

# USING A 3D TIME-OF-FLIGHT RANGE CAMERA FOR VISUAL TRACKING

Ulrich Reiser \* Jens Kubacki \*

*\* Department of Robot Systems, Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, GERMANY*

**Abstract:** We propose the usage of a novel 3D-Sensor for Visual Servoing in order to avoid the problem of depth estimation that is to be solved in most servoing schemes. The sensor used in this paper is based on the time-of-flight principle and returns a depth image along with an intensity image. As these are acquired by the same CCD array, no calibration between them is necessary. As an application example a position-based eye-in hand servoing scheme is presented with the sensor mounted on the end-effector of a pan-tilt unit. An experimental tracking task was defined as follows: a sphere moving with respect to the pan-tilt unit has to be kept in the optical centre of the camera. We implemented velocity-based control of the manipulator joints and obtained the following results: the pan-tilt unit was operated at a control frequency of 17.5 Hz and — starting from a skewed pan and tilt — the sphere could be regulated to the optical center in less than one second.

**Keywords:** Visual Tracking, Time-Of-Flight Sensor

## 1. INTRODUCTION

In the last decades Visual Servoing has been a very active topic in robotics. The term Visual Servoing means the control of manipulators using visual information as feedback (Hutchinson *et al.*, 1996). Commonly visual servoing methods are categorised by the camera configuration of the vision system, i.e. the number and position of the used cameras, the servoing scheme, i.e. position-based, image-based or hybrid, and the form of the data, calibrated or uncalibrated. Furthermore, systems can be divided into End-point Closed Loop (ECL), i.e. the manipulator end-effector is tracked along with the desired target position and End-point Open Loop (EOL) otherwise (Baeten and de Schutter, 2004). The image-based servoing approach has undergone growing popularity, because the error is determined directly in the image. Thus errors in sensor modeling and camera calibration are eliminated (Corke, 1993; Chaumette and Hutchinson, 2006). The method is based on the extraction of stable feature

points in the image. A feature is defined generally as any measurable relationship in an image and examples include moments, local intensity patterns, relationships between regions or vertices etc. There is, however, the danger of the tracked feature points leaving the image. Furthermore the camera trajectory may not be as direct as with position-based servoing. Position-based servoing on the other hand needs very accurate camera and object modeling. The possible combinations of the attributes described above and the diversity of specific applications led to various visual servoing schemes, each of which has certain advantages and drawbacks. One of the problems observed in most of these methods, however, is the necessity to determine the depth of a visually controlled feature point. Depth information is often obtained from multiple views or by knowledge of the geometric relationship between several feature points on the target (Corke, 1993). In this paper the usage of a novel range sensor is proposed that measures the depth coordinate which potentially relieves many servoing methods.

## 2. RELATED RESEARCH

In the literature many approaches can be found to obtain depth of an object perceived by a camera. A common approach to determine the depth of a target is the use of multiple cameras. The most applied configuration using more than one camera is stereo, which allows for the geometrical calculation of depth by triangulation, provided that good extrinsic and intrinsic calibration is available (Andersson, 1987; Allan *et al.*, 1991). The more cameras are used, the more information is available on the target. On the other hand it is more difficult, however, to have a big zone of intersection containing the target, and the calibration effort is bigger (Malis *et al.*, 2000). Furthermore, in order to be able to calculate depth of a feature point by triangulation, the correspondence of this point in both cameras must be assured. Cipolla *et al.* (Cipolla and Hollinghurst, n.d.) propose an uncalibrated stereo camera system which is used to reconstruct the position of the tracked object and the gripper. The system uses an affine camera model instead of a projective model, which simplifies the control law but may result in depth estimation being valid only locally, if the camera displacement is too large. Hager *et al.* (Hager *et al.*, 1995) present a stereo vision system for hand-eye coordination. This position-based approach tries to regulate the target-goal disparity to zero. Again extrinsic calibration is necessary and depth must be estimated. Nevatia presents a depth from motion approach (Nevatia, 1976) with a hand-mounted camera. Depth is obtained from sequential monocular views, while the manipulator is moving. The limitation, however, is that the objects in the scene should not move significantly. Therefore this approach constitutes only limited feasibility for a tracking task. Papanikolopoulos *et al.* (Papanikolopoulos *et al.*, 1993) present visual tracking of a moving target with one camera. The system requires depth of the object moving in 2D-plane perpendicular to the optical axis to be known. Systems have been developed using a range sensor for depth measurement. Agin (Agin, 1985) describes a set-up with a light stripe projector in conjunction with a camera, mounted in a cage carried by the robot arm. Depth is calculated by triangulation and therefore careful calibration is needed. In (Venkatesan and Archibald, 1990) an approach for a tracking system of moving objects is presented, that is similar to the set-up described in this paper. Two laser range profile scanners, fixed at the manipulator end-effector, measure the distance to the object to be tracked. The manipulator is controlled to keep the end-effector at the center of the object in a certain distance. The sensor system returned two 2D scans lying perpendicular to each other — one in the X-Z plane and one in the Y-Z plane — such that the object possibly could not be detected if not intersecting the scanning planes. The sensor used in this paper, however, returns a full 3D scan at a resolution of over 20000 pixels and an aperture of 50 degrees both horizontally and vertically.

