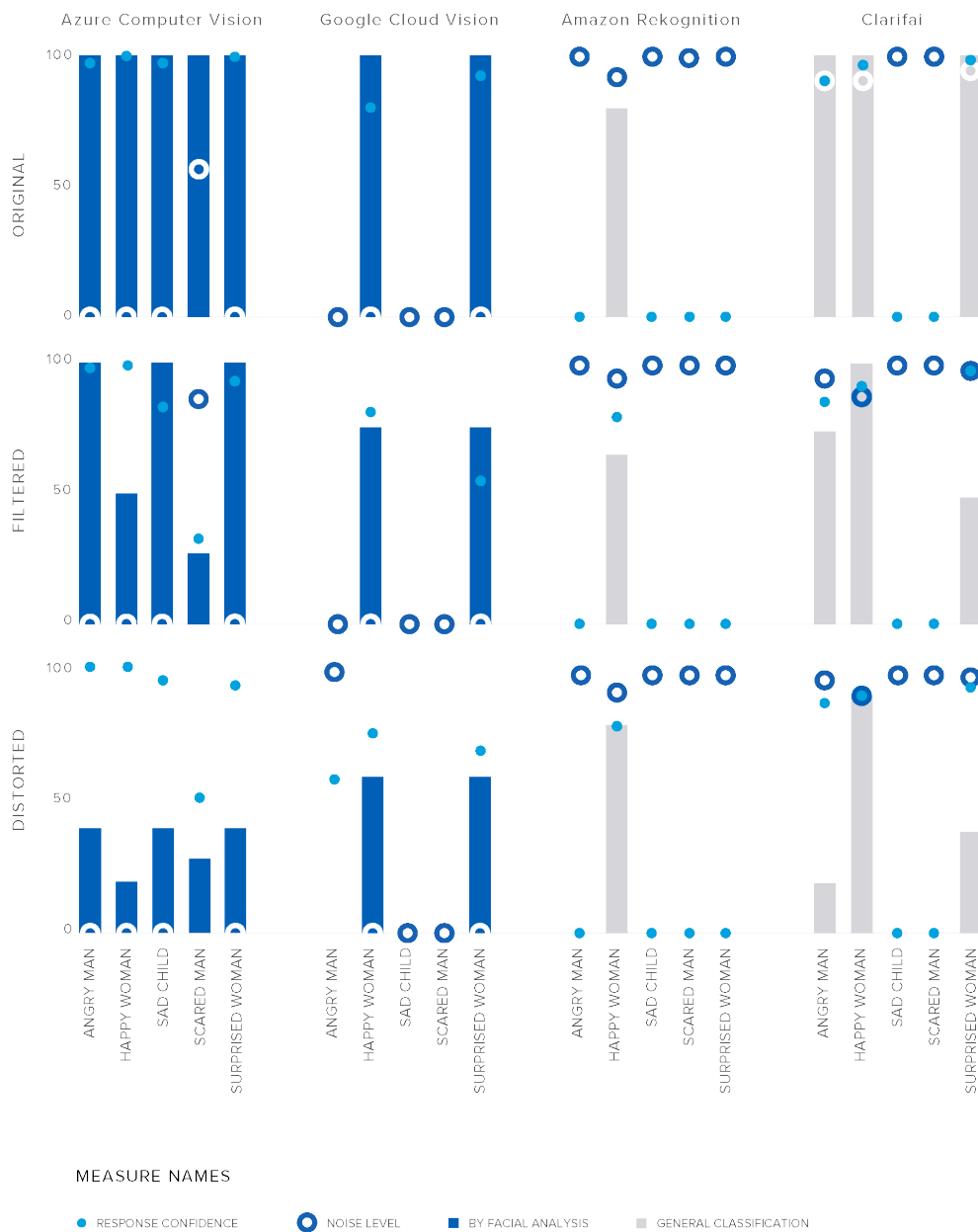
We evaluated 145 image variations representing anger, happiness, sadness, fear, and surprise. Mood indicators were determined through facial landmark analysis or general classification, sometimes with associated likelihood factors. We assessed accuracy by determining whether the service properly recognized the mood depicted with at least a 50% confidence.

Mood Analysis

Mood analysis tests measured the ability for the recognition services to appropriately classify the exaggerated emotions shown in the test images.

Mood prediction is still in its infancy based on our tests, and Azure's Computer Vision service wins hands down.

For determining happiness alone, general classification systems also fair well despite the lack of domain-specific knowledge, though the inherent noise of the classification responses leads to a significantly lower per term accuracy over the facial analysis services.



