

---

# Structured Light 3D Surface Sensing

---

**Ian J. Maquignaz**

Electrical & Computer Engineering  
Queen's University  
Kingston, ON K7L 3N6  
ian.maquignaz@queensu.ca

## Abstract

The physical world is an enigma to a vast majority of the computing systems. Though this remains of little consequence for many applications, for a growing number of systems from personal devices to manufacturing, a visual acclimation is an emerging and inherent requirement. In recent years, the machine vision community has worked diligently to meet this growing requirement and has created a new generation of devices and methodologies to provide computing platforms an acuity for 3D sensing. This work provides a literature review of standalone and embedded sensors, amalgamating research with pertinence to 3D surface imaging using structured light (SL). To provide context and insight into the future, the review summarizes forerunner research and introduces 3D surface imaging as a candidate application which may benefit from a paradigm shift towards unified machine and deep learning solutions.

## 1 Introduction

The physical world is an enigma to a vast majority of the computing systems. Though this remains of little consequence for many applications, for a growing number of systems from personal devices to manufacturing a visual acclimation is an emerging and inherent requirement. The machine vision community has worked diligently to meet this growing demand and in recent years has created a new generation of devices and methodologies to provide 3D visual sensing.

Once prominent and popular devices such as the Microsoft Kinect [1] are now passé, and are rapidly being replaced by specialized standalone and embedded sensors. As equivalents, Intel has created the RealSense depth sensor family which includes stereo-vision, stereo-infrared (IR), and LiDAR variants [2] with significantly improved capacities. Industrial and commercial sensors are growing in popularity, with specialized devices available from Zivid, Photoneo, Sony, Aeye, Basler, LMI, LUCID, Keyence and many others manufacturers [3–10]. Sensors embedded in personal and portable devices are also becoming more prominent, with mobile phones using depth mapping technology for applications from biometric security to image segmentation [11, 12].

But the paradigm of computer vision is changing. Parameters and coefficients, once manually tuned by an omniscient designer, are now a harrowing reminder of the circumstantial environments and inherent bias with which they were conjured. Reliable and versatile designs are now paramount in markets where industry and consumers alike are increasingly intolerant of errors and malfunctions (for example the iPhone X release [13] and fears of misidentification by government and law agencies [14, 15]). Machine Learning (ML) provides an opportunity for a paradigm shift in computer vision through example-based learning and discovery of data features and characteristics. Advances in computational

platforms now allow for the training and testing of models across previously unfathomable stores of data. To this avail, companies such as Silicon Software offer integration services for the embedding of deep learning machine vision solutions into FPGAs [16].

Methodologies and devices for Structured Light (SL) sensing of the physical world are heavily driven by hand-designed features and prior knowledge injected by the designer. Such designs operate with the presumption and prejudice of an omniscient creator, with the implementation of a particular and potentially circumstantial solution. Through ML, the potential exists to explore adaptive, abstract and alternative solutions across extensive data, offering the potential for innovation in the encoding and decoding pipeline of SL sensing. Though complete 3D sensing without occlusion is likely to remain unachievable without multiple sensors and/or multi-spectrum analysis, the potential exists to create a new generation of 3D surface sensors for projection mapping (dynamic & static) and 3D surface measurement. By offering higher resolution, accuracy, refresh-rates, and versatility, the sensors enable a multitude of different applications, including (but not limited to) part inspection, self-driving vehicles, and augmented reality displays.

The following literature review is an amalgamation of research with pertinence to 3D surface sensing using structured light (SL). Literature will summarize forerunner works and research in structured light sensing, and introduce deep neural network (DNN) models which could be integrated to create unified DNN & SL solutions. Where possible, sensing methodologies implemented with ML are reported and emphasized.

## 2 3D Surface Imaging

Traditional cameras and image sensors sense the physical world as two-dimensional (2D) planar images with implicit loss of three-dimensional (3D) depth information. Through multitude of different techniques depth can be recovered using mono-, stereo-, and multi-camera/sensor configurations. In its simplest of forms, the 3D representation of a scene can be inferred from correspondences within the content of a scene, but it should be emphasized that any recovery of 3D representation is partial.

3D surface imaging systems are design-oriented around measuring the  $(x, y, z)$  coordinates of points on an object's surface, creating point clouds of the form  $P_i = (x_i, y_i, z_i), i = 1, 2, \dots, N$ . In this regard, no volumetric or internal structure data is collected, and data is often best interpreted as a depth-map. For true 3D imaging and the generation of data isomorphic to CAD models, the 3D representation of an object can be recaptured through systems such as ultrasound, X-ray, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), which capture volumetric pixels (voxels) and/or internal structure through transmissive probing (looking at what passes through an object rather than what reflects).[17, 18]

Despite this limitation, 3D surface imaging systems can include capacities beyond those of the aforementioned true 3D sensors, enabling augmentation of point-cloud data. Simple versions of augmentation include the capacity to measure surface albedo (reflectance  $f$ ) and surface color (as scalar  $(r_i, g_i, b_i)$  if using the RGB color model), producing complex point clouds such as  $P_i = (x_i, y_i, z_i, r_i, g_i, b_i, f_i), i = 1, 2, \dots, N$  [17, 19]. Further extensions have been proposed by I.J.M, allowing for the recovery of perspective transformation coefficients of a projected pattern, creating points clouds of the form  $P_i = (x_i, y_i, z_i, \chi_{xi}, \psi_{xyi}, \tau_{xi}, \chi_{yi}, \psi_{yxi}, \tau_{yi}, \chi_{zi}), i = 1, 2, \dots, N$  [20].

The following subsections summarize methodologies and research in 3D surface imaging. To improve coherence and better encompass the field, methods are broadly categorized as passive correspondence based (PCBV), active correspondence based vision (ACBV), or other. For the purpose of this work, the focus and emphasis is on projector-camera systems (PROCAM) operating in the visual light spectrum. Though empirically 3D surfacing sensors can be created using a variety of different technologies, versatile and flexible hardware platforms often offer cost-saving advantages to both consumer and industry applications.

## 2.1 Passive Correspondence Based Vision

As described by Luhmann et al. [21], a singular calibrated camera can be used to create a 3D model by inferring the spatial correspondences created using a probing device. This is possible, given a calibrated camera has precisely known intrinsic and extrinsic parameters, including the camera's principle point, image plane, and radial distortion, allowing for mono- and stereoscopic measuring of discrete points (see Figure 1 for reference) [22]. Calibration parameters are known constants in metric cameras used in photogrammetry, but conventional non-metric cameras can be manually calibrated using a reference field (such as a chessboard) and methodologies such as Zhang's method [23]. To quantize the physical space in the scene of a singular image, probing devices must be used to create correspondence points with known positions and orientations in 3D space. These correspondences create triangles allowing for the solving of metric-depth by geometric properties, enabling a broad field of methods often referred to in literature as triangulation-based methods.

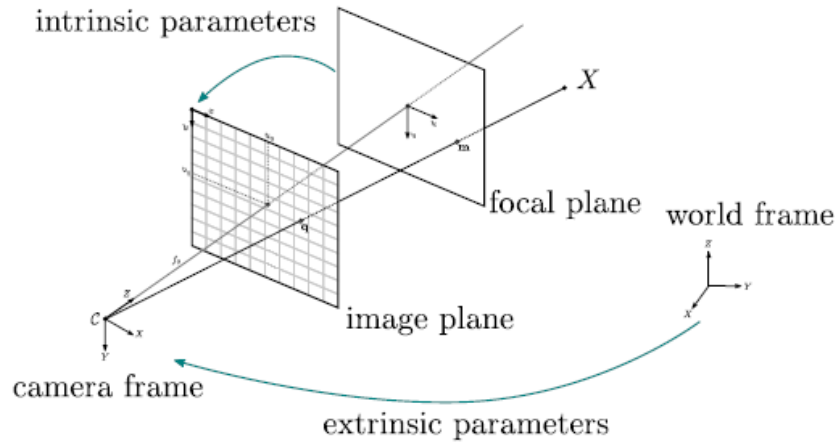


Figure 1: Pinhole Camera [24]

Though precise, the requirement for a physical probing device to create spatial correspondences is impractical and prohibitive for applications such as automation. Methodologies which infer correspondences directly from a scene are preferable and can be achieved through a multitude of approaches. When using a singular camera, the depth information in a scene is lost when a 3D structure is captured as a 2D retinal plane, but can be recovered from correspondences created using an aforementioned probing device, or, structure from motion. When using two or more cameras together in a stereo- or multi-camera configuration, the 3D representation of a scene can be recovered by referencing the 2D retinal planes against each other to identify correspondences [25]. The configuration of stereoscopic methods is analogous to those illustrated in Figure 2 through cameras  $C$  and  $C'$  with respective image planes  $I$  and  $I'$ .

