

Whitepaper

3D Vision Enables Everyday Devices to “See”

Why 3D vision is necessary for mass-market electronic vision applications

Salih Burak Gokturk

Abbas Rafii

Carlo Tomasi

This paper identifies and describes why 3D vision is necessary to enable a machine or digital device to “see.” The capabilities of existing 2D imaging solutions are explored and found to have significant limitations for practical mass-market applications. The benefits of 3D vision, particularly single-chip 3D vision, are then explained, along with capabilities that would otherwise be impossible or expensive to deliver with traditional 2D methods. The paper concludes that embeddable 3D vision solutions, like Canesta’s, are essential to enable everyday devices the ability to truly understand and interact with the world around them.

Table of Contents

Introduction 3

Why 3D Imaging Is Important..... 3

 Segmentation & separation of foreground from background 3

 Invariance to lighting and other environmental variances 4

 Sizing objects 4

 Object identification and tracking 5

Limitations of Common (2D) Digital Camera Approaches 5

 Ambient light disrupts the intensity image 6

 Same object can have many different colors and textures 7

 But color is useful to distinguish objects from each other: 8

 Impossible to determine size..... 9

 Impossible to determine the direction of motion..... 9

Conclusion 10

Introduction

With all their sophistication, today's computers still understand their environment poorly, and interact with it in relatively clumsy ways. They store and display images and video efficiently, but are utterly blind to the meaning of the billions of pixels neatly arrayed in their memories.

To address these limitations, Canesta has developed a new low cost electronic perception technology that enables everyday devices to "see" and understand their environment as objects. This small, single-chip solution forms three-dimensional images of its nearby surroundings in real time, providing the raw information necessary to enable a device to reliably perceive the world around it so that it may react and interact with it. There are many applications for this technology including automobile safety systems that can detect the size and position of a passenger so the airbag deploys appropriately in a crash; videogames and remote controls that can interact with the user through gestures or body movements; facial recognition systems that use three-dimensional shape to identify their subject more accurately, and many more.

The information contained in this document is intended to describe why 3D imaging, like Canesta's solution, is important and ultimately necessary to enable mass-market vision applications.

Why 3D Imaging Is Important

Traditional imaging systems produce two dimensional (2D) intensity images, which are appropriate for electronic recording and display of images. Such images are intended to be viewed and interpreted by human eyes. However, when the same images are processed by a computer to recognize, interpret and track objects, the task becomes immensely difficult because the contextual information is no longer available. ***The key to "see" objects is not to analyze the scene merely in terms of colors and textures, rather to segment it in terms of real-world objects and their spatial and temporal relationships in the scene.*** Imaging in 3D dramatically simplifies key techniques necessary for computer vision which ultimately makes mass market deployment possible. These computer vision essentials are described below.

Segmentation & separation of foreground from background

In a 3D image, the objects are easily mapped to different planes giving a Z-order (order in distance) for each object. The Z-order enables the computer program to distinguish the foreground objects from the background. Using this knowledge, the application can focus on foreground regions eliminating the wasteful processing of background material.

Segmentation is a major preprocessing task for various computer vision applications. For instance, the face should be segmented from the rest of the image before applying any face recognition application. Segmentation is a challenging task with intensity images (Figure 2).



Figure 1. An intensity image and grayscale-coded depth image of a person. Since the background is similar color to the person's shirt, it is difficult for the computer to segment the person from the background using the intensity image. The segmentation is trivial using the depth image.

The research on segmentation has been going on since the evolution of computer vision. Currently, color (which is known to be deceptive as described below) and motion clues are used for this task. Unfortunately current tracking systems fail over long video sequences of objects. Contour finding and segmentation becomes an easy task with range images. For instance, objects can be grouped by their distances to the camera. This way, the very challenging segmentation task is dramatically simplified.