Text Recognition (OCR)

Text recognition, sometimes called optical character recognition (OCR), is the process of converting typed or handwritten text into character values as traditionally managed by computers; for example, a picture of an A would translate to the ASCII* value 65.

If you can't explain it simply, you don't understand it well enough.
-Albert Einstein

For businesses and government agencies, text recognition is marketed as a way to help drive greater efficiency and cost-effectiveness; for example, it may make the filing of text documents easier for the user.

Basically, text recognition enables organizations to receive documents from less technically advanced sources and make them machine-readable. Unfortunately, the performance of the four providers that we tested for this function was mediocre. Google Cloud Vision performed the best on a short piece of typed text; however, the 75% to 80% degree of accuracy we observed is not reliable enough to be used in the absence of a person who can validate the results.

To help customers prepare their returns for the tax year 2016, H&R Block has been providing this functionality. OCR functionality reads relevant information that appears on W-2 forms (which summarize earnings from employers), reducing the need for manual entry. However, given the potential for recognition errors as well as the consequences of submitting an erroneous tax return, it is essential that taxpayers double-check results of the OCR functionality before submitting tax returns to the government. Our tests with this W-2 functionality yielded frequent errors in the generated data.

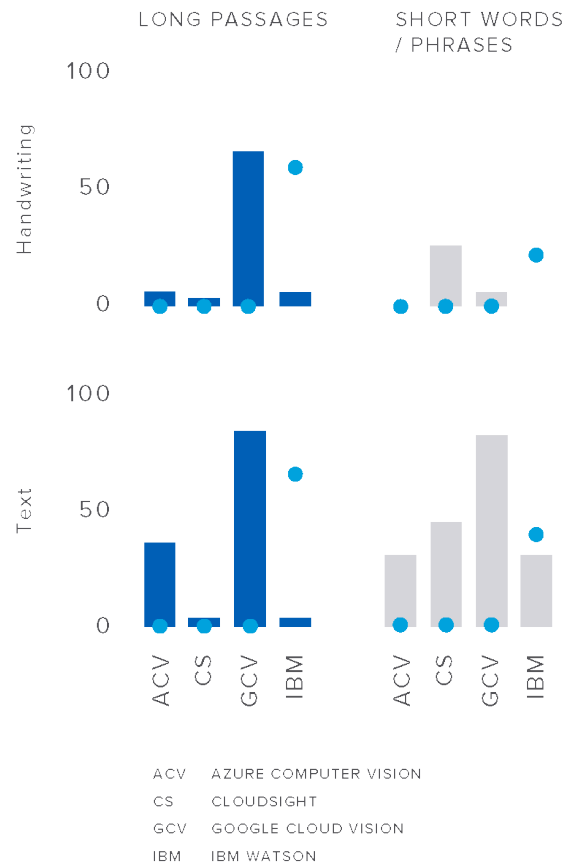
Instead of using any of the four services we tested, we recommend that businesses and government agencies that need text recognition functionality opt for a specialized OCR service such as ABBYY or an in-house implementation of the Tesseract open-source project.

Methodology:

We tested 100 image variations containing handwriting in print and cursive. We also tested 174 image variations containing typesetting, ranging from plain text on a plain background to captioned images. Service responses included individual or contiguous blocks of recognized text. We assessed accuracy by gauging the similarity of the responses to the expected text, based on a modified Levenshtein distance method (which determines how many characters must change to turn one word into another word).

Text Recognition (OCR)

OCR tests measured the ability for the recognition service to properly determine the exact text shown in the image. Test images included both handwritten and typed formats, with both long and short phrases.



* ASCII, or American Standard Code for Information Interchange, is a format for text files that is widely used in computers and on the internet.

Adult Content Detection

This functionality determines whether an image contains sexually suggestive or objectionable material. Businesses might use adult content detection to verify that user-provided content will not be objectionable to other users or to customer support staff.

We validated this function by using actual objectionable content as well as content that could be mistaken for objectionable but is not (e.g., a picture of Michelangelo’s statue of David or a picture of two objects that could be mistaken for female breasts).

Clarifai was 100% accurate in identifying objectionable material. Clarifai also was 100% accurate in categorizing classical works of art and other non-offensive content as non-objectionable. In contrast, we found that Azure and Google were stymied by simple image filtering; for

example, overexposure or underexposure of images.

Before adopting a service, we recommend that organizations first test each service against organizational requirements and sensitivities.