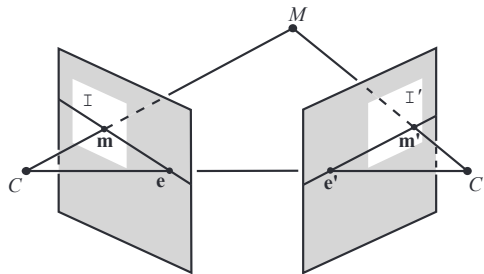


Figure 2: Epipolar Geometry

To accurately reconstruct the 3D representation of a surface, many factors need to be taken into consideration including the vergence (angle between  $C$  and  $C'$ ) and baseline (distance between  $C$  and  $C'$ ), but also the characteristics of the scene. Given structure from motion and the referencing of 2D retinal planes rely on scene content for correspondences, unfavorable morphology and surfaces exhibiting limited (or devoid of) texture are problematic [26]. Furthermore, as illustrated by the stereoscopic image pair in Figure 3a, identifying correspondences between 2D retinal planes can be a complex task.

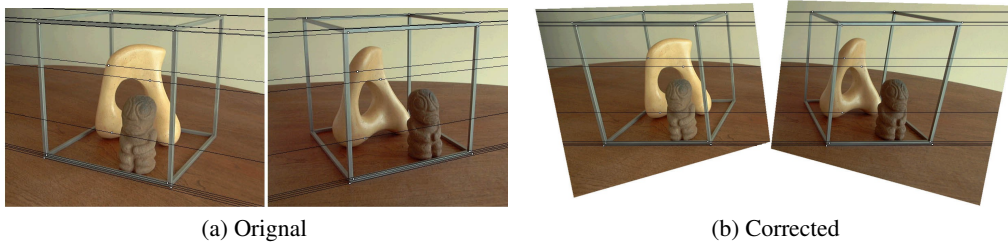


Figure 3: Rectifying Homography [27]

Numerous approaches have been proposed to reduce the complexity and computational expense of stereoscopic image analysis, including Relative Stereo Disparity (RSD) [25] and phase-based algorithms [28]. With relevance to recent works in ACBV in Subsection 2.2, Loop and Zhang’s method of rectifying homographies is of particular interest. As illustrated through Figure 3a, lens calibration can be used to independently rectify image planes for radial and tangential distortion. This ‘flattens’ the image planes, but does not converge the images to a common viewpoint and therefore correspondence searches along epipolar lines follow skew lines in image space. Taking advantage of known epipolar constraints (as shown Figure 2), Loop and Zhang proposed the application of rectifying homographies, allowing for epipolar lines to be parallel and aligned with the coordinate axis [27]. This is illustrated in Figure 3, where a rectifying homography has been applied to the images in Figure 3a to produce the images in Figure 3b with epipolar lines aligned to the horizontal axis. Stereoscopic analysis of Figure 3b is thus simplified, as in the absence of skewing, disparity is a measure of displacement along a 1D epipolar scan-line. The approach used by Loop and Zhang to produce the rectifying homography is by decomposing each homography into a specialized projective, similarity and shearing transform. This process is prone to erroneous results in part due to the accounting for distortion through forced affine qualities on homographies [29]. Mallon and Whelan improved on this approach, proposing inference of the rectifying homography from fundamental matrices and taking into consideration more than one localized region of an image. This proposed improvement exhibits a rectification accuracy equal to the error in the fundamental matrix estimation [29].

Though older at this stage of evolution in 3D surface imaging sensors, passive correspondence based vision systems (PCBV) have not lost their prominence in industry and consumer products. PCBV systems are versatile and can be tailored to provide accurate real-time operation on limited hardware [25][28]. As a result, PCBV systems are affordable and convenient solutions, persisting today through products such as the PlayStation Camera, enabling tracking for hands-free system navigation and facial recognition for auto-login on PS4 [30].

## 2.2 Active Correspondence Based Vision

Active Correspondence Based Vision (ACBV) methods are often analogous to their passive counterparts, exhibiting similar operating characteristics and limitations. In contrast to passive methods, ACBV mitigates dependence on correspondences scene content by actively projecting spatial or spatiotemporal patterns by modeling projectors as inverse cameras. As a benefit of this close relationship, the calibration of ACBV sensors generally mirrors that of previously mentioned PCBV sensors. The following subsections summarize approaches structured light to 3D surface imaging and the characteristics which enable their imperceptible operation.

### 2.2.1 Structured Light

Structured light is the active illumination of a scene with specially designed spatially and/or temporally varying intensity patterns. Though active illumination is often synonymous with projection, it is important to dissociate projection from the intuitively insinuated

consumer projector technology. As demonstrated by many proposed sensors and consumer products, projection can be achieved in ways other than the presupposed LCD, DLP or LED projector technology. For example, the Kinect V1 uses a IR laser shone through a diffraction grating to project a set of static IR dots on a scene [1]. In terms of a more radical design, Radu et al. propose using a catadioptric camera to sense lines emitted by an omnidirectional laser source to sense depth over a much wider field of view [31].

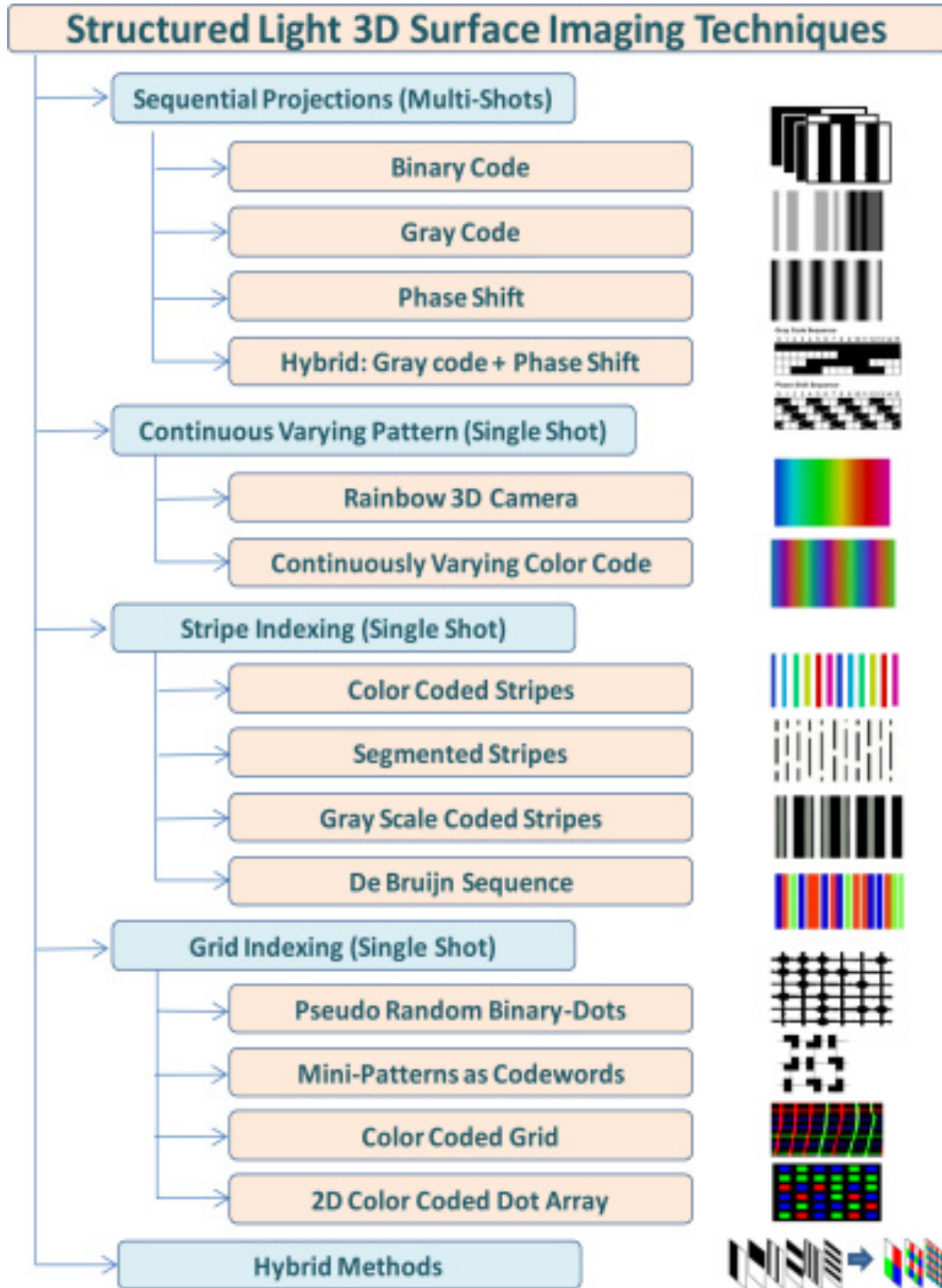


Figure 4: ©2011 Optical Society of America [17]



As illustrated in Figure 4 by Geng [17], active illumination can be achieved through a multitude of approaches which can be broadly classified into a number of pattern families. These include grid (M-array), stripe, continuous and coded patterns, which can be discretely, spatially and/or temporally multiplexed. Despite the variety in approaches, no family of patterns is indicative of an optimal depth sensor and each family offers varied density of recovered depth points, ability to handle moving subjects, and computational efficiency. As such, families of patterns can be mixed to create hybrid approaches which draw on the strengths of each pattern family. Differences between pattern types is exemplified by Salvi et al. in Figure 5, where patterns for profilometry are further categorized as either discrete or continuous [26]. As illustrated, Salvi et al. compiled a table summarizing works from 1982 to 2009, including methodology characteristics, including:

