OBJECT RECOGNIZATION FROM STRUCTURAL INFORMATION USING PERCEPTION

K. Vindhya Gupta[1], B. Akshitha[2], Samhita. A.S.S[3], K Radha[4]

[1,2,3]B.TECH III-Year, CSE, GITAM UNIVERSITY, Rudraram, Hyderabad
[4]Asst Professor, CSE, Rudraram, GITAM UNIVERSITY

Abstract - *Machine perception is the ability to use input from sensors such as cameras visible spectrum or infrared, microphones, wireless signals, and active sonar, radar, and tactile sensors to deduce aspects of the world. Applications include speech recognition, facial recognition, and object recognition. Computer vision is the ability to analyze visual input. Such input is usually ambiguous, a giant, fifty-meter-tall pedestrian far away may produce exactly the same pixels as a nearby normal-sized pedestrian, requiring the AI to judge the relative likelihood and reasonableness of different interpretations, for example by using its object model to assess that fifty-meter pedestrians do not exist. Classifiers are functions that use pattern matching to determine a closest match. They can be tuned according to examples, making them very attractive for use in Artificial Intelligence. These examples are known as observations or patterns. In supervised learning, each pattern belongs to a certain predefined class. A class can be seen as a decision that has to be made. All the observations combined with their class labels are known as a data set. When a new observation is received, that observation is classified based on previous experience.*

Key Words: Image formation, object recognition, 3D world, object recognition, structural information.

1. INTRODUCTION

By the interpretation of response of sensors, perception provides the information about the world. A sensor can be used as an input for an agent program and could be as simple as switch. There are varieties of sensory modalities which are available to artificial agents. Modalities which are available to human include vision, hearing and touch whereas modalities which are not available to the unaided human include radio, GPS, infrared and wireless signals. Active sensing is performed by some robots which send signals such as ultrasound or radar.
A sensor model consists of two components: Object model and Rendering model. Object model describes the objects that reside in visual world such as people, buildings, trees, cars etc. It includes a accurate 3D geometric model and indistinguishable constraints. Rendering models are quite accurate but ambiguous which describes the physical, geometric and statistical processes that produce the stimulus from the world.

A vision-capable agent problem is: *Which aspects should be considered to help the agent make good action choices and which are to be ignored?* There are 3 approaches to overcome this problem: Feature extraction, Recognition, Reconstruction. The feature extraction emphasizes simple computations which apply to the sensor observations, as exhibited by drosophila (fruit flies). The Recognition approach an agent draws distinction among the objects, encounters based on visual and other information. The reconstruction approaches an agent builds a geometric model of the world from an image or set of images.

2. IMAGE FORMATION

Imaging distorts the appearance of objects. For example, if you hold your hand infront of your eye, you can block out the moon, which is not smaller than your hand. As you move your hand back and forth or tilt it, your hand will seem to shrink and grow in the image, but it is not doing so in reality.

2.1 Images without Lenses

The pinhole camera Image sensors gather light that are scattered from objects in a scene and create a 2D image. The image is formed on retina which consists of two types of cells: about 100million rods and 5million cones. Cones are essential for colour vision. In cameras, the image is formed on an image plane which can be a rectangular grid of a few million photosensitive pixels, each a complementary metal oxide semiconductor or charge-coupled device. When a photon arrives at the sensor, it produces an effect whose strength depends on the wavelength on the photon. Through a pinhole camera, we can view stationary objects which helps to form a focused image. It consists of a pinhole opening, O, at the front of a box and an image plane at the back of the box. If the photons are small enough to pass through pinhole, then nearby photons in the scene will be nearby in the image plane, and the image will be in focus.

2.2 Lens Systems

A drawback in the pinhole camera is that we need a small pinhole to keep the image in focus. If we gather more photons by keeping the pinhole open longer it will get motion blur. If we can't keep the pinhole open any longer, we can try to make it bigger. A large opening is covered with a lens that focuses light from nearby object locations down to nearby locations in the image plane. The lens systems have limited depth of field which can focus light that lies within the range. Objects outside this range will be out of focus in the image.

2.3 Light and Shading

The 3 main causes of varying brightness are overall intensity,reflectandshading.Overall intensity of light: A white object in shadow may be less brighter than the black object in direct sunlight but the eye can distinguish the relative brightness and perceive the white object as white. Reflect: Due to different points in the scene, it may reflect moreorlessofthelight.