Invariance to lighting and other environmental variances

Mass-market applications require operation that is robust over the variety of conditions that practical user scenarios present. Color and texture of clothing and background conspire to fool color camera algorithms. For instance, living room applications may range from brightly lit to dim and include subjects with highly textured clothing to completely plain or texture-less (e.g. all white) clothing. The color of skin dramatically changes under different lighting types (e.g. fluorescent vs. incandescent). Robust applications require solutions that are invariant to these environmental and scene dependencies. Consumers are frustrated by applications that do not work all the time or impose restrictions on ambient lighting. There should be no artificial discontinuity due to the method of data collections. The discontinuity in data degrades the performance of algorithms and causes erroneous interpretation of data, making solutions that work in some controlled environments completely impractical in others. This variance is unacceptable for mass-market, consumer application where the environment and scene cannot be controlled.

Sizing objects

The 3D information in an image enables the computer program to determine the size of the objects. In a 2D image, a foreground child may have the same outline as a background adult. However, using 3D information, the computer program can discern the actual size of each person regardless of his or her relative positions. For instance, in an application to determine the type and size of a car occupant for controlling the deployment force of an airbag, the 3D information can be used to determine the size (e.g. child vs. adult) of the occupant. Using the depth information, the application can estimate the true size of a shape that is captured by a block of

pixels and make application reasoning based on the size of the object (e.g. automatically activate parental control for children's TV viewing options.)

Object identification and tracking

Compared to humans, computers lack the visual processing abilities of the visual cortex of the brain. Computers are also at a disadvantage because of the limited form of the visual data that are typically available to them. With 3D imaging, however, computers are able to make the necessary computations to identify and track objects in real-time regardless of their orientation. Using mathematical expressions, real-time 3D images can provide the orientation of an object and the direction of its movement. Furthermore, the 3D data can be analyzed to determine the position of the camera. These properties become important for identifying and tracking objects in real-time, keeping objects separated from each other when many objects are being tracked and stitching together successive images.

In an augmented reality game, the real world 3D space in front of the player can be mapped to a virtual 3D world created by the application. The superimposition of these two real and imaginary 3D spaces can create fantastic opportunities for the player to interact with computer animated objects displayed on a large screen.

Intensity (2D) based trackers can be deceived when an object does not contain enough texture compared to its background, or when it makes an abrupt position or orientation change. Knowing the geometrical model of an object simplifies this task. The task of tracking becomes much simpler with 3D information since the 3D geometry of the object is explicitly given in the data. A tracker using 3D information is invariant to color, texture and pose variations of the objects.

Limitations of Common (2D) Digital Camera Approaches

Since the invention of the first television camera, machines have been able to form electronic images of the world around them so they may be transmitted or stored so that *humans* can see them. This is certainly a valuable capability, supported by the size of the still picture and video recording industries and underscored by the popularity of digital cameras from webcams to cell phones.

What machines have not been able to do, however, is cost-effectively resolve such images into sets of objects in three dimensions; a process that the eye and brain do seemingly without effort. Enabling a machine to “see” requires the ability to collect and process *depth* information as described above and is the core reason 2-dimensional imaging technologies do not suffice. This section explains these limitations in greater detail.

Let us first investigate the meaning of an intensity image versus range image. Both of these images are constructed as a 2-D array of pixels. In an intensity image, each pixel denotes the amount of light received from the scene. The value of each pixel is a quantized value of this amount. In a range image, each pixel contains the range of an object from the camera. Using this image and simple lens equations, one can simply construct the 3-D information of the scene.

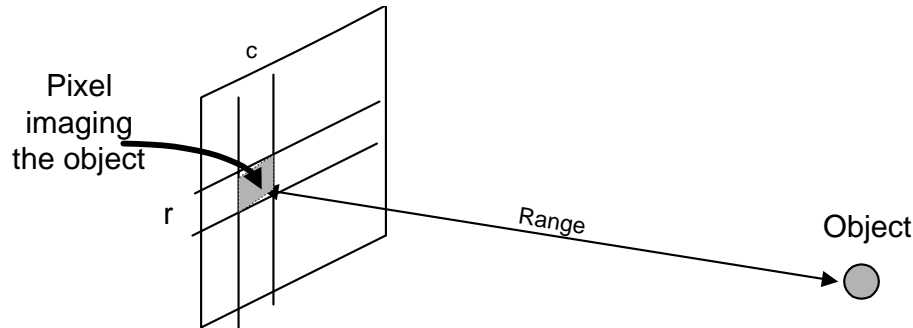


Figure 2. Range information is kept in every pixel of the Canesta sensor.