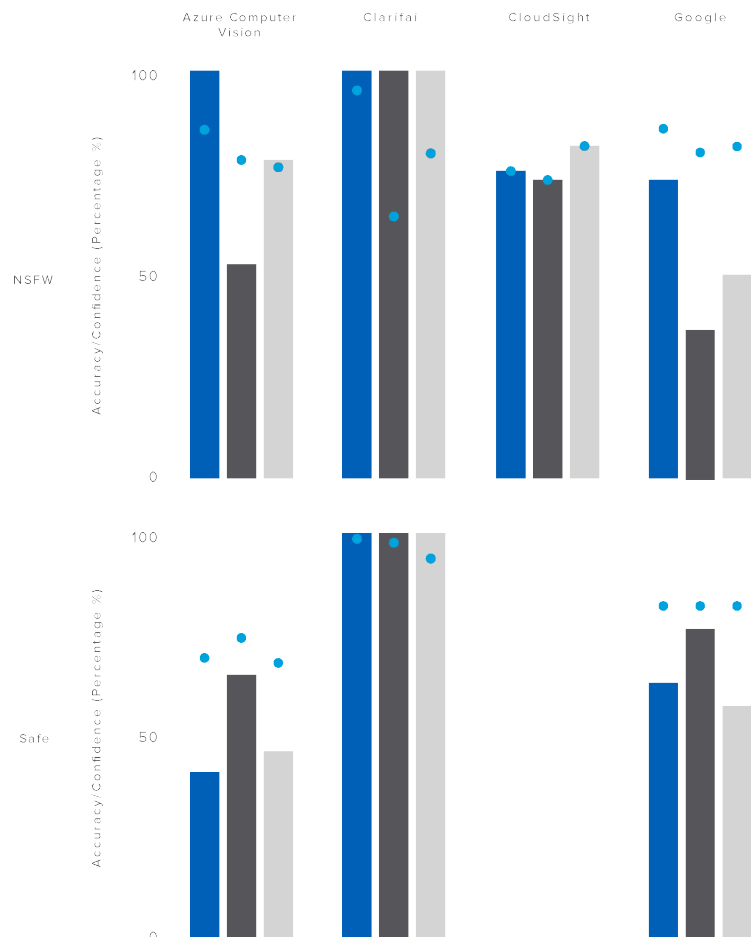
Methodology:

We tested 348 “not safe for work” (NSFW) softcore and hardcore pornographic image variations as well as 145 “safe” image variations containing nude classical art, misleading shadows and shapes, and suggestive advertisements. Responses included positive/negative indicators of explicitness. We assessed accuracy by determining whether the services correctly specified whether an image was NSFW or safe.

Adult Content Detection

Explicit recognition of images as either NSFW or Safe. The safe images evaluated contained deceptive shapes meant to fool the target services, such as classical works of art, the concentric circular shadows of two lamps, and suggestive advertisements.

Clarifai is a clear winner here, with 100% accuracy in determining whether an image is safe for work, even with filtering and distortion applied under which Azure’s strong accuracy suffered significantly. In addition, Clarifai’s average responses were at least 60% faster than the other services’.



Vendor Summary

While many vendors today offer partial image recognition solutions, focusing narrowly on such functions as text recognition or facial recognition, our tests covered only vendors that offer general services. This section of the paper summarizes our findings regarding each of these vendors, noting the specific functions tested and the strengths of each vendor.

Amazon Rekognition

If you are already an Amazon customer, the Amazon Rekognition platform, which Amazon released in autumn 2016, may be the solution of least resistance. Like all the services we tested, Rekognition is undergoing continual improvement.

The Rekognition service does not perform item identification in a manner similar to the Amazon mobile app. In the Amazon mobile app, customers can point their phone's camera at a product and it will identify that product in Amazon's inventory with sufficient specificity to order the item. Going into the tests, we had hoped to find that type of branded product and logo recognition but were disappointed that the functionality was not provided.

Amazon Rekognition struggled with smaller images. The other services sometimes succeeded with smaller images, but Amazon had a higher threshold for image size.

Because Amazon Rekognition does not do branded product or logo recognition, we did not test these functions. It is possible that Amazon has segregated this functionality to prevent others from replicating the company's consumer-facing product recognition service. If there is one company that should do product recognition, it is Amazon.

Features Tested

Item categorization, face detection, facial recognition, and mood analysis

Strengths

- Amazon Rekognition performed well in face detection.
- It was also strong in facial recognition, especially with distorted images.
- General item categorization was strong, although confidence was often misplaced.