Shading: This effect is caused when surface patches facing the light are brighter than the surface patches tilted away fromthelight.Most of the surfaces reflect light by a process of diffuse reflection. It scatters the light evenly across the directions leaving a surface, so the brightness of diffused surface doesn't depend on the viewing direction.Paints, cloths,rough stone, vegetation and rough wooden surfaces are diffuse. In case of mirrors, it depends on the way or direction you look at the mirror and hence are not diffuse. The behavior of a perfect mirror is known as specular reflection. On some surfaces like brushed metal, plastic and wet floor, we observe small patches where specular reflection has occurred known as specularities. Distant point light source is defined as the parallel rays travelling from the main source of illumination i.e sun. The angle theta describes the amount of light collected by the surface patch. A part of light collected is reflected by diffuse surface patch known as diffuse albedo. Shadow is the surface which cannot be reached by light. The constant ambient illumination is used to predict intensity by using effects like inter reflections and shadows.

2.4.Color

The principle of trichromacy states that for any spectral energy density there can be another spectral energy density constructed using a mixture of 3 colors (usually red, green and blue) so that humans can't differentiate them. The above principle helps making computer algorithms easier. The 3 different albedos R/G/B helps in modeling every surface and similarly every light source. The humans are able to identify the color of the surface under the white light ignoring the effects of different coloured lights, this is known as color constance.

3. OBJECT RECOGNITION BY APPEARANCE

Appearance means what an object tends to look like. For example, football is rather round in shape. It is important to know every class of images with a classifier. Taking the example of faces, looking at the camera-every face looks similar under good light and perfect resolution. A strategy called sliding window includes computing features for an object and present it to a classifier. One strategy is to estimate and correct the illumination in each image window and another is to build features out of gradient orientations. To find faces of different sizes, repeat the sweep over larger or smaller versions of the image. Then, we post process the responses across scales and location to produce the final set of detections. Postprocessing have several overlapping windows that each report a match for a face. To yield a single high quality match, we can combine these partial overlapping matches at nearby locations. Therefore, it gives a face detector that can search over locations and scales.

3.1 Complex appearance and pattern elements

Since, several effects can move features around in an image of the object, many objects produce much more complex patterns than faces do. Effects include: Foreshortening: which causes a pattern viewed at a slant to be significantly distorted

Aspect: which causes objects to look different when seen from different directions.

Occlusion: when some parts are hidden from some viewing directions. Self-occlusion can be defined as objects can occlude one another or parts of an object can occlude other parts.

Deformation: where internal degrees of freedom of the object change its appearance. An object recognizer is then a collection of features that can tell whether the pattern elements are present, and whether they are in about the right place. With a histogram of the pattern elements that appear can be considered as the most obvious approach to represent the image window. This approach does not work particularly well, because too many patterns get confused with one another.

3.2 Pedestrian detection with HOG features

Each year car accidents kill about 1.2 million people, and to avoid this problem cars should be provided with sensors which detect pedestrians and result in saving many lives. The most usual cases are lateral or frontal views of a walk. In these cases, we see either a "lollipop" shape-the torser is wider than the legs, which are together in the stance phase of the walk-or a "scissor" shape-where the legs are swinging in the walk. Therefore, we need to build a useful moving-window pedestrian detector. To represent the image window, it is better to use orientations than edges because there isn't always a strong contrast between a pedestrian and the background. Pedestrians can move their arms and legs around, so we should use a histogram to suppress some spatial detail in the feature. When breaking up the window into cells, overlapping occurs and hence, build an orientation histogram in each cell. Through this feature, we can determine whether head-and-shoulders curve is at the top of the window or at the bottom, but will not change, if the head is moved bit slightly.

As orientation features are not affected by illumination brightness, we cannot treat high contrast edges specially. We can recover contrast information by counting gradient orientations with weights that reflect how significant the gradient is compared to other gradients in the same cell. HOG feature (Histogram of Gradient Orientations): This feature compares the gradient magnitude to others in the cell, so gradients that are large compared to their neighbours get a large weight. This feature construction is the main way in which pedestrian detection differs from face detection. The detector sweeps a window across the image, computes features for that window and present it to a classifier.

