# Introduction

***Visual Servoing and its applications:*** With all the various rapid developments in the technological field, we have started to make computers see things. However, they still lack in understanding what they capture or see. Giving machines and robots the ability to see and comprehend the surrounding is called Visual Servoing. Visual Servoing, also known as vision-based robot control, is a technique to control a robot using the information provided by a vision system.

Three main processes in Visual Servoing are:

VS is a mixture of research areas:

Visual Servoing is used in rich variety of applications such as lane tracking for cars, navigation for mobile platforms and manipulation of objects. Vision can be used to control disparate dynamic systems like for example vehicles, aircrafts and submarines. They can increase the flexibility and the accuracy of robotic systems.

# Visual Servoing Techniques.

1. Position-Based Visual Servoing (PBVS) / 3D Servoing

This technique retrieves the 3D information about the scene where known camera model is used to estimate the position and the pose (orientation) of the target with respect to the camera (world, robot) coordinate system. The positioning or tracking tasks are defined in the estimated (3D) pose space.

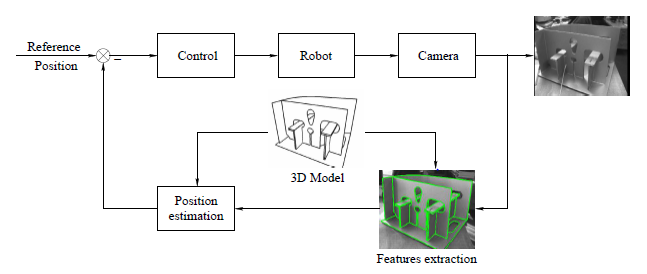


Figure 1 Model-based 3D Visual Servoing

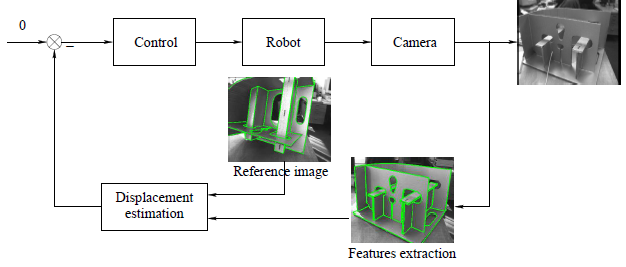


Figure 2 Model-free 3D Visual Servoing

1. Image-Based Visual Servoing (IBVS) / 2D Servoing

Proposed by Weiss and Sanderson, 2D image measurements are used directly to estimate movement of the robot. The control law is based on the error between current and desired features on the image plane, and does not involve any estimate of the pose of the target. The features may be the coordinates of visual features, lines or moments of regions.

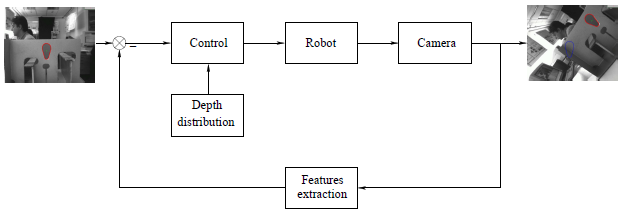


Figure 3 2D Visual Servoing

1. Hybrid approaches
   1. 2 ½ D Servoing
   2. Motion partition-based

# A vision system can further be classified based on variation in each component of the Visio system, such as

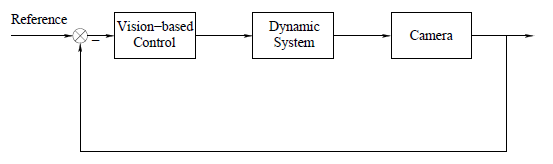
1. The number of cameras
   1. Single camera
   2. multi-camera
2. Camera position.
   1. In-hand or Eye-in-hand or end-point open-control-loop

The camera is attached to the robot/end-effector.

