# Computer Vision: How it works and where it is headed

Ryan McCormick
Colorado State University
Fort Collins Colorado USA
ryanbmcc@rams.colostate.edu

Tony Phan
Colorado State University
Fort Collins Colorado USA
tonyphan@rams.colostate.edu

## ABSTRACT

Computer vision has been around since the 1960's and has dramatically improved over the last couple of decades. We wanted to delve deeper into the history of computer vision to understand the initial hopes and discoveries of the field and explain how it has changed over time. We delve into how modern-day computer vision programs work on a base level prior to being able to understand how computer vision is applied in modern day applications and potential future applications of this field. Through our research, we were able to identify how much computer vision has changed since its initial development and how the field has rapidly expanded to other areas of study.

## KEYWORDS

computer vision, artificial intelligence, model, object detection, facial recognition, security

## 1 INTRODUCTION

To find the information contained in this paper, we researched the history of computer vision. Alongside this we also investigated multiple examples of how computer vision is applied. Below are the overviews of each section.

In Section 2 of this paper, we discuss the background of computer vision. In this section we also discuss one of the first neural networks that inspired Convolutional Neural Networks. Section 3 describes our motivation behind why we decided to conduct this research. Section 4 delves into sample projects that currently use computer vision in real-world applications. Section 5 covers examples of how computer vision can be applied to other fields and potential implementations in the future. Finally, Section 6 summarizes our findings.

## 2 BACKGROUND

Computer vision, a field of Artificial Intelligence that was created in the late 1960's [1], began with one simple goal in mind; to mimic human vision via computers and be able to identify what they are seeing. The hope was to be able to automate image processing so that one day computers won't need any manual human intervention to be able to quickly and accurately describe what an image is showing. With this goal in mind, Larry Roberts (an American Engineer and Computer Scientist) began his career researching this idea and was often credited as the father of Computer Vision.

Roberts hypothesized that it would be possible to draw out 3D information from 2D images of blocks. This thought spurred the creation of what is known as Blocks World. This environment allowed users to interact with blocks on a table of different shapes and sizes. This allowed users to build columns of blocks by typing instructions to the computer. This showed that a computer could be told how to visualize and understand 3D items. From here computer scientists began trying to find a way to make computers take images and understand what was being presented to them.

Soon computers were able to create a 3D model from 2D photos. While this proved that computers were capable of understanding and replicating 3D images, it still did not allow computers to correctly identify objects in images. The introduction of the "Neocognitron" by Kunihiko Fukushima in 1979 would push the field even further towards the true purpose of computer vision.

The Neocognitron was one of the first neural networks that could read handwriting. This was accomplished by teaching the network what characters could appear and then using these characters to compare with what was being given as an input. This was the first time a computer was given the ability to replicate a human's neuron structure in such a way that it could truly learn as it ran and identify full words/sentences. Convolutional neural networks allow us to teach computers to process images better than we previously could have. The next sections of this paper describe how computer vision is used by those in and outside of the field including examples and analysis of certain computer vision projects.

## 3 MOTIVATIONS

Artificial Intelligence (AI) is slowly incorporating itself into other fields aside from computer science. With this, we decided to explore computer vision (a subfield of AI) as this field is now used in many of our daily lives (unlocking phones, self-driving cars, etc.).

We wanted to better understand computer vision by exploring how it works on a fundamental level. The hope of this research was to gather a deeper understanding of where this field currently is and where it can be in the future.

# 4 CURRENT CAPABILITIES

## 4.1 The General Public

*4.1.1 Teachable Machine Testing*

Google's "Teachable Machine" allows anyone to create a machine learning model via a webcam, still image input, and even audio files. This website can train three different types of models which are Image Projects, Audio Projects, and Pose Projects [2]. Each project can contain as many classes as you would like to train for. Within each class you can upload as many samples as you would like. For those using a webcam it is quick and easy to add enough samples that the website has nearly no problem training the model.

To show an example of this website we made a Pose Model with the hopes of identifying four different Tai Chi poses. The decision behind this was because there are a lot of physical overlaps between each pose. This allowed us to see if the model could correctly identify what it was seeing based on the samples it was given. The following images show the classes and samples used to create and train the model.