4. RECONSTRUCTITING THE 3D WORLD

This section shows how to go from 2D image to a 3D representation of the scene. A fundamental question arises which is how do we recover 3D information when given all the points in the scene that fall along a ray to pinhole are projected to the same point in the image? The two solutions are,

- If we have two or more images from different camera positions, then we can triangulate to find the position of a point in the scene.
- We can exploit background knowledge about the physical scene that gave rise to the image. Given an object model P(scene) and a redering model P(image |scene).We can compute a posterior distribution P(scene|image).

For a scene reconstruction, we survey 8 commonly used visual cues because there is no single unified theory for it.

4.1 Motion Parallax:

We state an equation understanding the concept of relation among the optical flow velocity and depth in the scene.

$$v_x(x,y) = \frac{-T_x + xT_z}{Z(x,y)}, \quad v_y(x,y) = \frac{-T_y + yT_z}{Z(x,y)},$$

The point in the scene corresponding to the point in the image at (x,y).Both components of the optical flow,$v_x(x,y)$ and $v_y(x,y)$, are zero at the point $x=T_x/T_z$,$y=T_y/T_z$ .This point is called focus of expansion of the flow field. If we change the origin in the x-y plane to lie at the focus of expansion, then the expressions for optical flow take on a particularly simple form. Let (x',y') be the new coordinates defined by $x'=x-T_x/T_z$,$y'=y-T_y/T_z$. Then,

$$v_x(x',y') = \frac{x'T_z}{Z(x',y')}, \quad v_y(x',y') = \frac{y'T_z}{Z(x',y')} .$$

4.2 Binocular Stereopsis

This idea is similar to motion parallax, but we use two or more images separated in space. Binocular stereopsis enables a wilder field of vision for the predators who have eyes in the front. If we super pose the two images there will be a disparity in the location of the image feature in the two images as a given feature in the scene will be in a different place relative to the Z-axis of each image plane. Using optical flow equations vector T acting for time δt, with $T_x=b/\delta t$ and $T_y=T_z=0$. Horizontal disparity is equal to the ratio of base line to the depth, and vertical disparity is zero i.e,H=b/Z,V=0. Humans fixate at a point in the scene at which the optical axes of the two eyes intersect nder normal viewing conditions. The actual disparity is δθ, we have disparity=$b\delta Z/Z^2$.

In humans, b(the base line distance between the eyes) is about 6cm. So for Z=30cm we get small value δZ=0.036mm which means at a distance of 30cm humans can differentiate the depths which differ by a little length as 0.036mm.

4.3 Multiple views

Most of the techniques that have been developed had make the use of the information available in multiple views, even from hundreds or thousands of cameras. There are few problems in multiple views which can be solved algorithmically:

- Correspondence problem: In the 3D world, identifying features in the different images that are projections of the same feature.
- Relative orientation problem: Finding out the transformation between the coordinate systems fixed to the different cameras.
- Depth estimation problem: Finding out the depths of various points in the world for which image plane projections were available in at least two views.

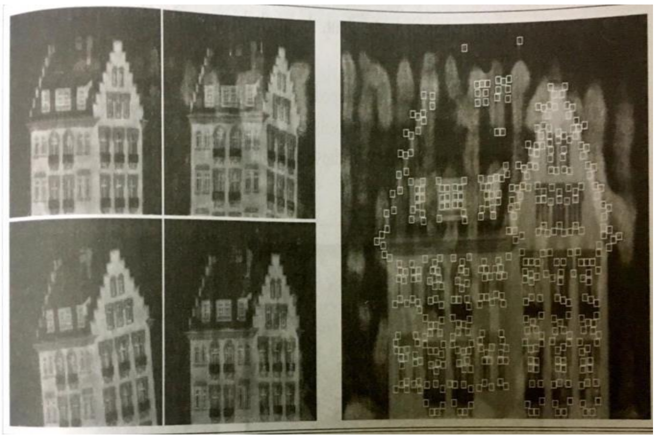

Fig.1. Binocular Stereopsis

Using numerically stable algorithms for solving relative orientations and scene depth and the development of robust matching procedures for the correspondence problem is one of the successful attempts in computer vision.