Azure Computer Vision

Azure Computer Vision is Microsoft's image recognition offering, and is part of the company's cloud service. For existing Azure customers, this would be a logical go-to service for image recognition. It outperformed other providers in mood analysis and facial recognition, but had problems with image size, faring poorly with smaller images.

Features Tested

Adult content detection, face detection, facial recognition, mood analysis, text recognition, and item categorization

Strengths

- Mood analysis and facial recognition were strong; however, facial recognition performance dropped off sharply when an image size fell below 480 pixels.

Clarifai

Clarifai provides a service unique among these vendors. It takes a domain-specific approach to image recognition, providing recognition domains for apparel, celebrities, color, faces, food, NSFW (or "not safe for work"), travel, and weddings. An additional domain is labeled "general." Clarifai will not divulge what is in that model but urges customers to try it and see if it meets their needs. Customers can also upload their own model images to train a new customer-specific domain. We did not test that functionality. Given Clarifai's performance with pre-packaged trained models, we suggest that you consider Clarifai for custom-trained models specific to your business problem.

One of the challenges we experienced with Clarifai is that the number of responses it provides varies widely. With other services, the customer sets the parameters or the vendor provides a set number of responses for each query. Clarifai's approach is to limit the responses based on confidence; therefore, it can be problematic to determine where to set that value for a given type of query.

Features Tested

Adult content, custom item recognition, face detection, facial recognition*, item categorization, and mood analysis

(For Clarifai, we tested facial recognition by building a custom model that included a set of faces we wanted Clarifai to recognize. Other services take care of this for the customer, so custom models are not necessary.)

Strengths

- Adult content (NSFW) detection: Clarifai was 100% accurate in identifying objectionable images and in distinguishing these from classical art and other non-offensive images.
- Clarifai achieved above-average scores for face detection and mood analysis.
- The service can build and manage custom detection models.
- Clarifai was one of the lower-cost vendors evaluated.

CloudSight

CloudSight appears to be a mechanical turk. In our tests, responses consistently took between 12 and 20 seconds, and many contained typographical errors. When we passed the same image to CloudSight multiple times, we received similar answers, but with variations in spelling. For example, we submitted this picture of television star Ellen DeGeneres:



The answers received included: Ellen DeGeneres, ellen the generous, ellen photo, helen degeneres, Ellen De Generes, and woman in a collared shirt. We initially saw this as a major weakness, but in tests where an image was heavily distorted, CloudSight outperformed other vendors. For example, when presented with an upside-down image of Channing Tatum, CloudSight recognized it, but other vendors did not.

We had some additional concerns. Among them: CloudSight does not assign a confidence level to its answers. The customer receives only one answer per query, whereas most competing services provide many more answers per query. In addition, CloudSight is by far the most expensive of the offerings, costing 30 times as much as the fully automated alternatives. Most notably, use of a mechanical

turk presents security concerns. If your app handles private financial, personally identifying, or health information, it is probably inappropriate to use this type of service.

Features Tested

Adult content detection, branded product identification, item categorization, logo detection, and text recognition

Strengths

- Although CloudSight suffers from slow response times, it is good at branded product identification and item categorization, probably because there are people behind the curtain doing the actual recognition.
- CloudSight outperformed other services in accurately recognizing short phrases of text; however, longer phrases tended to have more typos, which reduced accuracy.

Google Cloud Vision

Cloud Vision, Google's offering in the image recognition API space, is not the same as the company's reverse image search. (To use the latter functionality, customers submit an image to Google, and it returns other images that are similar.) Cloud Vision includes a large suite of recognition functions via a straightforward API.

Features Tested

Adult content detection, branded product identification, face detection, item categorization, logo detection, mood analysis, and text recognition

Strengths

- Cloud Vision provided consistent response time performance across recognition requests.
- It performed best for general item classification and OCR functionality.
- Cloud Vision offered a good set of features, but different features may return slightly different format messages, making integration more difficult; for example, if you ask it for item categorization, the response format will be slightly different from that used in logo recognition.

IBM Watson

Image recognition is part of IBM's heavily marketed Watson artificial intelligence product line.

The service provided widely varying response times, from 1 to 9 seconds. In addition, Watson had the smallest file-size limit of any service tested, disallowing larger images. This complicates integration and could limit the accuracy of the service, as all services experience decreased accuracy as image size decreases.