The visual servoing schemes above have to rely either on a very good calibration or on an accurate depth estimate. The use of the range sensor described in this paper relieves the problem of depth determination very much, as will be presented in the next sections. We have not found a visual servoing approach in the literature using this type of sensor so far.

## 3. SENSOR AND MANIPULATOR

### 3.1 3D-Sensor

For the acquisition of 3D data we use a time-of-flight (TOF) range sensor, depicted in fig. 1. The working principle is similar to that of radar: electromagnetic waves are sent out and undergo reflections and other wave propagation phenomena on obstacles. In case of the considered TOF-sensor the used wavelength is different to that of radar, of course. Around the objective an LED array is arranged that emits 20 MHz-modulated light, such that it can be distinguished from background illumination. A CCD-chip is located behind the objective which basically measures the time of flight of the modulated light for a pixel array of 176 x 144 size. For each pixel, the intensity, amplitude and phase of the infrared light ray incoming at the respective angle is measured. From four phase measurements, the time of flight and therefore the distance of the remitting object can be calculated. As the infrared light is modulated at 20 MHz, we get a non-ambiguous measurement for one period, i.e. 500 nano seconds. In this time, the light travels 15 metres which defines the range of unambiguousness to be 7.5 metres for the sensor. This is usually enough for indoor applications.

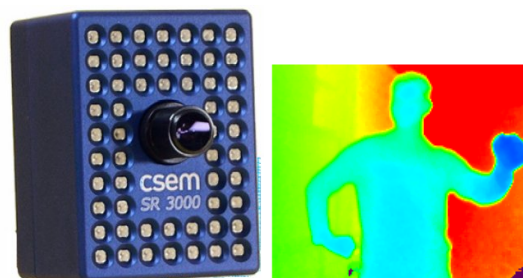


Fig. 1. Left: Time-of-flight range sensor. LEDs, propagating infrared light, are positioned around the objective, that collects them again. Right: Depth image provided by the sensor. Different colors indicate different distances of the objects from the sensor.

An example for such a depth or range image is shown in fig. 1. If we have the intrinsic parameters of the camera (focal length, pixel offset, distortions, etc.), the 3D coordinates (with respect to the camera frame) can be calculated for each pixel. Consequently a 3D point cloud can be provided in each scan. In contrast to laser scanners we get a full 3D image at one shot, not only a single 2D line. Currently the sensor is able

to acquire 3D images at a rate of up to 30 Hz in the mentioned resolution. This has the advantage that the camera usually provides a sufficient part of the scene without the necessity to move the sensor during the scan. Note that the time-of-flight measurement described above is performed not only once for one infrared ray but several times during the so-called integration time. If we increase this integration time, we are able to average the measurements and thus get less susceptible to noise on the one hand, but obtain a lower acquisition rate on the other hand. Besides the depth image the camera provides in addition an intensity image, which gives information about the intensity of the reflected light rays for each pixel. It is thus quite similar to a one layer grey value image acquired by a color camera. Fig. 2 shows both depth and intensity image of a chessboard. Note that both intensity image and depth image are acquired by the same physical CCD array, which makes the mapping between light intensity and depth information trivial for each pixel.