4.4 Texture

Texture is used to estimate distances and for segmenting objects. The texture elements are also known as texels.

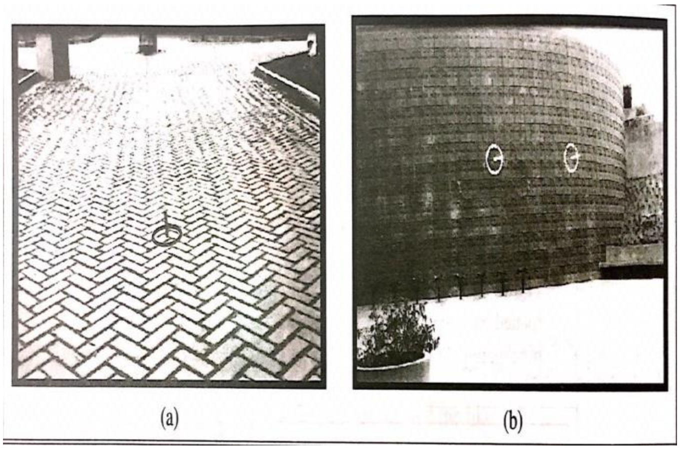

Fig.2. Texture

As shown in the above figure, the paving tiles are identical in the scene but appear different in the image for 2 reasons.

- Differences in the distances of the texels from the camera.
- Differences in the foreshortening of the texels.

Various algorithms have been developed to make use of variation in the appearance of the projected texels as a basis for finding out the surface normals but they were not accurate as much as the algorithms which were used for multiple views.

4.5 Shading

From different portions of a surface in a scene, we receive variation in the intensity of light called shading. It is determined by the geometry of the scene and by the reflectance properties of the surfaces. It has been proved to be difficult to invert the process i.e to recover the geometry and reflectance properties given the image brightness. If a surface normal points towards the light source, the surface is brighter and if it points away, the surface is darker. The surface might have low albedo because it changes quickly in images rather than shading which changes slowly. The solution for this problem is when we assume that albedo is known at every surface point.one of the measurement is brightness and another is normal. We cannot solve for normal, therefore we can consider that the nearby normal might be similar as most surfaces are smooth. There is a quite difficulty with the interreflections because of mutual illumination, when consider an indoor scene, makes it difficult to predict the relation between the normal and image brightness. Two surface patches with the same normal might have quite different brightness's as one receives light reflected and other facesonly a dark scene.

4.6 Contour

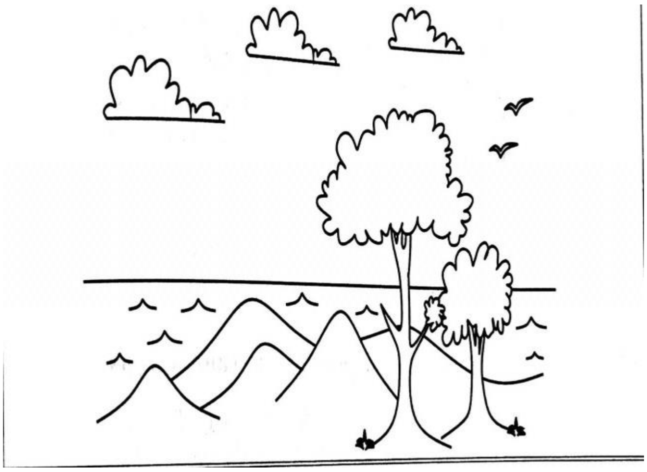

Fig.3. Contour

Considering the above line drawing, we get 3D shape and layout which is the combination of recognition of familiar objects in the scene and the application of generic constraints such as following:

- **Figure-ground-problem:** figuring out which side of the contour is nearer and which is farther. At an occluding contour, the line of sight is tangential to the surface in the scene.
- **T-junction:** It results in image when one object occludes another, by which the nearer ones look opaque and farther one is interrupted.

Considering a scenario, the projection of objects of different heights is at different locations on the ground plane. Suppose an eye or a camera is at a height $h_{c,}$, above the ground plane. Consider an object of height δY resting on the ground plane, whose bottom is (X,-h_C,Z) and top is at (X,δY-h_c,Z).The bottom projects to the image point(f X/Z,-fh_C/Z) and the top to (f X/Z,f(δY-h_C/Z).The bottoms of nearer objects project to points lower in the image plane, farther objects have bottoms closer to the horizon.