Watson offered text recognition but was surprisingly bad at it. We submitted the following image:

And we got this nonsensical response:

***Maw m li1vrwm HI Mil mi 1h HA h My
1111mm m i. N MW\ nH M W4 1W1 h
mum mi h i In I li Mimi I w I***

English Sample

Transcription: M. Avinor

We walked across the landing, and in the hall below the grandfather clock softly chimed. There was a smell of poliu, the landing was dim and cool. Beneaù her bare feet the rugs were soft. Urough the gloom ue could make out the carved mahogany of the banisters, spirals and curlicues. Miss Digg was waiting for her, the music open on the piano. There were roses in a bowl and a smell of roses in the room. We played the Bach, the Minuet in G. "You've practiced," Miss Diggs said. "I can tell you've practiced, Elizabeù." We went on playing. The notes came easily and we couldn't understand it because ue hadn't practiced at all.

Features Tested

Branded product identification, custom item recognition, face detection, facial recognition, item categorization, and text recognition (OCR)

Strengths

- Watson performed better than other services in custom item recognition, slightly edging out Clarifai. But, we feel that the increased cost of providing the training image set gives other vendors the edge in overall value.
- It performed above average in face detection and item categorization.

More About Costs

Performance and accuracy are not the sole factors to consider in selecting an image recognition service. The costs of set-up (i.e., system training) and operation also are key.

For these seven functions (i.e., Branded Product Identification, Item Categorization, Adult Content Detection, Face Detection, Mood Analysis, Text Recognition, and Logo Recognition), vendors charge based on the number of requests made, not the kinds of requests made. For a given vendor, 1,000 requests for adult content detection, for example, would cost the same as 1,000 requests for item categorization.

However, Google and Amazon offer a tiered pricing model, with a volume discount that can reduce these costs. Using a test case of 30,000 requests per month (the largest value that we could calculate with publicly quoted prices across all vendors), we found Amazon and Clarifai were the least expensive at \$30, followed by Google Cloud Vision at \$43.50, Microsoft Azure at \$45, and IBM Watson at \$60. The most expensive — costing 30 times more than Amazon and Clarifai — was CloudSight at \$900.

Note that as request volumes increase, the vendors with tiered pricing become more advantageous, as shown by the graph below.

Facial Recognition and Custom Item Recognition

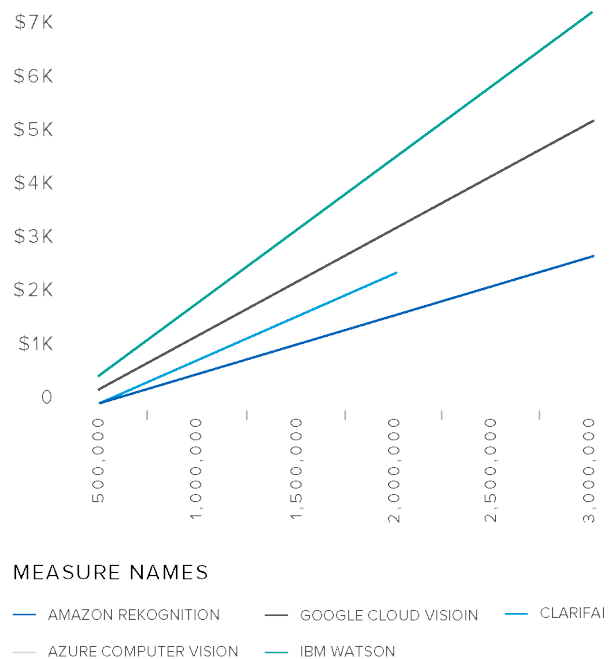
These two functions are priced differently from the previous seven. The services retain the same per-request charge for each request made, but Amazon, Azure, Clarifai, and IBM Watson charge an additional fee to train the recognition machine. For each image that goes into a facial recognition or custom item recognition training set, IBM Watson levies a one-time fee of 10 cents per image. That can add up quickly, as a reliable system needs thousands of training images. Moreover, many organizations will want to add to the training image set frequently, further driving up costs.

Amazon and Azure charge a monthly fee of \$0.00001 and \$0.00005 per training image, respectively. Clarifai charges a classification and storage fee of \$0.0008 per image per month for each image used to train the recognition machine, with the first 10,000 images being free.

Conclusion

The six image recognition services that we evaluated are generally good, but not one is complete in itself. If you adopt image recognition services, plan on supporting multiple services and on being able to switch services quickly. For many businesses, the expenditure will be cost-effective: a few thousand dollars per month for 3 million image recognition requests. That could prove highly valuable — depending on which features you add to your app and on your app's ability to deal with the inaccuracies that characterize these services.