* 1. Out-hand or Eye-to-hand or end-point closed-control-loop

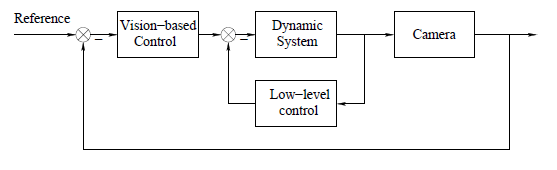
The camera is fixed in the world and observes the target and the robot.

1. Control scheme
   1. Direct-Visual Servoing System



In this control scheme, the vision-based controller directly computes the input of the dynamic system, as a result, they are very fast.

* 1. Indirect-Visual Servoing System



In this scheme, the vision-based control computes a reference control law which is sent to the low-level controller of the dynamic system. Most of the VS proposed in the literature follows an indirect-scheme called “dynamic look-and-move”.

# Manipulation task

A large amount of work as been done in visual servo control for manipulation. A manipulation task typically consists of following subtasks:

1. Recognize an object
2. Approach the object
3. Grasp the object
4. Manipulate the object

The general approach to perform a manipulation task in IBVS can be stated as:

* Defining the task function in image measurements.
* Feature detection and description
* Depth estimation
* Kinematics

# Literature Analysis

The aim of survey is to gain a general idea about the concepts and terminologies along with the present researches and advancements. Summary of most state of the art research algorithms are described below.

### Ezio Malis, “Survey of vision-based robot control”, presented at ENSIETA 2002

The aim of Visual Servoing approach is to control a robot using the information provided by the visual system. Vision systems are generally classified depending on the number of cameras and on their position.

Classification based on design of control scheme

1. Direct visual servo system

Vision-based controller directly computes the input of the dynamic system.

Very fast (at least 100Hz, with rate of 10ms)

1. Indirect visual servo sys / Dynamic look-and-move

The vision-based control computes a reference control law which is sent to low-level controller of dynamic system. In this case, the servoing of the inner loop must be faster than the Visual Servoing.

Vision System

Using pin-hole camera, the perspective projection of a 3D point to the image plane. Image plane is a matrix of light cells with (u,v) coordinates, the camera measures the intensity of light.

The intrinsic parameters of the camera are often only roughly known. Precise calibration is a tedious procedure. It is thus preferable to estimate the intrinsic parameters without knowing the model of the observed object. If several images of any rigid object are available it is possible to use a self-calibration algorithm to estimate the cam intrinsic parameters.

Feature extraction: Generally, use points as visual features like points, straight lines, ellipses, contours, etc. and algorithms like Harris detector and Canny

Matching features: The matching problem consist in finding the features in two or more images. Difficulties increases based on the displacement of camera and light conditions.

1. model based approach: match the current image to a model
2. model-free approach: match features between initial and reference views

Visual Servoing approaches

Based on the knowledge camera parameters.

1. Calibrated Visual Servoing: camera parameters are known
2. Uncalibrated Visual Servoing: roughly known

Based on the priori knowledge of target

1. Model-based Visual Servoing: if 3D model of target is available.

If at least 4 points, possible to compute desired camera pose and current camera pose, camera parameters must be perfectly known.

1. Model-free Visual Serving: 3D model of the target is unknown, “teaching by showing” approach. Error function is decided using reference image and current

Vision-Based control

1. Position-based (3D)

Control law directly uses the error on the position of camera. Depending on the available visual features in the image, one can compute the pose error.

Advantages: works directly in Cartesian space

Limitations: Since no control in the image, image features used for pose estimation may leave the image.

If camera is coarsely calibration or error in 3D model of target, current and desired camera pose inaccurate.

PBVS approaches

1. Model-based
2. Model-free
3. Image-based (2D)

A model-free approach in which error function is expressed directly in 2D image space.

Adv: It is robust to camera and robot calibration.

Limitations: it convergence is theoretically ensured only in a region around desired position.