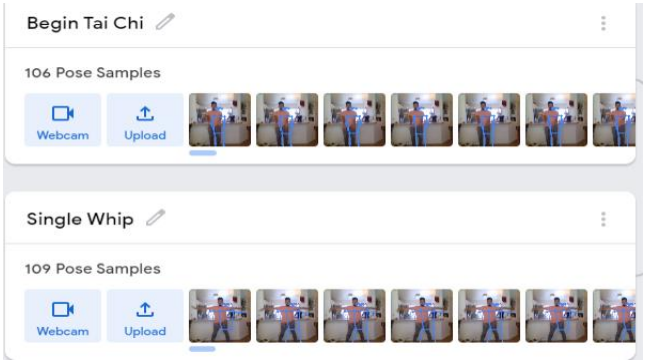


**Figure 1. Classes for Begin Tai Chi (106 Pose Samples) and Single Whip (109 Pose Samples).**

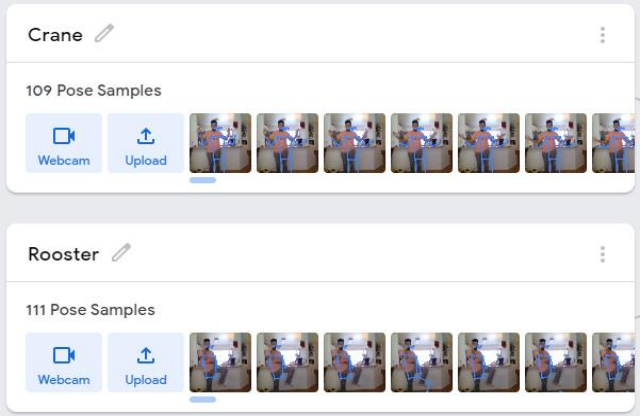


**Figure 2. Classes for Crane (109 Pose Samples) and Rooster (111 Pose Samples).**

You can see in the sample photos contained in Figures 1 and 2 that the computer has attached a skeleton to the body. It decides where the borders of the human body are and even where it is seeing joints and their angles. The computer takes this skeleton and translates it into a form it can understand. This includes lengths of the lines, the angles at the joints, and the distance of the figure from the camera. By using these pose samples, the teachable machine trains and learns the similarities between the Tai Chi poses and uses them to identify other poses (even in live video feed).

Googles program showcases how quick and efficient a computer can describe what it is seeing. Because the teachable machine is free and requires no coding knowledge it offers a great gateway into machine learning and computer vision. For those who chose to try this program for more than learning purposes, Google has also made it possible to download the trained model so that it may be used in other applications and programs. These downloadable files allow for not only professional level companies but also by beginners who are just getting started to gain more knowledge on computer vision, showing how accessible the field truly is.

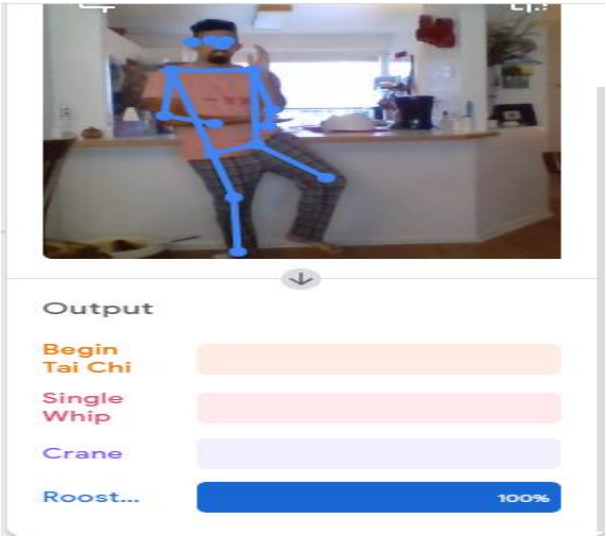


**Figure 3. The trained model can properly derive the Rooster pose from what it is seeing through the webcam.**

*4.1.2 Object Detection Testing*

Moving forward from beginner level programs we wanted to test a type of computer vision that is used more frequently used in real-world scenarios, object detection [3]. This can be described as the ability for a computer to identify objects in an image or video so that it can count, locate, and label them accordingly. After conducting extensive research, we found that this can be done in python primarily with the use of OpenCV, TensorFlow, and YOLOv3 (You Only Look Once). To get better understanding of what each of these libraries do and how they work, we found multiple tutorials and code examples that explained how to install each and run a trained model that shows how the computer identifies the objects it is presented with. Figure 4 shows a capture of the output we received after successfully running the example code we used to test how object detection works and if it could be run on a low-end computer.