Fig. 2. Intensity (left) and depth (right) images of a scene containing a chess board and some little objects. The grey values mean different luminance in the intensity image and different distances in the depth image. Both are acquired by the same chip, such that we have exact and known pixel correspondences.

Intrinsic calibration of the range sensor can be performed by means of the intensity image.

### 3.2 Manipulator and camera configuration

The TOF sensor described in the section above is mounted on a 2-DOF pan-tilt unit, a PowerCube wrist module PW070. The apparatus is fixed on a table to conduct the tracking experiments (see fig. 3). In the following, we use  $\phi$  and  $\theta$  to describe the current angular values of pan and tilt, respectively. This setup is also intended for use as camera head for a service robot. Thus we have an eye-in-hand camera configuration. It is also end-point open loop, because the TCP is not tracked. The joints of the manipulator are controlled by the input of the camera data, which will be described in more detail in the next section. The joint controller interface accepts both absolute angles and angular velocities.



Fig. 3. The 2-DOF pan-tilt unit, a PowerCube PW70, with the mounted TOF-sensor. The available degrees are pan and tilt with the corresponding angles  $\phi$  and  $\theta$ .

## 4. POSITION-BASED VISUAL SERVOING

The tracking task is defined as follows: regulate the manipulator joints such that the object is situated on the optical axis of the TCP mounted camera in a certain reference distance. This states an endpoint open loop tracking system.

The following steps have to be performed:

- Extraction of features of the moving object
- Localisation of the object, i.e. getting its 3D position
- Servoing the difference of the current 3D position of the object and a target position to zero

### 4.1 Feature extraction and localization

As the focus of the paper is on demonstrating a servoing task with the range sensor, we chose a very simple object to track, a sphere of ca. 10 cm diameter, mounted on a pole (see fig. 4). The ball is perceptible as disc from the range sensor's view—the background can be separated by means of distance segmentation. After smoothing and edge filtering, the center of this circle is determined by common 2D-ellipse fitting algorithms, e.g. (Fitzgibbon *et al.*, 1999). Note that both depth and intensity image can be used for image extraction: in the depth image the sphere constitutes a disc with a certain grey value due to a respective distance, and in the intensity image we also have a disc, but due to a certain reflectivity of the sphere.

Having found the pixel coordinates of the center of the ball,  $p_b = [u_b \ v_b]^T$ , the 3D coordinate can be derived in the following way (the subscripts mean: b ball, p pixel coordinates, n normalized coordinates): Firstly, the intrinsic parameters of the camera that have been obtained by calibration are needed to determine the coordinates of the normalized plane:

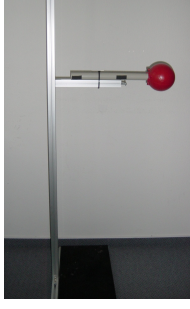


Fig. 4. A red ball on a pole is used as feature to be tracked. This set-up allows for easy distance segmentation, as the ball is the most prominent part.

$$p_{b,n} = K_{np}^{-1} \cdot p_{b,p} = \begin{bmatrix} X_b & Y_b \\ Z_b & Z_b & 1 \end{bmatrix} \quad (1)$$

with the normalized image coordinates  $p_b = [x_b \ y_b]^T$  and the intrinsic parameter matrix  $K_{np} = \begin{bmatrix} f & fs & u_0 \\ 0 & fr & v_0 \\ 0 & 0 & 1 \end{bmatrix}$ .

$p_{0,p}$  marks the principal point,  $f$  the focal length,  $r$  the aspect ratio and  $s$  the skew. As depth is measured for every pixel, we get the 3D coordinates  ${}^C P_b = [X_b \ Y_b \ Z_b]$  of the ball in the camera frame with

$${}^C P_b = Z_b \cdot p_{b,n}. \quad (2)$$

In the following the index for the camera frame will be omitted as all coordinates will be with respect to this frame. We have now determined the position of the object to be tracked with respect to the camera frame without the need for depth estimation.