- *Shots*: Number of patterns necessary to profile a surface
- *Cameras*: Number of cameras required by the methodology
- *Axis*: Number of axes referenced in pattern generation
- *Pixel Depth*: Binary(B), grayscale (G), or Color(C) pixel representation
- *Coding Strategy*: Periodic (P) or absolute (A) pattern positioning
- *Subpixel acc.*: Precision level as either sub-pixel (Y) or meta-pixel (N) accuracy
- *Color*: Support for textured objects (Y) or surfaces devoid of texture (N)

Some difference may be noted between illustrations of structured light techniques, including the portrayal of phase shift as a multi-shot technique in Figure 4 and as a single- or multi-shot technique in Figure 5. The categorization by Salvi et al. is correct, as multi-phase patterns are often multi-shot, yet single-shot phase-shifting and solid-state fringe patterns have been previously described in many works. This includes the fringe pattern proposed by Gong and Zhang to achieve very high depth frame rates [32]. In a broader sense, what is illustrated by the differences in classification between the two figures is the lack of universal truths, where spatial and/or temporal patterns are often designed to meet specific operating characteristics. This phenomenon is illustrated in Figure 5, where the characteristics exhibited by proposed sensors have little correlation to pattern classification. As such, patterns do not necessarily conform to any fixed categorization and, along with the sensors they enable, are not easily compared or benchmarked against one another. This is will be discussed further in a later section.

Though seldomly comparable, there are a number of design characteristics which should be observed in ACBV sensors. Depending on the desired precision, throughput, and hardware-costs, selections can be made to increase performance. In terms of maximizing precision, pattern selection should prioritize spatial multiplexed, continuous and other patterns which maximize usage of the spectrum (color, if using an RGB projector) and offer the capacity for sub-pixel precision in depth measurements. Inversely, throughput is maximized by selecting patterns requiring minimal processing, thus monochrome single-camera with simple patterns such as grid (M-array), stripe, binary and others with correspondences searches along a singular axis. Alternatively if hardware costs are to be minimized, selection should be inline with maximized throughput though prioritizing solely single-shot patterns such that projection can be performed by a static mechanical emitter.

These design characteristics are not universal truths and as such there are a number of weaknesses to ACBV which designers should be aware of. Much like any flashlight, structured light 3D imaging systems have limited energy in projection [17]. As such, ACBV sensors can only sense the physical world touched by projections with sufficient intensity. In time-of-flight (ToF) sensors, sensor range can be scaled by controlling emitter intensity, but no such varied-intensity ACBV sensor is known to have been successfully demonstrated. In addition to challenges with distance, most ACBV sensors are sensitive to textured and specular surfaces to a varying degree. Though radiometry can be applied to static scenes, no successful solution is known to have been demonstrated for dynamic scenery. As observed by Li et al., projector positioning greatly influences performance including clarity and illumination [33]. This also inherent to the parallax in triangulation-based 3D surface imaging, where 3D surface features can be occluded due to perspective [17]. Though this

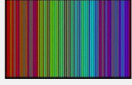
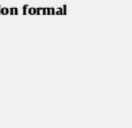
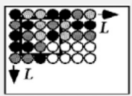





Discrete									
Spatial multiplexing									
De Bruijn									
	Boyer	1987	1	1	1	C	A	Y	N
	Salvi	1998	1	1	1	C	A	Y	Y
	Monks	1992	1	1	1	C	A	Y	N
	Pages	2004	1	1	1	C	A	Y	N
Non formal									
	Forster	2007	1	1	1	C	A	Y	N
	Fechteeler	2008	1	1	1	C	A	Y	N
	Tehrani	2008	1	1	1	C	A	N	Y
	Maruyama	1993	1	1	2	B	A	N	Y
	Kawasaki	2008	1	2	2	C	A	N	Y
	Ito	1995	1	1	2	G	A	N	Y
	Koninckx	2006	1	1	2	C	P	Y	Y
M-array									
	Griffin	1992	1	1	2	C	A	Y	Y
	Morano	1998	1	1	2	C	A	Y	Y
	Pages	2006	1	1	2	C	A	Y	N
	Albitar	2007	1	1	2	B	A	N	Y
Time multiplexing									
Binary codes									
	Posdamer	1982	> 2	1	1	B	A	N	Y
	Ishii	2007	> 2	1	1	B	A	N	N
	Sun	2006	> 2	2	1	B	A	Y	Y
N-ary codes									
Caspi	1998	> 2	1	1	C	A	N	N	
Shifting codes									
Zhang	2002	> 2	1	1	C	A	Y	N	
Sansoni	2000	> 2	1	1	G	A	Y	Y	
Guhring	2001	> 2	1	1	G	A	Y	Y	
Continuous									
Single phase									
Shifting (SPS)									
	Srinivasan	1985	> 2	1	1	G	P	Y	Y
	Ono	2004	> 2	1	1	G	P	Y	Y
	Wust	1991	1	1	1	C	P	Y	N
	Guan	2004	1	1	1	G	P	Y	Y
Multiple phase									
Shifting (MPS)									
	Gushov	1991	> 2	1	1	G	A	Y	Y
	Pribanić	2009	> 2	1	1	G	A	Y	Y
Frequency multiplexing									
Single coding frequency									
	Takeda	1983	1	1	1	G	P	Y	Y
	Cobelli	2009	1	1	1	G	A	Y	Y
	Su	1990	2	1	1	G	P	Y	Y
	Hu	2009	2	2	1	C	P	Y	Y
	Chen	2007	1	1	1	C	P	Y	N
	Yue	2006	1	1	1	G	P	Y	Y
	Chen	2005	2	1	1	G	P	Y	Y
	Berryman	2008	1	1	1	G	P	Y	Y
	Gdeisat	2006	1	1	1	G	P	Y	Y
	Zhang	2008	1	1	1	G	P	Y	Y
	Lin	1995	2	1	1	G	P	Y	Y
	Huang	2005	> 2	1	1	G	P	Y	Y
	Jia	2007	2	1	1	G	P	Y	Y
	Wu	2006	1	1	1	G	P	Y	Y
	Spatial multiplexing								
Grading									
	Carrhill	1985	1	1	1	G	A	Y	N
	Tajima	1990	1	1	1	C	A	Y	N
			Shots	Cameras	Axis	Pixel depth	Coding strategy	Subpixel acc.	Color

Figure 5: Structured Light Classification by Salvi et al. [26]

parallax can be mitigated by incorporating multiple cameras and emitters, most ACBV sensors are incompatible with multiple overlapping projection sources.

As previously mentioned, the calibration of ACBV sensors mirrors that of traditional PCBV, though the process can be less sympathetic. Autofocus has been implemented in cameras since 1977 with the first 35mm SLR in 1981 [34] and yet, despite it existing, off-the-shelf projectors seldomly offer the sensory hardware necessary for self-calibration. The cameras and projectors in ACBV sensors often achieve the required acclimation for calibration through the intervention of a human expert, but today's requirement for precision, versatility, and dynamic operation often make that incompatible. As such, commercial and industrial ACBV sensors such as the Intel RealSense and Zivid One come with fused optics, controlling only minimal hardware characteristics such as exposure and aperture [2, 19]. Overcoming fixed orientation is an active area of research, with examples of dynamic self-calibration and its enabling characteristics for robotics explored by Wieghardt and Wagner [35].

For a curated list of commercially available state-of-the-art including ACBV systems, *Vision Systems Design* publishes an annual ranking of disparate and innovative technologies, products, and systems found in machine vision and imaging [10]. This publication is a thorough and comprehensive list of technologies commercially available to industry on a global perspective, providing great insight into emerging trends in 3D surface imaging and machine vision.

## 2.2.2 Imperceptible Structured Light

ACBV sensors operating in industrial settings often require unrestricted and uncompromised use of the illuminated spectrum to sustain desired precision, resolution and frame rates [19]. In the theater of an office or living room, such operation is hindering and prohibitive of user experience with Augmented Reality (AR) and other applications. To overcome this limitation, ACBV sensors have been adapted to offer imperceptible operation by increasing operating speeds and projecting in spectrums which exceed the range of human perception.

As demonstrated by consumer products such as the Kinect V1 and RealSense D435, a convenient approach to imperceptibility is shifting operation from the visible spectrum to alternatives such as the infrared (IR) spectrum [1, 36]. Operating at a wavelengths beyond human capacity, IR operation offers some significant advantages over sensors operating in visible light. Imperceptibility aside, IR sensors are insensitive to ambient lighting and textured surfaces as a result of artificial light containing limited IR radiation and few indoor textures being IR reflective [1, 37, 38]. As described by Laudau et al., IR sensors suffer from drawbacks with specular surfaces similarly to most depth sensors. In this scenario, reflecting light results in erroneous depth information, effectively 'washing out' texture and structured light [39]. Though there is no exclusivity to the IR spectrum, examples of ACBV sensors in alternate spectrums are predominantly IR.