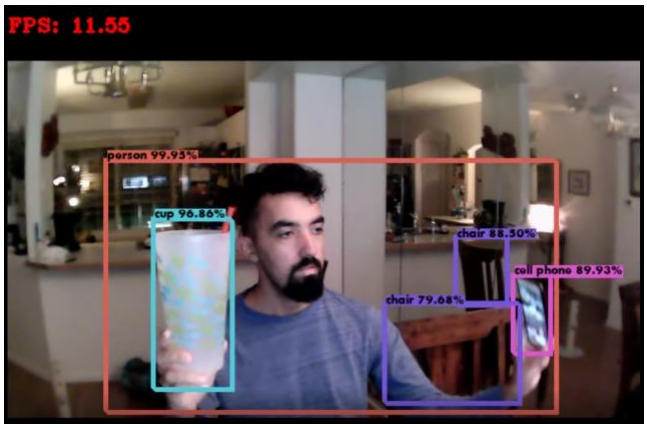


**Figure 4. Bounding Boxes generated by the example that was run with classification and accuracy displayed.**

```
# by default VideoCapture returns float instead of int
width = int(vid.get(cv2.CAP_PROP_FRAME_WIDTH))
height = int(vid.get(cv2.CAP_PROP_FRAME_HEIGHT))
fps = int(vid.get(cv2.CAP_PROP_FPS))
codec = cv2.VideoWriter_fourcc(*FLAGS.output_format)
out = cv2.VideoWriter(FLAGS.output, codec, fps, (width, height))
```

**Figure 5. Setting how we receive width, height, and fps through cv2.**

Looking at the code and reading the documentation [4] for each library we were able to understand more about how this output was generated. TensorFlow was used to train the model, however this step had already been done prior to us testing said model. OpenCV-Python allowed us to connect a webcam to the computer and control how its attributes were handled by the program (height/width of the output video, color correction, and text on screen). We found that using YOLOv3 allowed us to have the computer detect what it was seeing through the camera in real time. This is possible because it is a convolutional neural network that views images as data arrays and compares patterns in it to any predefined classes it has access to. The example we ran had 80 predefined classes that were used to train the program. These classes were provided by Microsoft's COCO dataset which is used to train models for object detection, captioning, and segmentation. With this we were able to see the computer identify multiple common household items with along with the confidence, or accuracy, that the computer had in what it was identifying.

Even though we received results with high accuracy it was not easy to do so. This was because of the weights, which are a parameter that transforms input and decides how much influence it has on the output, that we were using from the COCO dataset. While the weights gave us high accuracy it also slowed down the program a lot, limiting us to twelve frames per second at the most. We found that if we switched to a dataset with less weights, we could run at a much higher frame rate, but it reduced the accuracy of the boxes as well as the accuracy of what the computer thought it was seeing. This was very interesting as it raised questions about how object detection is done in real world scenarios. Is accuracy more important or is the frame rate and speed more important? This led us to take a deeper dive into how object detection is used in real situations which will be talked about in later sections.

## 4.2 Real-World Use

*4.2.1 Facial Recognition*


**Figure 6. Examples of object detection and classification from elevated positions**

In 2010 Facebook rolled out an update to their platform that they hoped would help users tag friends in their pictures. This was done using facial recognition on the images a user posted and allowing the AI do take a dive into the user's friend network and match the photo with that of another user. This was an astonishing use of facial recognition as it showed that an AI could not only focus on a person's face in a photograph but also match it with the same face in multiple different pictures. This of course was not the first use of facial recognition, but it was one of the first times it was used outside of a set of test data that was still able to match two separate photos of the same person and return what it had found without manual intervention. Facebooks use of facial recognition was a driving force behind other companies taking a deeper dive into computer vision for facial recognition and identification purposes, causing a large spike in research and progression [11].

Moving forward from recognizing faces within images, many companies began exploring different ways facial recognition [5] could be used in other applications as well. With the introduction of Face ID in 2017, facial recognition began playing a significant role in the security of individuals. Initially, this new facial recognition system was only used to unlock smart phones and devices but eventually it also made its way into confirming payments and verifying the identity of individuals. With computer vision becoming a part of daily activities, it also found its way into further enhancing protection and security.

*4.2.2 Protection and Security*

Computer vision is also used in a variety of ways by governments and law enforcement agencies whether it's the local police or the FBI. Some examples of how it is used is facial recognition for identifying people of interest and even object recognition to try and predict crimes or even find missing people in dense areas. The next section will be covering the latter of the examples.

Some countries such as Germany and Spain use object detection by ariel drone surveillance for their police officers [6, 7]. They use it specifically in this field as it allows them to track their officers and identify them against pedestrians and other objects so they can make informed decisions during active situations. With normal drone surveillance, the operator must manually check every person in the camera frame and identify the officers. This process can instead be simplified using computer vision by creating training data on features that officers have compared to pedestrians (I.e. officers have certain badges/uniforms they must wear). The surveillance drones can then compare the footage in real time to the training data and correctly identify officers among a ground of people. Within the United States, however, there are stricter privacy laws which prevent the use of surveillance drones around the public. For this reason, surveillance drones are used primarily for search and rescue and disaster management.