Solution of depth estimation

1. Estimate interaction matrix on-line
2. Approximate the Interaction matrix
3. Estimate the depth distribution

### D Kragic, et al., “Survey on Visual Servoing for manipulation”, ISRN KTH/NA/P–02/01–SE, 2002

Using visual feedback to control a robot is called as Visual Servoing. Visual (image based) feature such as points, lines and regions can be used to, for example, enable alignment of manipulator. Three main processes: Initialization, tracking and robot control. Most general applications require detection of object, segmentation, recognition, servoing, alignment and grasping.

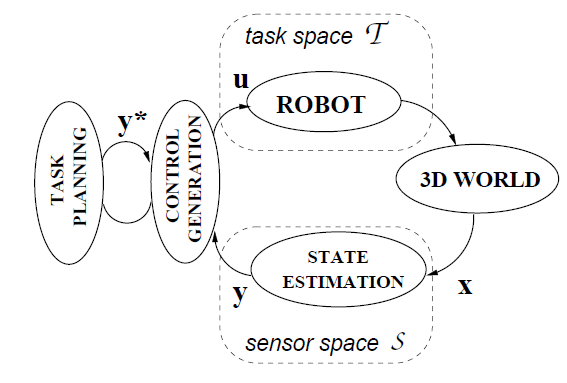
* Different ways of using visual info:
  + Open-Loop
  + Visual Servoing

1. Indirect visual servo sys (Dynamic look-n-move)
2. Direct visual servo sys

* Image info: PBVS, IBVS, 2 1/2D visual servo system

Visual Servoing is mixture of research areas like robot modelling (geometry, kinematics, dynamics), real-time system, control theory, system (sensor) integration, computational vision (image processing, structure-from-motion, camera calibration).

There are many different ways of classifying VS system: based on sensor configuration, number of camera, generated motion command, scene interpretation, etc. So, following is a generalized taxonomy.



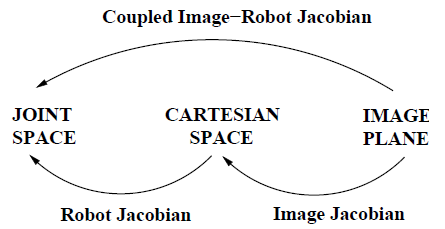
A relative Euclidian motion of the robot is defined as the input **u** in the *task space* of the robot, T. **x** represents the state vector, **y** is the measurement vector and **y**\* is the vector of desired measurements in the *sensor space*, *S*.

* + Visual-Motor model estimation

(Corke & Good, 1996) visual kinematics deals with how a manipulator should move in response to perceived visual information while visual dynamic control accounts for the dynamic effects that usually occur in robotic system.

* + State estimation: issues related to visual measurements (optical flow, position, orientation or features) are addressed: camera config, number of camera, image processing techniques. Sensor-Robot config include eye-in hand, stand-alone or combination.
  + Control generation

*Visual-Motor model estimation*: If the robot forward or inverse kinematics are known, the differential changes between the joint and Cartesian space are computed using robot Jacobian. These systems are classified as systems where visual-motor model is known *a-priori*. Depending on the feedback representation mode and level of calibration between the camera and the robot frame, the visual servo systems are classified as position based, image based or 2 1/2 D systems.



*Obtaining Visual measurements*: Visual servo systems in general employ visual feedback to obtain measurements either:

1. directly in the image plane using correlation-based methods, optical flow techniques, image differencing, or
2. use camera parameters and some pre–knowledge about the observed image features (CAD models) to determine the pose of the object. The adopted techniques usually depend on the number of cameras used, camera configuration, the level of calibration and some pre-knowledge about the scene.

Binocular vision

Two cameras in a stereo arrangement may be used to provide complete 3D information about the scene. One of the common approaches is to estimate the disparity which is then used for depth estimation. However, correspondence problem is a major obstacle.