Operation in the IR is both an advantage and a drawback. In terms of hardware, visible light sensors offer versatility to AR systems where, in addition to sensing, hardware can display and capture video for human consumption. Therefore, it can be said that visible light sensors offer reduced hardware costs for AR systems, whereas systems using IR sensors (such as the Kinect) would require additional hardware for displaying and capturing content [1]. That being said, though not all visible light sensors can sense imperceptibly whilst simultaneously displaying content, all PROCAM hardware inherently offers this dual functionality. As an example of this, iLamps projects a chessboard to calibrate before projecting digital content for user consumption [40].

To retain this dual functionality in the visible spectrum, alternative modes of operation have been proposed using high-speed projection, dithering, and flicker fusion, which modulate light faster than perceivable by human psychophysical responses. To achieve this, a threshold is necessary for maximum human capacity to interpret visual stimuli. In 2007 Kuroki et al. performed a psychophysical study on motion-image quality, finding that blur and jerkiness from motion were no longer perceptible at frame rates upwards of 250 fps [41]. This is impractical, as standard cinema video operates at 24 fps and most off-the-shelf



projectors have a refresh rate between 30fps and 180fps. Furthermore, high-contrast patterns can be visible to human observers at greater than 60Hz, resulting in visual stress and distraction [42]. As a result, most works employ flicker fusion techniques which can have low contrast and lower operating frequencies.

---

**Equation 2.1** Flicker Fusion [43]

---

$$I_1^+ = I + \Delta Pattern \quad (2.0.1)$$

$$I_2^- = I - \Delta Pattern \quad (2.0.2)$$


---

As shown in Equation 2.1, flicker fusion works by decomposing an image  $I$  into complementary images  $I^+$  and  $I^-$  injected with a weighted ( $\Delta$ ) pattern. When the images are projected in sequence at a high frequency, the images aggregate and the human observer perceives the unaltered image  $I$  [43]. Though flicker fusion has been demonstrated as a method, there remains incongruities between works with the minimum operating frequency and maximum pattern weights. For example, some works ([44, 45]) cite Raskar et al. [46] and target aggregated uniformity in projection at 60Hz, while others ([43, 45]) cite Park et al. for critical fusion frequency at  $>75\text{Hz}$  [47]. That being said, it should be noted that Raskar et al. provide no citation or validation, and Park et al. made an error in their citations, though the apparent citation of a 1986 manual on human physiology by Andrew Watson contains no mention of ‘critical fusion frequency’ or 75Hz [48].

Using the same principle as flicker fusion, dithering was proposed in 2004 by Cotting et al. to take advantage of digital mirroring devices (DMD) core to DLP projector technology [44]. By controlling the flipping of DMD micro-mirrors, Cottin et al. proposed modulating patterns into the DLP’s aggregated rendering pipeline, enabling encoding with no discernible impact on visualized content. Though limited by the technology available in 2004 and unable to demonstrate the full potential of the concept, Cotting et al. embedded reflected binary codes (RBC) within an image with a reduced dynamic color range. This successfully embedded the patterns imperceptibly but degraded the visual quality of projected media. DLP technology remains an active area of interest for hardware ‘tricks’, with recent works such as Cole et al. taking similar advantage of the DLP hardware to project three patterns per visualized frame by recognizing the DLP projector color-wheel frequency [45].

### 2.3 Other Methodologies

Given the impracticality of correspondences based vision in environments with unfavorable scene content, alternative approaches mitigating the requirement for geometric congruity have been sought. Though not all are common or known to be commercially available, alternative approaches have been proposed for 3D surface imaging using moving optics, plenoptics cameras and time-of-flight devices.

**Moving Lenses:** Approaches to depth sensing using moving lenses is an interesting and emerging area of research. In 2015, Amin and Riza introduced a system capable of reconstructing depth by using an electronically controlled lens to vary focal length [49]. This proposed system used a laser source and was compared to a standard 1024x768 Philip’s LC4345 3LCD projector with its focus fixed to 56.8cm from the camera. Over a depth of 30cm to 100cm, Amin and Riza found their approach estimated depth with 6%-10% error in comparison to 10%-40% error with the conventional projector. A similar methodology was proposed by Iwai et al. in 2015, placing an electronically controllable lens in front of a 3LCD Epson EMP-1710 1024x768 projector and capturing the result with a 10.1MP Cannon EOS Rebel XTi [50]. This approach produced a mean depth difference of 0.21mm in comparison to a 0.26mm mean depth difference from the same configuration with fixed optics and grey-code patterns.

**Plenoptic Lenses:** Plenoptic cameras (or light-field cameras) are a niche and possibly re-emerging area of research [18, 49]. Introduced by Adelson and Wang in 1992 as a means

of achieving ‘single lens stereo’, plenoptic lenses have a unique conoidal structures which create a parallax and allow for the simultaneous capture of multiple viewpoints [51]. The traditional design of plenoptic lenses creates a defocused array of images with as few as one in-focus pixel, which is greatly impractical. As a result, this design was improved in 2009 by Lumsdaine and Georgiev to enable the capture high resolution images alongside other applications [18, 52]. As noted by Amin and Riza, one specialized function of plenoptic cameras is the capacity for the degree of focus/defocus to be measured for axially separated objects from a single image [49]. This provides an aptitude for autofocus exceeding SLR cameras [34].

**Time-of-Flight:** Time-of-flight (ToF) technology is an alternative approach to 3D surface imaging with predominance in outdoor consumer and industrial settings, but emerging recently as a viable sensors for indoor consumer applications. With that in mind, Intel just released the RealSense LiDAR L515, offering up to a 1024x768 depth resolution over a range of 0.25m to 9m at 30FPS (23.6 million depth points per second)[2].

ToF devices function by illuminating a scene with a modulated light source and inferring depth from analysis of reflected light. This can be achieved through pulsed or continuous-wave signal modulation though, given light travels as  $3 \times 10^8$  m/s, 1mm precision would require the generation of impractical 6.6 picosecond pulses. As a result, continuous-wave signal modulation is favorable, with multi-frequency modulation necessary to sustain disambiguated operation at high frequencies. ToF devices offer many advantages over ACBV sensors, including an invariance to scene morphology and a capacity to control sensor range through active control over illumination energy. As a result, where stereo-vision sensors experience a depth resolution error as a quadratic function of increasing distance, ToF sensors can offer scalable range by controlling illumination and using the reflected intensity as a measure of confidence. These operating characteristics and others are compared to stereo-vision and structured light as illustrated in Figure 6. One known drawback of ToF design is that frequency is inversely proportional to sensing distance. As such, low frequencies can be used for long-range scanning, but high frequencies (and thus fast electronics) are necessary for short-range applications such as indoor use. This relationship can be cost-prohibitive, and is likely correlated to the late emergence of commercially available ToF devices for consumer applications. [53]

CONSIDERATIONS	STEREO VISION	STRUCTURED-LIGHT	TIME-OF-FLIGHT (TOF)
Software Complexity	High	Medium	Low
Material Cost	Low	High	Medium
Compactness	Low	High	Low
Response Time	Medium	Slow	Fast
Depth Accuracy	Low	High	Medium
Low-Light Performance	Weak	Good	Good
Bright-Light Performance	Good	Weak	Good
Power Consumption	Low	Medium	Scalable
Range	Limited	Scalable	Scalable

Figure 6: Comparison of 3D Imaging Technologies. Copyright© 2014, Texas Instruments Incorporated [53]

**Discussion:** As noted by Amin and Reza, measurements using moving and plenoptic lenses are dependent on scene morphology and lighting for successful operation [49]. As a result, moving and plenoptic lenses are unlikely to replace ACBV sensors in the near future. ToF devices on the other hand are rapidly growing in popularity and decreasing in price. The technology is also important to mention as the operating speeds of projector technology increases, as someday soon it may offer the capacity to modulate light for ToF devices. More specifically, DMD dithering on DLP projectors as proposed by Cotting and

Fuchs could potentially be used to modulate a signal to enable a ToF device, creating an unprecedented high-speed and high-resolution 3D surface imaging sensor [44].

## 2.4 Performance Assessment & Qualitative Metrics

Throughout the works perused, a variety of different metrics have been reported to characterize and quantize the operating performance of proposed systems. For the purposes of qualitative assessment and comparison, uniformity and conformity in reported metrics is a desirable quality. Idealistic as that concept may be, the metrics commonly reported for proposed 3D surface imaging sensors often have heavy bias towards hardware and are poor measures of a methodology's performance.

**Performance Assessment:** Performance of proposed systems is frequently quantized as a measure of throughput in recoverable depth Points Per Second (PPS) as expressed in Equation 2.2.1. Drawing for optical communication, this is reflective of the baud rate (symbol rate) as expressed through Equation 2.1.1 [54].

---

### Equation 2.2 Optical Communication Quality Assessment

---

Baud Rate (Symbol rate):

$$Bd = Bits * Channels * OperatingFrequency \quad (2.1.1)$$

Spectral Efficiency:

$$SpectralEfficiency\left(\frac{bps}{Hz}\right) = \frac{ChannelThroughput(bps)}{ChannelBandwidth(Hz)} \quad (2.1.2)$$


---

Though PPS and Baud metrics provide insight into speed of operation, there is an inherent reflection of hardware operation and misrepresentation of methodology performance. Specifically, there is a bias towards works using platforms with higher computational performance, higher resolution optics and support for higher framerates. Drawing again from optical communication to provide clarity, spectral efficiency as expressed in Equation 2.1.2 is a metric used to quantize the rate at which information can be transmitted over a given bandwidth [55]. This metric can be adapted to represent density of points as shown in Equation 2.2.2.