#### 4.2 Tracking

As the tracking goal is to position the camera such that the object lies in the image center, i.e. on the optical axis, we can define the error

$$E = P_b - P_0 = [X_b \ Y_b \ 0]^T \quad (3)$$

with  $P_0 = [0 \ 0 \ Z_0]^T$ . The controller has to derive the necessary manipulator movements to regulate this error as stable and as fast as possible to zero. Basically this can be achieved by either position-based control or velocity-based control of the manipulator joint angles. We have implemented velocity-based control because of the better results (Baeten and de Schutter, 2004).

In the following, the general approach for position-based visual servoing is given that can be applied to any manipulator. Here we need the direct kinematics for the inverse jacobian matrix to determine the necessary angular movements to regulate the tracking error to zero. Furthermore hand-eye calibration is required to express the error with respect to the TCP frame (see fig. 5). Hand-eye calibration can be achieved, e.g. with

the calibration method of Tsai (Tsai and Lenz, 1989) or Wei et al. (Wei *et al.*, 1998). The error vector from equation 3 must contain at least as many components as degrees of freedom to be controlled, such that in general more than one feature point must be extracted. We get the following general equation:

$$\dot{\phi} = -K \cdot J^{-1} \cdot {}_C^{TCP} T \cdot {}^C E \quad (4)$$

with the proportional feedback gain matrix  $K$ , the jacobian matrix  $J$ , the hand-eye transformation  ${}_C^{TCP} T$ , the error vector with respect to the camera frame  ${}^C E$  and the joint angular velocity vector  $\dot{\phi}$ .

Because of the relative simple geometry we do not use kinematics of the manipulator, but map the found position error with respect to the camera frame to respective joint angles, as will be presented below. Furthermore the experimental results were obtained without hand-eye calibration, as the camera is mounted such that camera and TCP frame differ only by a short translation that has no effect on the angles  $\phi$  and  $\theta$ . As our set-up consists of a manipulator with merely 2 degrees of freedom, we calculated the angular velocities of the pan and tilt joint in a simpler way. In fig 5 the current and target positions of the ball with respect to the camera are depicted. The camera frame must be rotated by the angles  $\phi$  and  $\theta$ .

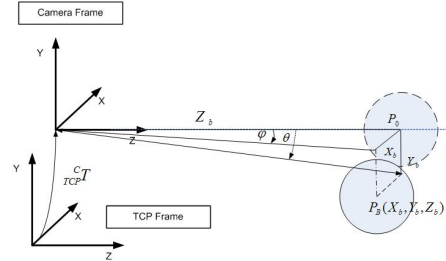


Fig. 5. Current position (solid line) and target position (dashed line) of the ball with respect to the camera frame.

From fig. 5 we obtain the angle deltas that are regulated to zero by

$$\Delta\phi = \arctan\left(\frac{X_b}{Z_b}\right) \quad (5)$$

$$\Delta\theta = \arctan\left(\frac{Y_b}{Z_b}\right) \quad (6)$$

and the corresponding angular velocities by

$$\dot{\phi} = -k_\phi \cdot \Delta\phi = -k_\phi \cdot \arctan\left(\frac{X_b}{Z_b}\right) \quad (7)$$

$$\dot{\theta} = -k_\theta \cdot \Delta\theta = -k_\theta \cdot \arctan\left(\frac{Y_b}{Z_b}\right) \quad (8)$$

with the proportional feedback gains  $k_\phi$  and  $k_\theta$ .

## 5. EXPERIMENTS

### 5.1 Set-Up

The tests have been conducted with the following set-up: the manipulator is fixed on solid ground as depicted in fig. 3. The camera is mounted on the end-effector. A sphere attached to a pole is chosen as object to be tracked. The pole is fixed by means of a stand (see fig. 4), which is placed such that the sphere still lies in the view of the camera but not on its optical axis. When the servoing is started, joint angles are recorded until the target position has been reached. The results are presented in the next sections.

### 5.2 Results

For velocity-based control the angular velocities from equ. 7 and 8 are directly passed to the joint controller at a frequency of 17.5 Hz. The reason for this frequency will be given below. Before the start of the servoing, the joint angles are in zero position, while the sphere is positioned such that  $\Delta\phi$  and  $\Delta\theta$  are at about -20 degrees. Fig. 6 shows the extracted features from the first image at the beginning of the experiment and the last image after the servoing. From the center of the detected circle in the image, the 3D position of the center of the sphere in cartesian coordinates can be calculated from equ. 2. The feature trajectory in image coordinates is shown in fig. 7.