### M. Eriksson, et al., “Vision based Servoing and grasping with multiple control schemes”, in *SIRS ‘89–7th International Symposium on Intelligent Robotic Systems*, pp. 51–60, July 1999

A manipulation task mainly consists of following subtasks:

1. Recon object
2. Approach object
3. Grasp object
4. Manipulate object

The general approach to perform object recognition can be divided into 4 different stages:

1. Figure Ground Segmentation
2. Feature Extraction
3. Indexing
4. Verification

The system starts by segmenting all the potential slots and pieces from the background. One of the segmented slots is then randomly selected. The recognition system then selects the correct piece. The recognition is based on matching geometric invariants on the contours. This builds on the fact that 2 contours resulting from two different projections of 2D shape differ only by a homography.

1. Figure ground segmentation

In order to do recognition, figure ground segmentation is generally needed in order to extract the potential objects from background. The author used color-based segmentation in this case. He manually sets the RGB value.

1. Feature Matching

Contours are extracted directly from the figure ground segmentation step, travelling one lap around each blob. Note that too small blobs are considered noise and the curves can be smoothed with Gaussian filter.

1. Shape matching

Most approaches to comparing 2 planar objects are based on projective invariants. Once the homograph is found, the projection matrix T can be determined by identifying 4 matching points. There are different ways to find matching points such as inflexion points, bitangents.

Servoing

The general problem to be solved when doing IBVS is how to map a feature vector extracted from image into a joint control vector to be applied to the set of joints on the manipulator. This mapping is generally referred to as the image Jacobian. By using the robot Jacobian, the joint angle velocities can be determined from the desired in Cartesian space, as long as the pose is within the workspace of manipulator, inverse kinematics.

1. Approaching the object

A large problem with IBVS in general is how to estimate the image Jacobian J, and how to avoid J to be ill conditioned. Also, small error in feature vector generates large perturbations in J. The author constrained the DoF to 2 and iterated the following steps until a threshold is reached.

1. Get position of object in image coordinate
2. Find pitch and yaw angles so that the object becomes centered in the image
3. Move towards the object (along the Z-axis in reference frame of the camera)
4. Pre-grasping

Pre-grasping involves finding a config of the arm so that the z-direction of the end-effector is perpendicular to the table-surface.

1. Update roll, pitch and yaw
2. Center the object in image without changing roll, pitch or yaw.

In order to grasp the object, a translation is required to move the gripper to the current position of camera. After this translation, end-effector is moved downwards until indicated by force-torque sensor.

Implementation

Nomadic XR 4000 platform with Puma560 on top. Image processing carried out under Linux, arm-control under Real-Time OS QNX. The computers are interconnected using TCP/IP implemented on top of SMA

### Muhammad Ahsan Sami, et al., ”Object tracking with a robotic manipulator mounted on a ground vehicle using IBVS”, 2017.10, 206-210 (5 pages), ICCAS 2017

There is a constant need for safer methods to reconnaissance or collect surveillance data from hazardous environment. Visual algorithms can help robots with human-like vision. Visual Servoing techniques can be implemented on the camera feed to control a robot. ***Visual Servoing*** *systems are of two types,* ***IBVS****(Image Based Visual Servoing) and* ***PBVS****(Position Based Visual Servoing).* In PBVS, robotic manipulators tracks object using video feedback. Although PBVS is better when displacement from the goal is large, IBVS are better in close proximities. IBVS takes continuous real-time images and compares it with previous images to move manipulator accordingly so that object to be tracked can always be in center of image. Involves less calculation and removes need of calculating coordinates of the object in contrast to PBVS. *Hence, IBVS increases robustness of system and makes it less dependent on camera calibration.* **Inverse kinematics** is used to find orientation of robotic manipulator with respect to object being tracked. Paper discusses the control of a 2DOF. Accuracy of resulting operation is improved by accuracy of robot motion and sensor as well as closed loop control is used for better efficiency. Visual Servoing technique and a PID controller is used for precise tracking. OPENCV, for identification of moving objects anomalies with respect to usual routine environment.