---

### Equation 2.3 Optical Communication Quality Assessment

---

Points Per Second (PPS):

$$PPS = Correspondences * Channels * framerate \quad (2.2.1)$$

Point Density:

$$PointDensity\left(\frac{PPS}{M * N * Channels}\right) = \frac{ChannelThroughput(bps)}{ChannelBandwidth(Hz)} \quad (2.2.2)$$


---

As a measure of consistency and accuracy in the operation of proposed systems, many works make use of plane-fitting. Using this technique, the proposed system is used to iteratively collect a point cloud of an ideal (likely lambertian) planar surface set at a variable depths. Once collected, a plane is fit to each point cloud and error is estimated as measure of point conformity to the plane. As an example, Fanello et al. compare five approaches to depth mapping by calculating the Root Mean Squared Error (RMSE) from plane fittings in a uniform environment and planes 20-350cm from the sensors [56]

Plane fitting provides insight into a proposed system's capacity to operate with surfaces at varying depths of field, but is not without bias. Inherent to the metric, it should be acknowledged that a multitude of methods exist for fitting a plane and none are impervious to subjectivity or is reflective of a ground-truth. In the case of RANSAC approaches,

the produced plane is greatly dependent on the consensus set size, data set (point-cloud) size, and search iterations [57]. If presented independently, any measure of error is strongly reflective of the optics and therefore metric error (mm, cm and etc) can be controlled in experimentation by choosing favorable optics and depths for the planar surface. Furthermore, plane-fitting is subjective of the planar surface, and is therefore ignorant of the sensor's capacity to handle surface discontinuity. This is a known issue with continuous structured light methods, and therefore it is possible for a sensor to achieve low error in plane-fitting and yet be incapable of generating a coherent point cloud for an object with a jagged surface such as a mechanical keyboard [17]. To overcome this limitation, works have elected to report supplementary fittings to non-planar surfaces. One example of such is the reporting of point cloud error and standard deviation for a cylindrical green tea can by Dai and Chung [43].

**Qualitative Metrics:** Assessing and measuring the quality of an imperceptible 3D surface sensor is a difficult task, for as previously discussed, there is little consensus in terms of what qualifies as imperceptible. Some works elect to validate by performing human studies, but this can be a challenging and subject to large amounts of bias. As discussed by Kuroki et al., human perception is impacted by viewing distance and angle, but also content resolution, contrast and color [41]. This is without mentioning physiological differences in visual acuity between different age groups, genders and other genetic dispositions.

As an alternative to visual inspection, system performance can be assessed through quantitative metrics applied to image pairs through measures such as MSE, PSNR, and SSIM. Each of these metrics, defined in Equation set 2.4, quantitatively expresses different measures of change in an image pair. By comparing the unaltered image projection to a captured flicker fusion image over known exposure (16.6ms for 60Hz), such metrics ascertain the integrity of projection.

---

**Equation 2.4** Qualitative Metrics MSE [58], PSNR [59], and SSIM [60][61]

---

Mean Squared Error (MSE):

$$MSE = \frac{1}{M * N * D} \sum_{x=0}^M \sum_{y=0}^N (I_{Sample}(x, y) - I_{Ref}(x, y))^2 \quad (2.3.1)$$

Peak Signal-to-Noise Ratio (PSNR):

$$PSNR = 10 * \log_{10}(\eta^2 / MSE) \quad (2.3.2)$$

Structural Similarity Index (SSIM):

$$SSIM(I_{Sample}, I_{Ref}) = \frac{(2 * \mu_{I_{Sample}} * \mu_{I_{Ref}} + (0.01 * \eta)^2)(\zeta_{I_{Sample}, I_{Ref}} + (0.03 * \eta)^2)}{(\mu_{I_{Sample}}^2 + \mu_{I_{Ref}}^2 + (0.01 * \eta)^2)(\zeta_{I_{Sample}}^2 + \zeta_{I_{Ref}}^2 + (0.03 * \eta)^2)} \quad (2.3.3)$$


---

The MSE, as shown in Equation 2.3.1, provides a simple measure of differences between two images. The measure quantitatively expresses the difference as the sum of squared differences between the pixels of a sample image  $I_{Sample}$  and a reference image  $I_{Ref}$ . The equation normalizes the quantitative measure by dividing the result by the image size  $M * N * D$  (or, in laymen terms, *columns \* rows \* depth*).

PSNR extends on MSE by expressing quantitatively measured noise as ratio in decibels. As shown in Equation 2.3.2, PSNR is a logarithmic representation of MSE which takes into account the peak signal value  $\eta$ . The value of  $\eta$  is respective of the image data type, where for an image composed of unsigned chars,  $\eta$  would equal 255.

SSIM provides a unique measure quantizing changes in image luminescence, contrast, and structure between a sample image  $I_{Sample}$  and a reference image  $I_{Ref}$  [61]. This metric is defined in Equation 2.3.3, where SSIM is calculated using the mean ( $\mu$ ), cross-covariance

( $\zeta$ ) and data type maximums (*eta*) of the images. Though expressed in Equation 2.3.3 as a singular metric, MathWorks notes that SSIM is separable into individual measures of change in luminescence, contrast and structure [60]. This is achieved, respectively, by quantizing the difference in image local means, standard deviations and cross-covariance. Equation 2.3.3 for SSIM is prominent in image processing qualitative metrics reporting, though it should be noted that alternative methodologies of measuring image structural similarity exist (such as *Complex wavelet structural similarity* by Sampat et al. [62]).

**Comparing Methodologies:** Despite these metrics, comparing methodologies is a surprisingly difficult task. To demonstrate, Table 1 illustrates the broad characteristics of recently proposed sensors by Qiu et al. [63] and Cole et al. [45], alongside proposed sensors from the past decade consisting of an M-Array approach by Dai and Chung [43] and a fringe pattern approach by Gong and Zhang [32]. Of these methods, the DLP projected grey-codes by Cole et al. and flicker fusion approach by Dai and Chung are imperceptible techniques.

Table 1: Proposed PROCAM Surface Imaging Sensors

	<i>CPU</i>	<i>Projector</i>	<b>Depth Resolution</b>
Qiu et al. [63]	Intel i9 7920X 2.9GHz	1280x800@30Hz	1280x800 <sup>†</sup>
Cole et al. [45]	AMD Ryzen 5 1600MHz	1920x1080@200Hz (DLP)	1400x256
Dai and Chung [43]	Intel i5-760 2.8 GHz	1024x768@120Hz	68x51
Gong and Zhang [32]	Unknown*	858x600@60Hz (DLP)	480x480

Table 2 was completed by calculating Points per Depth Frame (DFP), Points per Projector Frame (FP) and Point Density in Projector Frame (FPD) using the reported metrics from each paper and three commercially available IR sensors. As denoted by asterisks (\*), operation specifics of the commercial products could not be found and were assumed to be single-shot. This assumption is presumably correct for the Kinect V1 and RealSense which are known to have static IR projection [1] [36, p. 54]. As denoted by a star (\*), the RealSense D435 can operate at 30fps for 1280x720 resolution or at 90fps for 848x480 resolution [36, p. 54]. This alternate mode of operation at 90 fps represents 36.6M PPS and 407,040 points per frame (PPF). It should be noted that after closer analysis it would appear that Qiu et al. may have incorrectly reported the depth resolution of their proposed sensor as 1280x800. The highest resolution reported in experimentation is 5x863, collecting 4,315 points per frame via dual pattern subtraction.

Table 2: Depth Metrics of Proposed PROCAM Surface Imaging Sensors.

	<i>Resolution</i>	<i>FPS</i>	<i>Shots</i>	<i>PPS</i>	<i>DFP</i>	<i>FP</i>	<i>FPD</i>
Kinect V1 [64, 65]	320x240	30	1*	2.3M	76,800	76,800*	1.0*
Kinect V2 [64, 65]	512x424	30	1*	6.5M	217,088	217,088*	1.0*
RealSense D435 [2, 36]	1280 x 720	90*	1*	27.6M	921,600	921,600*	1.0*
Qiu et al. [63]	1280x800 <sup>†</sup>	2,720 <sup>‡</sup>	1	11.7M <sup>‡</sup>	4,315	4,315	0.004
Cole et al. [45]	1400x256	22.2	9	7.96M	358,400	39,822	0.019
Dai and Chung [43]	68x51	60	2	208,080	3,468	1,734	0.0022
Gong and Zhang [32]	480x480	4k	1	921.6M	230,400	230,400	1.0

Points Per Second (PPS), Depth Frame Points (DFP), Frame Points (FP), Frame Point Density (FPD)  
 \*=missing, \*-=conditional, †=wrong, ‡=theoretical