Computer vision plays a vital role in disaster management as object detection is how drones can distinguish people from wreckage. Object detection has come far enough that it can still differentiate objects and separate them into their correct classes, despite viewing an area hundreds of feet above the ground (seen in Figure 6 below). This makes it possible for drones to spot people in need of help even if they are surrounded by trees and rubble after any form of disaster. If there is enough of a person visible, the drone can classify them and make it easier for first responders to know where they are amongst the wreckage. When computer vision was first developed its goal was to make a computer see like a human can, with where the technology is being used now you could argue that in some respects it can see even better.

## 5 INTO THE FUTURE

We researched some of the most recent and upcoming examples of computer vision. This includes areas such as construction, manufacturing, delivery, and automated driving. This section will cover some of the information we gathered.

### 5.1 Construction

Construction work is one of the most physically taxing jobs around, even with the machinery that is being used to help. On top of this it is also incredibly time-consuming as it requires constant

inspection of the worksite and any machinery being used. As such it is only natural that there is a large demand for help in the field. Therefore, an engineering and robotics company named Boston Dynamics created a robot called Spot [8] to assist those working on construction sites with general tasks.

Spot is a partially autonomous robot with five cameras on its body giving it a full 360-degree topographical view of its surroundings. The use of these cameras gives Spot the ability to view the terrain around it so that it can make conscious decisions about how to traverse it. On top of the cameras allowing Spot to traverse the terrain around it they also allow it to finish "missions" it is given by its users. These missions are set via fiducials, which are like QR codes, being placed anywhere Spot is needed and then manually running it through the order you would like it to see the fiducials. After Spots path is recorded it can be told to run the mission autonomously.

Some examples of how Spot is used for construction include carrying objects to and from areas of constructions sites, moving rubble out of the way, and measuring distances between support beams quickly and accurately. These examples are all possible with Spots many attachments, the most important being the robotic arm addition. This arm has a camera and IR lasers built into the claw at the end of it that allows Spot to correctly see how far an object is from him and how it needs to interact with the objects it needs to lift or move. During these missions if Spot is presented with obstacles or difficult terrain in its way, the cameras will also allow it to analyze the situation and decide how to move around the obstacle to complete the mission. These jobs are also not limited to only walking from place to place to pick up or drop off loads that it is carrying. Spots cameras also allow for data analysis. An example of this is that Spot can be told the levels that pressure gauges should remain at and can identify if a machine is malfunctioning by viewing the gauges and comparing it to what it has been told is safe.

## 5.2 Automated Vehicles

Multiple automotive companies are attempting to solve the problem of autonomous vehicles and it is leading to great strides in how computers can process images. While there are a few examples of autonomous vehicles (such as Nuro's R2 delivery car [9]), we found that many of the vehicles use deep learning and computer vision already. With this, however, most of these vehicles also must utilize lidar to make up for any mistakes their deep learning algorithms make. By using lidar to make up for accuracy issues, these vehicles become unscalable and inefficient when put in new environments. This is where computer vision shines as Tesla (the leading electric vehicle and self-driving manufacturer) is attempting to create autonomous cars through computer vision alone. These cars have eight cameras on them, three in the front, two on the sides facing forwards, two on the sides facing rearward, and one on the back facing rearward as well [10].

These cameras are set to constantly take videos of their surroundings and to relay them back to the onboard computer. From here the computer can identify what it is looking at by cross referencing the video feeds with the classes it has been given. This form of object recognition works in the same way our example program did in Figure 4, but on a much larger scale. This is where Tesla is working to improve computer vision in their vehicles. With such a large scale it is hard to properly determine distances between the vehicle and its surroundings without using lidar. To solve this issue, they began treating it as a supervised learning problem and are now training their cars' on-board computers to make quick on-the spot calculations based on what they see and the extensive annotated data they have already. This implementation allows much more scalability and consistency than its lidar counterpart and shows yet another implementation of how computer vision is used.

## 5.3 Medical

When a patient's identity gets confused with a different patient, there becomes a risk of giving the wrong treatment/medicine to them (putting the patients' lives at stake). In the future, however, medical industries will be able to use facial recognition systems to better identify patients and reduce this risk.