Specification of movement of robot so that its end effectors achieves the desired task is known as **Motion Planning**. To obtain desired motion, **interaction matrix** was linked with the manipulator kinematics (DH parameters and joint constraints).

*Control (manipulator) = Control (pitch & yaw motors)*.

The turret sys involves a remote operator and turret itself. User commands and turret makes movement.

* Image processing is done in Visual C++ and object tracked is continuously monitored through cam feed. Object position and orientation are continuously monitored and feedback data is updated to control yaw and pitch.
* Detection and tracking is done by VS library called OpenCV (Computer Vision). A threshold edge image is the starting point of **Hough transform** (algorithm in OpenCV).

Edges are very imp in an image as they carry some imp visual info since they delineate elements of image.

Drawbacks of simple binary edge map:

1. Edges detected are unnecessarily thick.
2. It is difficult at the same time to find a threshold which is high enough to not include too many insignificant edges and low enough to detect all imp image edges.

This trade-off problem was solved using **Canny algorithm**.

The paper uses Hough transformation for detection and tracking the object in binary edge map of image and uses canny algorithm for trade-off.

### Ehab Salahat, Murad Qasaimeh, “Recent advances in feature extraction and description algorithms: a comprehensive survey”, IEEE 2017

Feature detection and description algorithms can indeed be considered as the retina of eyes of machines and robots. The concept of feature detection and description primarily aims towards object detection, analysis and tracking from a video/image stream to describe the semantics of its actions and behavior. Most of the proposed algorithms require intensive computation. Hence, hardware accelerators, such as DSP, FPGA, SoC, ASIC and GPU, with massive processing capabilities are required in real-time applications. However, hardware-constraints such as memory, power, scalability and format interfacing constitute a major bottleneck. Processing Algorithms will make substantial contribution to resolve these issues. To ensure robustness of Vision Algorithm, an essential prerequisite is that they are designed to cover a wide range of possible scenarios with a high-level of repeatability and affine-invariance. Key factors influencing real-time performance include the processing platform, monitored environment, and the application of interest.

Local image features (also known as interest points, key pts and salient features) can be defined as a specific pattern which is unique from its immediate close pixels. It includes edges, corners, regions, etc. These local features are then converted into numerical description, representing unique and compact summarization. Feature extraction plays the role of an immediate image processing stage between different CV Algorithm. Applications: real-time surveillance, image retrieval, vid mining, object tracking, mosaicking, target detect and wide baseline matching et al.

In general, local features have spatial extent. Ultimately, it is infeasible to localize all such features as this will need the prerequisite of high-level frame understanding. Ideal features should typically have certain important quality: Distinctiveness, locality, quantity, accuracy, efficiency, repeatability, invariance, robustness. CV has to decide the trade-offs.

MSER derivatives

* N-dim extension
* Linear-time MSER Algorithm
* X-MSER
* Parallel MSER

Others include the Extremal Regions of the Extremal Levels Algorithm and the Tree-based Morse Regions (TBMR).

SIFT derivatives

SIFT algorithm has a local feature detector and local histogram-based descriptor. It detects sets of interest points in an image and for each point it computes a histogram-based descriptor with 128 values. Number of algorithms tried to reduce the SIFT descriptor. Others used different window size and histogram computation pattern around interest pts.

* ASIFT
* CSIFT
* n-SIFT
* PCA-SIFT
* SIFT-SIFER retrofit

Other derivatives include SURF, SIFT CS-LBP Retrofit, RootSIFT Retrofit, and CenSurE and STAR algorithms.

### Dineesh Mohan, Dr. A. Ranjith Ram, “A review on Depth estimation for computer vision applications”, IJEIT volume4, issue 11, May 2015

When looking out the side window of a moving car, the distant scenery seems to move slowly while the lamp posts flash at high speed. This effect is called Parallax and it is used to extract geometrical information from a scene. The distance of the objects is obtained by using multiple captures of the same scene at different viewpoints. Disparity means pixel displacement between corresponding points in the multi-view images or stereo image.