From reported measures of PPS, the proposed sensor by Gong and Zhang outperforms all with the achievement of 921.6M pps. This is due to its 4,000fps camera which allowed for depth sensing on moving fan blades. Isolating operation to a single depth frame, it can be seen that the RealSense D435 has 4x more points per depth frame (DFP). After camera isolation, Qiu et al.'s dependence on camera frame rate is shown, reflecting a 0.004 density of depth points per projected frame (FPD). Dai and Chung's approach produces a lower FP and FPD, but uses flicker fusion for imperceptibility which effectively halves the projection rate from 120Hz to 60Hz. If the same approach was followed using visible artifacts, it is presumable that Dai and Chung's approach would achieve 3,468 FP and a 0.0044 FPD. As



denoted by daggers (‡), it is worth mentioning performance reported by Qiu et al. is theoretical, deriving frame rate (2720 fps) and PPS (11.7M) solely from computational time. This is unprecedented, with no reported metrics or hardware specification for a prototype system. In contrast, Gong and Zhang provide a detailed report of the operation of a prototype sensor and only in conclusion discuss the theoretical limits of operation. As such, Gong and Zhang discuss how modifying active projection of the proposed sensor could increase operating speed up to 12,500Hz by either lowering projector resolution or replacing it with a mechanical grating. On another note, the hardware used for analysis of computational time by Qiu et al. is near state-of-the-art consumer hardware, with the reported i9 processor offering a 263% higher 8-core benchmark score than the i5 reported by Dai and Chung [66]. As such, the claim that increased bandwidth is achievable by further hardware optimization is quite arguable, particularly in an academic setting. As suggested by the metrics reported in Table 2, the operating bottleneck appears to lie in the sparse projected pattern, which could be replaced by a continuous fringe pattern to dramatically increase point density (FPD). It should be further mentioned that Qiu et al.’s account of the metrics and operation of the RealSense D435 is inaccurate.

The metrics in Table 2 emphasize the desirability for higher data throughput at the lowest possible operating frequency. As a general observation, these characteristics enable the lower hardware costs of commercially available products, disfavoring multi-shot approaches. This is exemplified by the 9-pattern approach used by Cole et al. which exhibits a low FPD and a lower FP than the Kinect V1. These metrics belie the added capacity offered by the proposed system, which takes advantage of DLP technology. Cole et al. were able to project three patterns per aggregated frame, achieving an effective pattern projection rate of 200Hz to collect depth at 22.2Hz while simultaneously maintaining coherent media projection at 66Hz. This gives the proposed sensors capacities beyond the Kinects and RealSense, though there is a significant lack of evidence to support this claim. The work by Cole et al. claim an ‘empirical’ measurement of imperceptibility from showing *one singular* static image to ten people [45]. Though derived mathematically, no qualitative or quantitative measurements are reported to support successful implementation. In contrast, Dai and Chung validated through experimentation using 500 random images with varied encoding intensities being ranked quantitatively by ten subjects of known demographics seated 1m away [43].

Understanding performance and qualitative metrics commonly reported for 3D surface imaging sensors is important to correctly interpreting reported results. As discussed in this section, many of the commonly reported metrics have inherent bias towards hardware and circumstance. Unfortunately, due to a lack of conformity, it is often up to the reader to interpret the meaning of reported results. It should be noted that all the metrics are generally performed in ideal environments, projecting on lambertian planar surfaces. None take into account environmental factors such as external sources of light, surface characteristics such as albedo, or hardware cost. As such, using the metrics above, comparing proposed PROCAM 3D surface imaging sensors or assessing their aptitude for a specific task is often futile.

### 2.4.1 Discussion

There are advantages to each 3D surface imaging approach, but as discussed in the previous section, differentiating between methods to select the best approach for a specific operating environment can be a complex task. Passive Correspondence Based Vision (PCBV) offers a versatile and inexpensive approach, but depth is dependent on disparity in correspondences which may or may not exist in scene morphology. Active Correspondence Based Vision (ACBV) methods work on the same triangulation principle but offer some invariance to surface morphology by actively projecting correspondence through visible light or an alternative spectrum (e.g IR). Such sensors are currently predominant in the consumer market, with imperceptible variations offering increased hardware versatility to accommodate multi-media projection. Such methods are diverse in approach, but it should be noted that multi-shot (coded or continuous) approaches are subject to motion blur in depthmaps and lower frame rates. Alternatives to triangulation-based 3D vision based on physical properties of light (such as ToF) mitigate the inherent parallax of multi-sensor systems and

offer the greatest invariance to scene morphology. As the cost of such sensors continues to decrease, it is quite likely that such sensors will continue to gain popularity and encroach on ACBV market share.

### 3 Machine & Deep Learning

In recent years, machine learning (ML) and deep learning (DL) have become incumbent tools rapidly being adopted into every field of science and engineering imaginable. This recent interest and adoption is paradoxically, given deep learning is considered a new and emerging field yet has roots almost a century old. As discussed by Ian Goodfellow, Deep Learning has had many names, starting with cybernetics in the 1940-60s, connectionism in the 1980-90s and finally deep learning in 2006 [67]. The adoption of ML and DL today is a reflection the present day, where for the first time history the necessary computational resources and datasets are available to sustain training and viable operation of models for a wide range of applications [67, 68]. As such, this is a pivotal moment in time where the paradigm of machine vision has changed and, arguably, been turned upside down by ML and DL.

Though today's largest neural networks have fewer neurons than a frog, the achievements that have been possible through ML and Deep Neural Networks (DNNs) is considerable [67]. As observed by Ling in 2017, DL solutions have become predominant in analysis for semantic understanding, outperforming traditional machine vision solutions for object detection, semantic segmentation, face recognition, and visual tracking [69]. These changes have been incremental, with catalyst works such as 'LeNet' in 1998 by LeCun et al. [70] and 'Alex Net' by Krizhevsky et al. in 2012 [71].

Throughout this literature review, only a handful of ML references were encountered, using ML for image- and post-processing. This includes a 2004 paper in stereo-vision by Sinz et al., which demonstrated that Gaussian processes could be used to learn mappings from image to spatial coordinates [72]. Sinz et al. claim their proposed approach could lead to higher depth accuracies and faster operation than classical calibration methods. In terms of ACBV, Dai and Chung implement ML for object recognition, detecting and recognizing primitive shapes used to create codewords and correspondences [43, 73]. Their work would progress to be demonstrated as a successful touchscreen interface [74, 75]. More recently in 2016, Fanello et al. proposed treating the depth triangulation correspondence problem as a classification-regression task which could be solved by cascading random forests [56]. Improving on the computationally expensive and localized matching approaches, Fanello et al. demonstrate a 375Hz depth camera based on IR dot patterns with lower error and noise characteristics than commercially available sensors.

Aside from these works, many commercial depth-sensing products targeting industry mention the keywords ML and DL throughout their product listings. Of those perused, all implemented ML and DL in post-processing or image analysis, suggesting no application of ML or DL towards 3D surface imaging techniques. This observation is mirrored in consumer products, where few references were found. Of tangible significance, the iPhone X was noted to use ACBV to project 30,000 invisible dots to create depth maps and infrared images of a user's face [11]. Using this information, the iPhone X's neural engine is capable of transforming the depth maps and infrared images to compare representations against enrolled facial data. No further specification on the system could be found and it is presumable that ML and DL are solely applied in post-processing through the neural engine.

To conclude, at this time there are no known works with pertinence to the designing, encoding and/or decoding of patterns for ACBV sensors using ML or DL. As will be discussed in the following section, there is potential for significant innovation in the field of ACBV. It is hoped that ML and DL may assist in making these changes a reality.

### 4 Machine & Deep Learning Potential

Though few works could be found with relation to correspondence based vision, it is important to recognize the potential and versatility offered by ML and DL models. Though

often designed and demonstrated around a singular problem, models can offer predictive capacities far beyond the original scope. As an example, Convolutional Neural Networks (CNNs) gained popularity after LeCun et al. demonstrated their capacity to recognize 28x28 images of handwritten digits from the MNIST database [70], but today CNNs are used for a plethora of different applications, including forecasting financial markets. This concept was successfully demonstrated by Sezer and Ozbayoglu [76] in 2018 and extended by Maratkhan et al. in 2019 to increase annualized returns from 13% to 29.54% on DOW-30 stocks [77]. To continue this example, alternative approaches to financial analysis using ML and DL have been tried with almost every model imaginable, including random forests [78], Reinforced Learning (RL) [79], and Long Short-Term Memory (LSTM) [80].

As demonstrated through DNNs designed for image classification, DNNs have the capacity to generalize beyond human reasoning. This concept is well illustrated by attempts to visualize incremental steps in DNN models for image classification, where the first layers appear humanoid in approach, but lose coherence and rationality at deeper layers. Though the logic in operation is indiscernible, it is known that deeper DNN models have increased capacity for generalization, handling input with varied resolutions, object sizes, illumination, and coloring. Models created by He et al. and Howard et al. demonstrate these capacities, with Howard et al. adding special hyper-parameters to control the trade-off between latency and accuracy [81][82]. That being said, the larger the network, the greater the computational expense, and significant research has been made in making networks more efficient. This includes works by Iandola et al. to create SqueezeNet, a replica of AlexNet with 50x fewer parameters and sub 0.5MB model size [83], and GoogleNet by Szegedy et al., which targeted increased model depth and width without incurring additional computational expense [84]. Similar approaches have been taken in object detection, where unified detection and recognition models (without transfer of learning) can be optimized end-to-end to increase model performance. This was demonstrated in works such as YOLO by Redmon et al. [85].