5. OBJECT RECOGNIZATION FROM STRUCTURAL INFORMATION

Human body parts are very small in images and vary in color, texture among individuals and hence it is difficult to detect them using moving window method. Some parts are very small to a size of two to three pixels wide. As the layout of body describes the movement, it is important to conclude layout of body in the images. whether the configuration are acceptable or not it can be described by a model called deformable template. The simplest deformable template model of a person connects lower arms to upper arms, upper arms to torso, and so on.

5.1 The geometry of bodies: finding arms and legs

A tree of eleven segments is used to model the geometry of body. Segments are rectangular in shape." Cardboard people" are few models which are used to assume the position and pose(orientation) of human body parts and segments in the images. Evaluation of configuration is done based on 2 criteria: First, an image rectangle should look like its segment. A function φ_i describes how well an image rectangle matches a body segment. Function ψ defines how the relations of a pair of rectangle segments meet the expected body segments. Each segment has only one parent because the dependencies between segments form a tree. The parent can be described by a function $\psi_{i,p,a(i)}$.

$$\sum_{i\in\text{segments}} \phi_i(m_i) + \sum_{i\in\text{segments}} \psi_{i,\text{pa}(i)}(m_i, m_{\text{pa}(i)}) .$$

There is an angle between segments mainly for the ankles and knees to be differentiated. If there are M image rectangles having the right torso as $O(M^6)$ for a model, then the best allocation of rectangles to segments will get slow. However this can be solved by using various speed-ups which are available for an appropriate choice of ψ. This model is usually known as pictorial structure model.

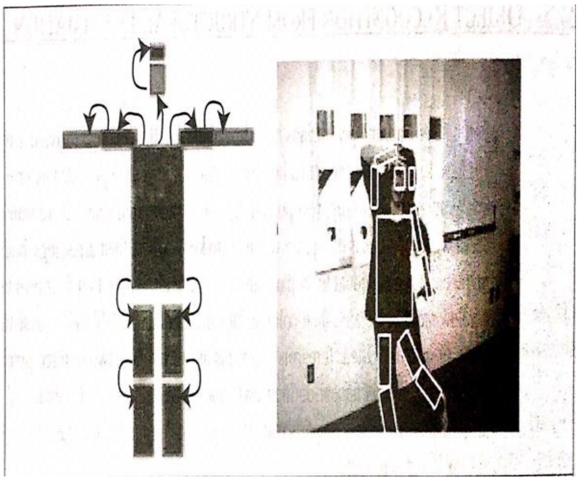

Fig.4. The geometry of bodies: finding arms and legs

We generally have to build a model of segment appearances when we don't have an idea of what a person looks like. Appearance model is the description of what a person looks like.

5.2 Coherent appearance: tracking people in video

By improving the concept of tracking people in a video we can get further in game interfaces and surveillance systems, but many methods haven't succeeded with the problem because people in the videos tend to move fast and produce large accelerations. The effective methods state the fact that, from frame to frame the appearances change. We assume the video to be a huge collection of pictures we wish to track. The people in the video can be tracked by detecting their body segments. In some cases, the segment detector may generate lots of false positives because the people are appearing against a near fixed background. This problem can be solved using an alternative which is quite in practice, by applying a detector to a fixed body configuration for all the frames. We can know when we found a real person in a video by tuning the detector to low false positive rate.

6. CONCLUSIONS

Perception requires a significant amount of sophisticated computation and for the extraction of information are needed for tasks such as manipulation, navigation, and object recognition.

- The physical and geometric aspects has well-defined the process of image formation. Using the described 3D scene, a picture can be produced from the position of arbitary camera.
- The primitive features from the image are extracted using image-processing algorithms.
- The 3D information is used to obtain various cues in the image. We get accurate interpretations as every cue relies on background assumptions about physical scene.

REFERENCES

[1] Artificial intelligence, a modern approach(third edition) by Stuart J.Russell, Peter Norvig

[2] Wikipedia:https://en.wikipedia.org/wiki/Computer_vision

[3] https://en.wikipedia.org/wiki/Speech_recognition