Analyzing intensity images for an understanding of the scene is a difficult task. Color or brightness information is the only information that a raw intensity image provides, and this information does not necessarily correspond well to different objects in the scene. As a consequence, the computer system needs to be trained to distinguish objects based on their colors or movements. Having the 3-D information greatly simplifies this task. For example, a foreground object can be simply separated from the background objects using its range information.

Ambient light disrupts the intensity image

The intensity image varies with environmental lighting conditions whereas the range image is not affected by the illumination changes. This is another main advantage of using range images. This is similar to the problem of face recognition applications using 2D images. The systems tend to recognize the same person in some illumination condition, yet not in some other illumination condition (Figure 3).



Figure 3. The image of the same person under various illumination conditions. An intensity based face recognition system works on the first image but not on the second image. Canesta sensor deals with this problem by providing illumination invariant output.

Similarly, many applications require operation under various lighting conditions. Such a system should be trained for every possible lighting condition (Figure 4), another burden that affects the practical use of vision systems. This is a significant problem for practical applications (such as living rooms) where lighting conditions vary widely and it is impractical to train the system for all possible situations. Variances in lighting are an “Achilles heel” for virtually all vision systems which depend only on 2D color images.

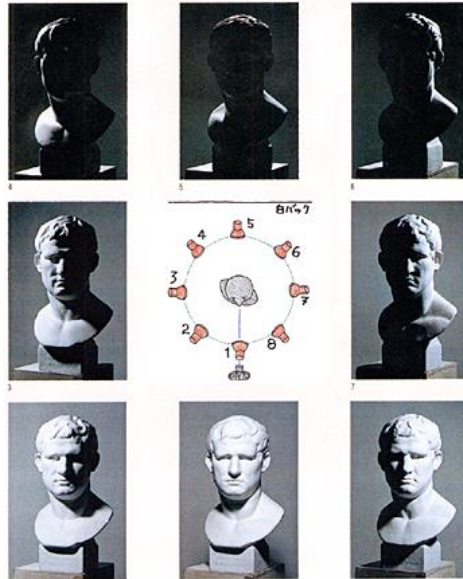


Figure 4. Image of a statue under varying illumination conditions. The intensity image varies under different illumination. This makes the image more difficult to process by a computer.

Same object can have many different colors and textures

In order to recognize an object, the system needs to distinguish between various colors, textures and poses of that object. For instance, a face detection system needs to be trained by all possible skin colors, and poses of the face (Figure 5). As mentioned before, the skin color also changes significantly under different lighting type compounding the problem. This constructs a tremendous amount of training data. Storage and processing of this data becomes a challenging dilemma. A 3-D image simplifies this task since the shape and size information about objects is already available in the data.



Figure 5. A face detection system based on intensity images should be trained for every possible skin color, texture, etc. A 3D based system can be trained on even a simple 3D face model and can work reliably.

But color is useful to distinguish objects from each other:

Intensity images can be deceptive for a computer. Very commonly, the quantized values of different objects project onto same or similar intensity values on an image. Especially, when we use a mono-color black and white camera, two completely different colors (red and blue for instance) can map to the same brightness level (Figure 6). On a color camera, however, the amount of data that needs to be processed increases by three fold. Therefore, color information can be a valuable addition to a system which employs 3D imaging, but is seldom capable of providing enough information to enable a robust end-use application. For example, colors change dramatically depending on the lighting (e.g. fluorescent or incandescent). This makes color a helpful variable to consider but impractical to depend on for core recognition capabilities.

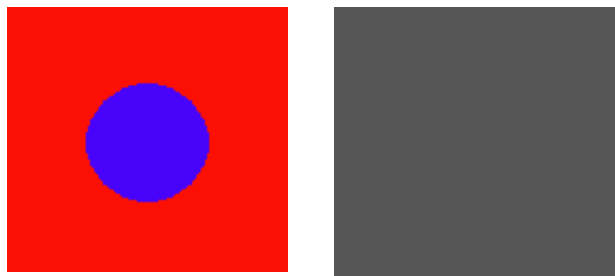


Figure 6. A color image and its grey scale correspondence. The same colored objects might map to same intensity level when we use intensity images.