To this avail, ML and DL offer an interesting potential towards a unified ACBV solution. ACBV and modulated approaches discussed in Sections 2.2 and 2.3 are designed pragmatically through human understanding and intuition of projection, scene illumination, and capture. Furthermore, all discussed sensors exhibit scene invariant operation, offering no dynamic sensing or control over illumination, object edges, or imperceptibility, leaving a significant gap between the current state-of-the-art and theoretical optimal. It is therefore possible that ML and DL may offer the toolset necessary to bridge this theoretical gap, creating a unified solution with capacities beyond procedural and rule-based human conceived models. Through a hybrid Generative Adversarial Network (GAN) and Autoencoder model, a new approach to ACBV with dynamic projection and maximized throughput in various operating conditions may be possible, offering the next echelon towards the theoretical optimal.

## 5 Conclusion

In this literature review, an amalgamation of research and commercial products surrounding 3D surface imaging has been presented and critiqued. Through these works, many industrial and consumer systems gain insight into the physical world, obtaining the capacity for qualitative measurement and intelligent interaction for applications from self-driving cars to biometric security on personal devices. Though these solution may not be accessible or practical for all applications, the summarized works demonstrate an evolving field which is striving to provide visual acclimation for a world growing in its requirement.

As the paradigm of machine vision changes, this work suggests an evolution in 3D surface imaging towards unified machine and deep learning solutions. Offering the potential for a new generation of reliable and versatile designs, it is anticipated that future works will explore adaptive, abstract and alternative designs operational in diverse theaters of operation. These changes will innovate on the current static encoding and decoding pipelines of structured light sensing, offering dynamic operation to meet the growing requirements for precision and quality.

To conclude, this work summarized research with pertinence to 3D surface sensing using structured light (SL). Literature included forerunner works and research in structured light sensing, and introduced the potential offered by machine learning and deep neural network (DNN) models towards integrated and unified DNN & SL solutions.

## References

- [1] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb 2012.
- [2] Intel Corporation. Intel® realsense™ technology. Accessed 17/1/2020. [Online]. Available: <https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html>
- [3] Zivid. Zivid. Accessed 17/1/2020. [Online]. Available: <http://www.zivid.com/>
- [4] Photoneo s. r. o. Phoxi 3d scanner. Accessed 19/1/2020. [Online]. Available: <https://www.photoneo.com/>
- [5] Sony Depthsensing Solutions. Sony depthsensing solutions. Accessed 19/1/2020. [Online]. Available: <https://www.sony-depthsensing.com/>
- [6] AEye, Inc. Think like a robot, perceive like a human. Accessed 19/1/2020. [Online]. Available: <https://www.aeye.ai/>
- [7] Basler AG. Basler 3d cameras. Accessed 19/1/2020. [Online]. Available: <https://www.baslerweb.com/en/products/cameras/3d-cameras/>
- [8] LMI Technologies. 3d smart sensors. Accessed 19/1/2020. [Online]. Available: <https://lmi3d.com/>
- [9] Keyence Corporation. Products. Accessed 19/1/2020. [Online]. Available: <https://www.keyence.com/>
- [10] J. Carroll. Camera intrinsics. Accessed 19/1/2020. [Online]. Available: <https://www.vision-systems.com/factory/article/16748308/vision-systems-design-announces-2019-innovators-awards>
- [11] Apple Inc. About face id advanced technology. Accessed 19/1/2020. [Online]. Available: <https://support.apple.com/en-ca/HT208108>
- [12] M. Levoy and Y. Pritch, "Portrait mode on the pixel 2 and pixel 2 xl smartphones," *Google AI Blog*, October 2017, accessed 19/1/2020. [Online]. Available: <https://ai.googleblog.com/2017/10/portrait-mode-on-pixel-2-and-pixel-2-xl.html>
- [13] "Apple explains face id on-stage failure," *BBC News*, September 2017, accessed 16/1/2020. [Online]. Available: <https://www.bbc.com/news/technology-41266216>
- [14] J. Goodrich, "Facial recognition faces more proposed bans across u.s." *IEEE Spectrum*, May 2019, accessed 16/1/2020. [Online]. Available: <https://spectrum.ieee.org/news-from-around-ieee/the-institute/ieee-member-news/why-san-francisco-banned-the-use-of-facial-recognition-technology>
- [15] Z. Istvan, "Facing up to facial recognition," *IEEE Spectrum*, Jan 2020, accessed 28/1/2020. [Online]. Available: <https://spectrum.ieee.org/computing/software/facing-up-to-facial-recognition>
- [16] Silicon Software GmbH. Touch the future. Accessed 19/1/2020. [Online]. Available: <https://silicon.software/en>
- [17] J. Geng, "Structured-light 3d surface imaging: a tutorial," *Adv. Opt. Photon.*, vol. 3, no. 2, pp. 128–160, Jun 2011. [Online]. Available: <http://aop.osa.org/abstract.cfm?URI=aop-3-2-128>
- [18] M. Matheis, "Evaluation of depth-camera-systems for usage in semi-controlled assembly environments," Master's thesis, University of Stuttgart, 2016.
- [19] Zivid. Zivid one+ technical specification. Accessed 19/1/2020. [Online]. Available: <https://www.zivid.com/hubfs/files/SPEC/Zivid%20One%20Plus%20Datasheet.pdf>
- [20] I. J. Maquignaz, "Imperceptible pattern embedding: Structured light steganography for augmented reality applications," Master's thesis, Queen's University, 2019.
- [21] T. Luhmann, S. Robson, S. Kyle, and I. Harley, *Close Range Photogrammetry: Principles, Techniques and Applications*. Wiley, 2007. [Online]. Available: <https://books.google.ca/books?id=oclpAAAACAAJ>



- [22] O. R. Kölbl, “Metric or non-metric cameras,” *Photogrammetric Engineering and Remote Sensing*, vol. 42, no. 1, pp. 103–113, 1976.
- [23] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.
- [24] OpenMVG. Pinhole camera model. Accessed 08/2/2020. [Online]. Available: <https://openmvg.readthedocs.io/en/latest/openMVG/cameras/cameras/>
- [25] W. Y. Yau and H. Wang, “Fast relative depth computation for an active stereo vision system,” *Real-time imaging*, vol. 5, no. 3, pp. 189–202, 1999.
- [26] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, “A state of the art in structured light patterns for surface profilometry,” *Pattern Recognition*, vol. 43, no. 8, pp. 2666 – 2680, 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S003132031000124X>
- [27] C. Loop and Z. Zhang, “Computing rectifying homographies for stereo vision,” Tech. Rep. MSR-TR-99-21, April 1999. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/computing-rectifying-homographies-for-stereo-vision/>
- [28] B. Porr, B. Nürenberg, and F. Wörgötter, “A vlsi-compatible computer vision algorithm for stereoscopic depth analysis in real-time,” *International Journal of Computer Vision*, vol. 49, no. 1, pp. 39–55, 2002.
- [29] J. Mallon and P. F. Whelan, “Projective rectification from the fundamental matrix,” *Image and Vision Computing*, vol. 23, no. 7, pp. 643 – 650, 2005. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0262885605000399>
- [30] Sony Interactive Entertainment LLC. Playstation camera. Accessed 10/2/2020. [Online]. Available: <https://www.playstation.com/en-ca/explore/accessories/playstation-camera-ps4/>
- [31] R. Orghidan, J. Salvi, and E. M. Mouaddib, “Modelling and accuracy estimation of a new omnidirectional depth computation sensor,” *Pattern Recognition Letters*, vol. 27, no. 7, pp. 843–853, 2006.
- [32] Y. Gong and S. Zhang, “Ultrafast 3-d shape measurement with an off-the-shelf dlp projector,” *Opt. Express*, vol. 18, no. 19, pp. 19743–19754, Sep 2010. [Online]. Available: <http://www.opticsexpress.org/abstract.cfm?URI=oe-18-19-19743>
- [33] D. Li, D. Wang, and D. Weng, “Non-planar projection performance evaluation and projector pose optimization,” *Journal of the Society for Information Display*, vol. 26, no. 6, pp. 352–368, 2018.
- [34] Y. Deshayes and L. Béchou, *Reliability, Robustness and Failure Mechanisms of LED Devices: Methodology and Evaluation*. Elsevier, 2017, pp. 40–42.
- [35] C. S. Wiegardt and B. Wagner, “Hand-projector self-calibration using structured light,” in *2014 11th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, vol. 01, Sep. 2014, pp. 85–91.
- [36] Intel Corporation. Intel®realsensetmd400series product family. Accessed 14/2/2020. [Online]. Available: <https://www.intel.com/content/dam/support/us/en/documents/emerging-technologies/intel-realsense-technology/Intel-RealSense-D400-Series-Datasheet.pdf>
- [37] D. M. P. Nagaraja, “Solar spectrum, variability, and atmospheric absorption,” <https://science.nasa.gov/science-news/science-at-nasa/images/sunbathing/sunspectrum>, 1 2010, (Accessed on 02/18/2018).
- [38] T. Rittler, F. Seitner, and M. Gelautz, “Structured-light-based depth reconstruction using low-light pico projector,” in *Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia*, ser. MoMM 2015. New York, NY, USA: ACM, 2015, pp. 334–337. [Online]. Available: <http://doi.acm.org.proxy.queensu.ca/10.1145/2837126.2837154>
- [39] M. J. Landau and P. A. Beling, “Optimal model-based 6-d object pose estimation with structured-light depth sensors,” *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 58–73, 2017.