The Human Vision System (HVS) is prepared for the depth perception. The most important source of depth perception is the two eyes, sharing a common area of vision. Parallax also provide information about distance of object. Brain also computes the distance of the focused plane. Distance of stars is calculated using constant illumination. The main difficulty is the correspondence problem, stereo-matching which determines the spatial displacement between each two corresponding pixels in stereo pair. Generally, Stereo-Matching algorithms are categorized into 2 major classes: local and global. Most of the methods are computationally expensive. Recently, the scene reconstruction problem was treated as a matching problem where the objective was to match points or features between 2 or more images. After the match is obtained, 3D position of point is determined using triangulation method assuming camera position are known.

Various methods for depth estimation:

1. Vergence

When both eyes are positioned in such a way that the optical axis intersect on the surface of an object. Thus, stereo fixation of the object is obtained. This type of movement is called vergence. And then triangulation is carried out.

1. Stereo disparity

An object which is not at stereo fixation and it projects to different locations on the Left & Right retina according to scene depth and horizontal baseline separating the eyes. The difference between these 2 locations is called Stereo Disparity (SD). Disparity maps are computed using block matching algorithm. Color-based segmentation methods are used here to obtain the disparities of the object in the disparity maps and the average was taken for depth estimation.

1. Stereo matching

The matching process consists of identifying each physical point within different images. However, matching techniques are not only used in stereo or multi-vision procedures but widely used for image retrieval or fingerprint identification where it is important to allow rotational and scalar distortions. There are various constraints that are generally satisfied by true matches thus simplifying the depth estimation algorithm, such as similarity, smoothness, ordering and uniqueness. The matching is, hence, implemented by means of comparators allowing different identification strategies such as minimum square errors (MSE), sum of absolute differences (SAD) or sum of squared differences (SSD).

1. Familiar size

The depth of an object can be estimated from the size of its projection on the camera images if the real size of the object is known.

1. Using Bayesian cue integration in combination
2. Using defocus cue

It uses two defocused observations of a scene that are captured with different camera parameters for the depth estimation. This method consists of two steps. In the first step, depth estimate is obtained for the focused image. In the second step, fast optimization is used for refining the solution

1. Convex optimization approach

A spatially varying multiplicative model is developed to account for brightness changes induced between left and right views

1. Using object-placement relation

### Dr. JL Mazher Iqbal, Shaik Shakir Basha, “Real time 3D depth estimation and measurement of Uncalibrated stereo and thermal images”, ICNTE-2017

To determine an object‘s distance relative to an observer is a common field of research in CV and image analysis. Currently existing approaches for stereo vision are classified into 2 categories such as monocular, binocular, trinocular or stereo sequence. The active stereo system such as ultra sound and laser gives depth directly. In passive system algorithm such as Sum of Absolute Difference (SAD), Random Sample Consensus (RANSAC) is used. The Stereo Vision analysis with calibration method seems to be complex process which does not satisfies time and space limits. The oriented camera captures uncalibrated images. The corresponding points in between 2 images are obtained from any 1 of the 3 methods such as 2D-2D relative orientation (recovery motion), 2D-3D image orientation (space resection or camera calibration) and 3D-3D abs orientation (3D similarity transformation).

Stereo vision

1. Triangulation

Once the stereo images have been brought into point-point correspondence, the depth estimation by triangulation is straight-forward. The simplest stereoscopic vision is the position of camera along one axis or translation axis. Thus, there is no rotational relationship between the 2 cameras and so they differ only by a shift system. In the coordinate system, the X and Y world coordinates are those that reside in the plane parallel to the image plane. Z is the range from the cam to the object of interest. The baseline b and the focal length f, is vital in uncalibrated stereo images.

1. Epipolar coordinates

Proposed algorithm used in this paper.