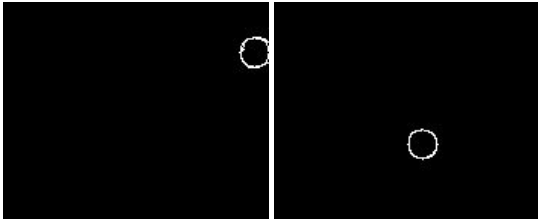


Fig. 6. Left: Detected feature at the beginning of the servoing. Right: Detected feature at the end of the servoing. Note that the images have been mirrored with respect to the x axis in order to have the image origin in the left bottom corner. This eases comparison with the image feature trajectory.

The error defined in equ. 3 is then regulated to zero. Fig. 8 shows that both  $X_b$  and  $Y_b$  converge smoothly to zero without any overshooting. As the 3D measurements of the camera underly noise, servoing is stopped when the error goes below 20 mm, which corresponds to  $\Delta\phi$  and  $\Delta\theta$  decreasing below the threshold angle of about 0.02 radians. The angular errors are depicted in fig. 9. As can be perceived in the figures, this is the case after less than a second, which corresponds to about 15 control steps. As was mentioned above, the camera may be operated at up to 30 Hz. At the same time the data acquisition time decreases, noise, however, increases, because the measurement may be

averaged over less time. Therefore we operated the camera at 18 Hz in the experiment to keep noise at an tolerable level. Together with feature extraction one complete control step lasts in average 57 milliseconds, which corresponds to 17.5 Hz. If we would need higher accuracy at a possible lower control frequency, we could just decrease the camera data acquisition rate to reduce noise.

In comparison with a stereovision approach, we have no computational effort for the generation of the 3D data.

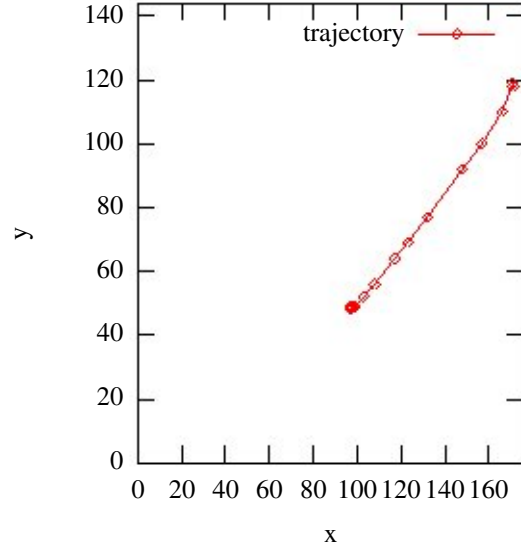


Fig. 7. Trajectory of the detected feature points in image coordinates.

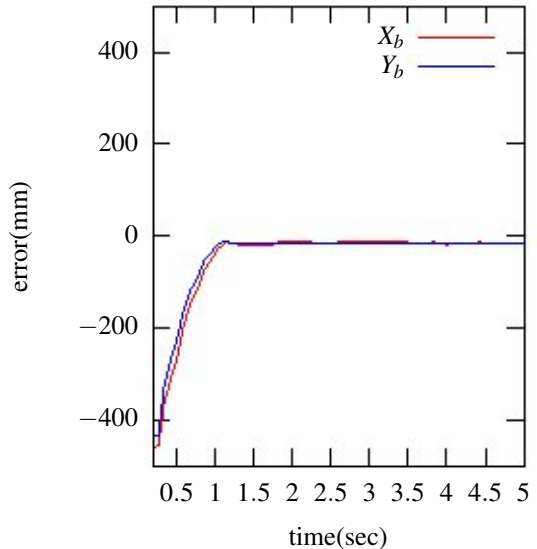


Fig. 8. Cartesian errors  $X_b$  and  $Y_b$  over time. Note that the error converges to 0 in less than 1 second.