Costs for Generate Recognition Functions



About CapTech

CapTech is ready to help clients develop strategies to leverage image recognition services and integrate the technology with new or existing systems to get the greatest value from these services and achieve a competitive advantage. CapTech brings 20 years of systems integration experience and machine learning expertise to help our clients leverage new cloud-based artificial intelligence services.

For more information on selecting or integrating the right image recognition services for your organization, contact Jack Cox at jcox@captechconsulting.com or call 1.844.373.4025.

About the Contributing Researchers:

Jack Cox, Fellow



Jack Cox is a software developer, systems architect, and a Fellow at CapTech Ventures, Inc., where he is responsible for the firm's mobile and device software practice. Recently he has been working with Fortune 500 companies to develop mobile and IoT apps, and define strategies for adopting cutting-edge technologies. Jack is a frequent speaker at professional conferences where he opines on all things technological.

Chris Heinz, Senior Consultant



Chris Heinz is a software engineer in CapTech's Technology Solutions practice area. He has developed and architected embedded systems, Java-based web service applications, and Android mobile apps. Chris is passionate about delivering the best possible software solution utilizing proven patterns and techniques.

Kevin Vaughan, Manager



Kevin Vaughan has developed software for 15 years and has focused on mobile development and cloud platform services for the past 7 years. He has delivered solutions across many industries including gaming, education, e-commerce, and supply chain management. His contributions to open source communities and Q&A sites such as StackOverflow have reached hundreds of thousands of other developers.

Appendix

Research Methodologies

CapTech researchers submitted more than 4,800 images to 6 image recognition services and evaluated their accuracy comparatively to each other and to expected results using methods specific to each recognition-type, including Adult Content Detection, Facial Detection, Facial Recognition, Mood Analysis Prediction, Text Recognition (OCR), Logo Recognition, Branded Product Identification, Item Classification, and Item Recognition.

Adult Content

Test Images	348 NSFW softcore and hardcore pornographic image variations 145 “Safe” adversarial image variations containing nude classical art, misleading shadows and shapes, and suggestive advertisements
Service Responses	Positive/negative indicators of explicitness
Accuracy Evaluation	Correctly specifying whether an image was either NSFW or Safe

Facial Detection

Test Images	90 image variations ranging from a camouflaged face to a group photo of 13 individuals of varying age, race, and focal depth
Service Responses	Bounding boxes around detected faces
Accuracy Evaluation	Matching the center of the returned boxes to expected face positions within a range of tolerance

Facial Recognition

Test Images	Training Set: 63 seed images for 21 people at different perspectives Test Set: 58 image variations containing subsets of those people
Service Responses	Bounding boxes around detected faces with associated matching face
Accuracy Evaluation	Matching the center of the returned boxes to expected face positions within a range of tolerance, with the correct associated matching face

Mood Analysis

Test Images	145 image variations representing anger, happiness, sadness, fear, and surprise
Service Responses	Mood indicators determined through facial landmark analysis or general classification, sometimes with associated likelihood factors
Accuracy Evaluation	Properly recognizing the mood depicted with at least a “possible/50% chance” equivalency of likelihood

OCR

Test Images	100 image variations containing handwriting in print and cursive 174 image variations containing typesetting ranging from plain text on a plain background to captioned images
Service Responses	Individual or contiguous blocks of recognized text
Accuracy Evaluation	Similarity to the expected text based on a modified Levenshtein distance method

Logo Recognition

Test Images	696 image variations of clean brand logos on a plain background 87 image variations containing multiple logos on a plain background
Service Responses	Free-text descriptions of any identified brands
Accuracy Evaluation	Matching the center of the returned boxes to expected face positions within a range of tolerance, with the correct associated matching face

Branded Product Identification

Test Images	464 image variations of single or groups of branded products
Service Responses	Free-text descriptions of any identified brands and products
Accuracy Evaluation	Matching the returned text to the combinations of any acceptable variation of the expected brand or company name as well as the type of product

Item Classification

Test Images	609 image variations of single or groups of generic items
Service Responses	Classification descriptors for each of the items identified in the image
Accuracy Evaluation	The most specific classification (e.g., mammal, dog, Golden Retriever) returned for each expected item was evaluated, with a correlation of specificity to accuracy

Item Recognition

Test Images	Training Set: 145 curated seed images of 46 items Test Set: 1,769 image variations of individual and grouped items with the quality and composition of photos quickly taken by phone.
Service Responses	Similarity rating of test image to images in training set
Accuracy Evaluation	Matching of most confident responses to expected items