Difficult or impossible to recognize objects

Suppose a computer program has identified the outline of an object taken by a regular 2D camera. Any of the objects in Figure 7 shown from the side view can map to the same image.

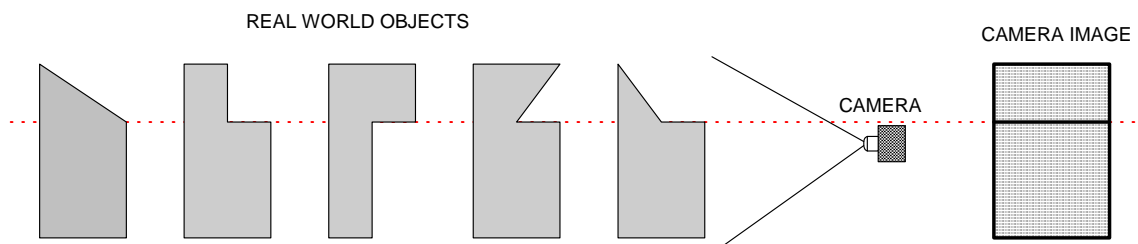


Figure 7. The image of different shapes can be the same.

Consider an application which is trying to recognize the difference between a closed fist gesture and a gesture where there is a closed fist with the forefinger pointing at the camera. It is virtually impossible to recognize this difference with a 2D image; yet trivial with a 3D image.

Impossible to determine size

It is difficult to determine the size of the objects by looking at the 2D images. A small object that is close to the camera, or a big object that is far away from the camera can project onto same sizes on the image plane (Figure 8). Therefore, it is not possible to know the exact size of objects without knowing their distance from the camera.

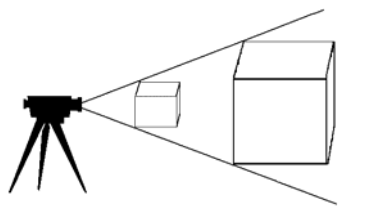


Figure 8. The image of a small and close object can be the same as the image of a bigger object that is further away.

Impossible to determine the direction of motion

The projection of 3-D world into 2-D image plane can cause ambiguities while determining the movement of objects. Let us look at this effect by an example. Figure 9.a gives an image of a circle and the axis that goes through that circle. Figure 9.b gives an image that could result from two different transformations of the circle in the first image. The circle might have deformed to an ellipse, or it might have rotated across its axis. It is not possible to resolve this ambiguity using 2-D imaging. A similar ambiguity is observed when reading the lips of a person. The movements of the lip can cause similar ambiguities using a 2-D image. The Canesta sensor provides the 3-D size, shape and range information directly, eliminating similar type of ambiguities that could arise.

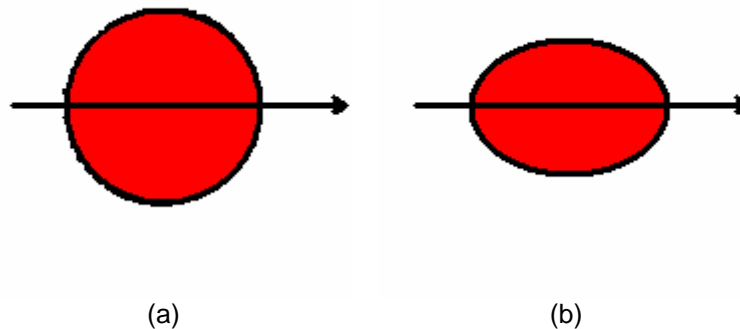


Figure 9. An ambiguous case. Has the circle rotated around the axis or has it deformed to an ellipse?

Conclusion

This paper presented the reasons why 3D imaging is crucial for enabling everyday digital devices to “see” in a sense that they can understand the meaning of these images. These reasons include:

- Robustness to environmental conditions (e.g. lighting)
- Robustness to scene conditions (textures and colors of objects such as clothing)
- Ease of segmentation, and separation of foreground from background (resulting in dramatically lower processing power)
- Accuracy in identification and tracking
- Ability to reliably determine object size

These key capabilities are essential to any system which endeavors to truly understand and interact with its surroundings. Such systems that do will have the ability to provide capabilities and experiences well beyond what is commonly available today.