- [40] R. Raskar, J. Van Baar, P. Beardsley, T. Willwacher, S. Rao, and C. Forlines, “ilamps: geometrically aware and self-configuring projectors,” *ACM Transactions on Graphics (TOG)*, vol. 22, no. 3, pp. 809–818, 2003.
- [41] Y. Kuroki, T. Nishi, S. Kobayashi, H. Oyaizu, and S. Yoshimura, “A psychophysical study of improvements in motion-image quality by using high frame rates,” *Journal of the Society for Information Display*, vol. 15, no. 1, pp. 61–68, 2007.
- [42] J. C. Lee, S. E. Hudson, J. W. Summet, and P. H. Dietz, “Moveable interactive projected displays using projector based tracking,” in *Proceedings of the 18th annual ACM symposium on User interface software and technology*, 2005, pp. 63–72.
- [43] J. Dai and C. R. Chung, “Embedding invisible codes into normal video projection: Principle, evaluation, and applications,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 12, pp. 2054–2066, Dec 2013.
- [44] D. Cotting, M. Naef, M. Gross, and H. Fuchs, “Embedding imperceptible patterns into projected images for simultaneous acquisition and display,” in *Third IEEE and ACM International Symposium on Mixed and Augmented Reality*, Nov 2004, pp. 100–109.
- [45] A. Cole, S. Ziauddin, and M. Greenspan, “High-speed imperceptible structured light depth mapping,” *Accepted in International Conference on Computer Vision Theory and Applications*, 2020.
- [46] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stessin, and H. Fuchs, “The office of the future: A unified approach to image-based modeling and spatially immersive displays,” in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. ACM, 1998, pp. 179–188.
- [47] H. Park, M.-H. Lee, B.-K. Seo, Y. Jin, and J.-I. Park, “Content adaptive embedding of complementary patterns for nonintrusive direct-projected augmented reality,” in *Virtual Reality*, R. Shumaker, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 132–141.
- [48] A. B. Watson, “Temporal sensitivity. handbook of perception and human performance,” *Sensory Processes and Perception*, edS Boff KR, Kaufman L., Thomas JP, editors.(New York: Wiley, pp. 1–43, 1986.
- [49] M. J. Amin and N. A. Riza, “Active depth from defocus system using coherent illumination and a no moving parts camera,” *Optics Communications*, vol. 359, pp. 135–145, 2016.
- [50] D. Iwai, S. Mihara, and K. Sato, “Extended depth-of-field projector by fast focal sweep projection,” *IEEE transactions on visualization and computer graphics*, vol. 21, no. 4, pp. 462–470, 2015.
- [51] E. H. Adelson and J. Y. A. Wang, “Single lens stereo with a plenoptic camera,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 99–106, Feb 1992.
- [52] A. Lumsdaine and T. Georgiev, “The focused plenoptic camera,” in *2009 IEEE International Conference on Computational Photography (ICCP)*, April 2009, pp. 1–8.
- [53] L. Li, “Time-of-flight camera—an introduction,” *Technical white paper*, no. SLOA190B, 2014.
- [54] L. Frenzel, “What’s the difference between bit rate and baud rate?” <https://www.electronicdesign.com/technologies/communications/article/21802272/whats-the-difference-between-bit-rate-and-baud-rate>, April 2012, (Accessed on 02/12/2020).
- [55] K. Kikuchi, “Fundamentals of coherent optical fiber communications,” *Journal of Lightwave Technology*, vol. 34, no. 1, pp. 157–179, 2016.
- [56] S. Ryan Fanello, C. Rhemann, V. Tankovich, A. Kowdle, S. Orts Escolano, D. Kim, and S. Izadi, “Hyperdepth: Learning depth from structured light without matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5441–5450.

- [57] M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, no. 6, p. 381–395, Jun. 1981. [Online]. Available: <https://doi-org.proxy.queensu.ca/10.1145/358669.358692>
- [58] MathWorks, “Mean-squared error,” <https://www.mathworks.com/help/images/ref/immse.html>, 2018, (Accessed on 02/16/2018).
- [59] —, “Peak signal-to-noise ratio (psnr),” <https://www.mathworks.com/help/images/ref/psnr.html>, 2018, (Accessed on 02/16/2018).
- [60] —, “Structural similarity index (ssim) for measuring image quality,” <https://www.mathworks.com/help/images/ref/ssim.html>, 2018, (Accessed on 02/16/2018).
- [61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [62] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, “Complex wavelet structural similarity: A new image similarity index,” *IEEE transactions on image processing*, vol. 18, no. 11, pp. 2385–2401, 2009.
- [63] Y. Qiu, J. Malcolm, A. Vattoo, S. Ziauddin, and M. Greenspan, “Inverse rectification for efficient procam pattern correspondence,” *Accepted in Winter Conference on Applications of Computer Vision*, 2020.
- [64] J. Park, H. Chao, H. Arabnia, and N. Y. Yen, “Advanced multimedia and ubiquitous engineering,” *Future Information Technology*, vol. 2, 2015.
- [65] M. Rahman, *Beginning Microsoft Kinect for Windows SDK 2.0: Motion and Depth Sensing for Natural User Interfaces*. Apress, 2017.
- [66] UserBenchmark, “Intel core i9-7920x vs intel core i5-760,” <https://cpu.userbenchmark.com/Compare/Intel-Core-i9-7920X-vs-Intel-Core-i5-760/m278103vsm717>, (Accessed on 02/15/2020).
- [67] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [68] TOP500.org, “Performance development,” <https://www.top500.org/statistics/perfdevel/>, 2019, (Accessed on 02/19/2020).
- [69] H. Ling, “Augmented reality in reality,” *IEEE MultiMedia*, vol. 24, no. 3, pp. 10–15, 2017.
- [70] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
- [72] F. H. Sinz, J. Q. Candela, G. H. Bakır, C. E. Rasmussen, and M. O. Franz, “Learning depth from stereo,” in *Pattern Recognition*, C. E. Rasmussen, H. H. Bülthoff, B. Schölkopf, and M. A. Giese, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 245–252.
- [73] J. Dai and R. Chung, “Embedding imperceptible codes into video projection and applications in robotics,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 4399–4404.
- [74] M. He and J. Cheng, “Self-adaptive coding-based touch detection for interactive projector system,” in *2014 4th IEEE International Conference on Information Science and Technology*, April 2014, pp. 656–659.
- [75] J. Dai and C. R. Chung, “Touchscreen : On transferring a normal planar surface to a touch-sensitive display,” *IEEE Transactions on Cybernetics*, vol. 44, no. 8, pp. 1383–1396, Aug 2014.

- [76] O. B. Sezer and A. M. Ozbayoglu, “Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach,” *Applied Soft Computing*, vol. 70, pp. 525 – 538, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1568494618302151>
- [77] A. Maratkhan, I. Ilyassov, M. Aitzhanov, M. F. Demirci, and M. Ozbayoglu, “Financial forecasting using deep learning with an optimized trading strategy,” in *2019 IEEE Congress on Evolutionary Computation (CEC)*, June 2019, pp. 838–844.
- [78] J. Zhang, S. Cui, Y. Xu, Q. Li, and T. Li, “A novel data-driven stock price trend prediction system,” *Expert Systems with Applications*, vol. 97, pp. 60 – 69, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0957417417308485>
- [79] P. C. Pendharkar and P. Cusatis, “Trading financial indices with reinforcement learning agents,” *Expert Systems with Applications*, vol. 103, pp. 1 – 13, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0957417418301209>
- [80] Y. Baek and H. Y. Kim, “Modaugnet: A new forecasting framework for stock market index value with an overfitting prevention lstm module and a prediction lstm module,” *Expert Systems with Applications*, vol. 113, pp. 457 – 480, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0957417418304342>
- [81] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [82] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
- [83] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size,” *arXiv preprint arXiv:1602.07360*, 2016.
- [84] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [85] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [86] Solomon Tehchnology Corporation. Think like a robot, perceive like a human. Accessed 19/1/2020. [Online]. Available: <https://www.solomon-3d.com/>
- [87] LUCID Vision Labs Inc. Discover the industrial helios time of flight camera. Accessed 19/1/2020. [Online]. Available: <https://thinklucid.com/helios-time-of-flight-tof-camera/>
- [88] MathWorks, Inc. What is camera calibration? Accessed 07/2/2020. [Online]. Available: <https://www.mathworks.com/help/vision/ug/camera-calibration.html>
- [89] Depthkit, “Kinect for windows v2,” <https://docs.depthkit.tv/docs/kinect-for-windows-v2>, (Accessed on 02/12/2020).
- [90] J. Vincent, “Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech,” *The Verge*, January 2018, accessed 16/1/2020. [Online]. Available: <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>
- [91] T. Jia, B. Wang, Z. Zhou, and H. Meng, “Scene depth perception based on omnidirectional structured light,” *IEEE Transactions on image processing*, vol. 25, no. 9, pp. 4369–4378, 2016.