## 6. CONCLUSION

We have demonstrated position-based servoing with eye-in-hand configuration for a tracking task using a

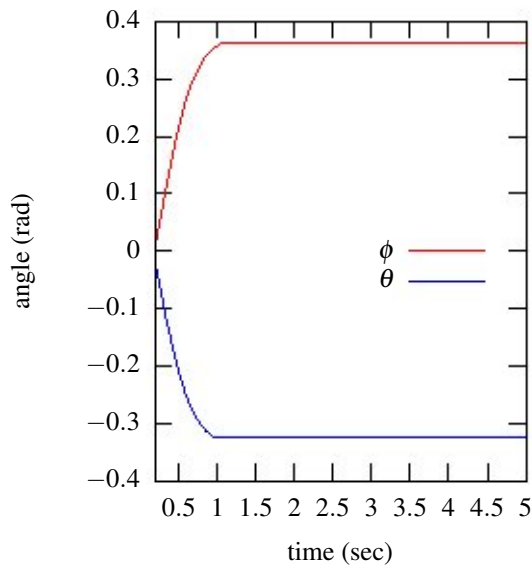


Fig. 9. Values of the joint angles during the servoing.

novel range sensor. The advantage of the approach is that depth can be acquired from the sensor as measured value and needs not be obtained by multiple views or estimated by a complex algorithm. Image features can be extracted from either depth or intensity image and the corresponding 3D coordinates calculated by means of the intrinsic camera parameters. The geometry of the tracked object needs not to be known. The approach shows good experimental results for velocity-based control. A thorough stability and accuracy analysis will be conducted in the ongoing work.

## 7. ACKNOWLEDGMENTS

This work was funded partly by the research project DESIRE by the German Federal Ministry of Education and Research (BMBF) under grant no. 01IME01.

## REFERENCES

- Agin, G.J. (1985). Calibration and use of a light stripe range sensor mounted on the hand of a robot. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. pp. 680–685.
- Allan, P.K., B. Yoshimi and A. Timcenko (1991). Real-time visual servoing. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. pp. 851–856.
- Andersson, R.L. (1987). Real Time Expert System to Control a Robot Ping-Pong Player. PhD thesis. University of Pennsylvania.
- Baeten, J. and J. de Schutter (2004). *Integrated Visual Servoing and Force Control*. Springer tracts in advanced robotics. Heidelberg.
- Chaumette, F. and S. Hutchinson (2006). Visual servo control part i: Basic approaches. *IEEE Robotics and Automation Magazine* pp. 82–90.
- Cipolla, R. and N. Hollinghurst (n.d.). Visually guided grasping in unstructured environments. *Journal of Robotics and Autonomous Systems*, 19:337–346, 1997.
- Corke, P.I. (1993). *Visual control of robot manipulators - a review*. pp. 1–31. Vol. 7 of *World Scientific Series in Robotics and Automated Systems*. World Scientific Press.
- Fitzgibbon, A., M. Pilu and R.B. Fisher (1999). Direct least square fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(5), 476–480.
- Hager, G.D., W. Chang and A.S. Morse (1995). Robot hand-eye coordination based on stereo vision. *IEEE Control Systems Magazine* 15(1), 30–39.
- Hutchinson, S., G.D. Hager and P.I. Corke (1996). A tutorial on visual servo control. *IEEE Transactions on Robotics and Automation* 12(5), 651–670.
- Malis, E., F. Chaumette and S. Boudet (2000). Multi-cameras visual servoing. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.
- Nevatia, R. (1976). Depth measurement by motion stereo. In: *Computer Graphics and Image Processing*. Vol. 5. pp. 203–214.
- Papanikolopoulos, N.P., P.K. Khosla and T. Kanade (1993). Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision. *IEEE Transactions on Robotics and Automation* 9(1), 14–35.
- Tsai, R.Y. and R.Y. Lenz (1989). A new technique for fully autonomous and efficient 3d robotics hand-eye calibration. *IEEE Transactions on Robotics and Automation* 5(3), 345–358.
- Venkatesan, S. and C. Archibald (1990). Realtime tracking in five degrees of freedom using two wrist mounted laser range finders. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. pp. 2004–2010.
- Wei, G., K. Arbter and G. Hirzinger (1998). Active self-calibration of robotic eyes and hand-eye relationships with model identification. *IEEE Transactions on Robotics and Automation* 14(1), 158–166.