Alongside accidentally mistaking patients' identities, often doctors and nurses also make mistakes when analyzing medical scans/images. With these mistakes, they can sometimes incorrectly identify a patient's diagnosis (whether it is a false positive or not) which often can be fatal. With computer vision, however, it can help assist doctors and nurses in identifying and classifying illnesses before they even appear on any medical scans due to the general nature of computer vision being good at identifying/classifying patterns in images. Computers can identify illnesses by receiving and analyzing the images of previous patients with similar illnesses and using these as training data to learn from. Computers then compare the training data to new scans of patients (testing data) and analyze any similarities and differences between the training and test data. If successful, computers can correctly classify if a patient has an illness quicker and more accurately than a medical professional. The only drawback to implementing computer vision within the medical industry is that it can incorrectly conclude a false negative case (did not detect an illness when it was present) so further research into computer vision is needed to increase the accuracy before medical fields can fully adopt it. With more data being obtained from patients (I.e., more scans are performed over time), however, the accuracy of computers correctly analyzing an illness using scans will undoubtably increase.

# 6 Conclusion

In this paper we have discussed the background of computer vision, some programs that can be used to better understand how it works, the current state of computer vision implementations, and how these implementations can be utilized in the future. The

research that was put together here was done to better understand how the field of computer vision came to be and why it will be so important moving forward.

From the creation of computer vision in the 1960's and its initial goal of mimicking human vision and identifying objects in images to where it is now in 2021, there has been many significant changes over the years. Nowadays, computer vision can not only identify objects within an image they are given but they can also do so at such a level that technology has been implemented into everyday use in many different professions. Whether a person is an expert in technology or barely has any knowledge of it, computer vision has evolved over time and now it has become much easier to use and understand so it can now be used in other areas of studies. Of these, one of the areas that benefits the most from the implementation of computer vision, is security (using facial recognition) because it currently helps rescuers and law enforcement agencies find people in environments that would be difficult to find them otherwise. With these significant benefits to security, the end of computer vision is nowhere in sight as there are plans to implement self-driving cars and the integration of it into the medical diagnostics of patients. Through our research of learning about the history, present, and future use of computer vision, we can expect to see more people (and other fields) to begin adopting it more.

## REFERENCES


[1] Pulsar Editor. 2019. A history of computer vision & how it lead to 'vertical ai' image recognition. (March 2019). Retrieved December 1, 2021 from https://www.pulsarplatform.com/blog/2018/brief-history-computer-vision-vertical-ai-image-recognition/#:~:text=When%20computer%20vision%20started%20to,to%20artificially%20intelligent%20image%20recognition

[2] theAIGuysCode. 2021. Build and Train a Machine Learning Model without Code in Under 10 Minutes. (January 2021). Retrieved December 1, 2021 from https://www.youtube.com/watch?v=7ryde8Cz-Cs

[3] theAIGuysCode. 2020. Real-time yolov3 object detection for webcam and ... - youtube. (March 2020). Retrieved December 1, 2021 from https://www.youtube.com/watch?v=p44G9_xCM4I

[4] theAIGuysCode. 2020. TheAIGuysCode/object-detection-API: Yolov3 object detection implemented as apis, using TensorFlow and Flask. (September 2020). Retrieved December 1, 2021 from https://github.com/theAIGuysCode/Object-Detection-API

[5] Anon. 2021. A brief history of facial recognition - NEC New Zealand. (May 2021). Retrieved December 1, 2021 from https://www.nec.co.nz/market-leadership/publications-media/a-brief-history-of-facial-recognition/

[6] Naveen Joshinav. 2021. How computer vision can aid law enforcement in smart cities. (June 2021). Retrieved December 1, 2021 from https://www.allerin.com/blog/how-computer-vision-can-aid-law-enforcement-in-smart-cities

[7] Michael. 2021. Your guide to computer vision in drone technology. (October 2021). Retrieved December 1, 2021 from https://keymakr.com/blog/computer-vision-in-drone-technology/

[8] Knowledge. 2021.(October 2021). Retrieved December 1, 2021 from https://support.bostondynamics.com/s/article/Getting-Started-with-Autowalk

[9] Anon. 2021. Nuro's technology. (2021). Retrieved December 1, 2021 from https://www.nuro.ai/technology

[10] Matt Pressman. 2020. Learn how elon musk has evolved Tesla's autopilot hardware over the years. (September 2020). Retrieved November 25, 2021 from https://evannex.com/blogs/news/the-history-of-tesla-s-autopilot-hardware-and-how-it-s-evolved#:~:text=Cameras%3A%20Eight%20total%20%E2%80%93%20three%20front,and%20one%20rear%2Dfacing%20camera.

[11] Jason Kincaid. 2010. Facebook uses face recognition to help tag photos. (December 2010). Retrieved November 23, 2021 from https://techcrunch.com/2010/12/15/facebook-uses-face-recognition-to-help-tag-photos/?guccounter=1