Editors

Editors-in-Chief
Bernd Michel  
Fraunhofer Micro Materials Center Berlin and Chemnitz,  
Technologie-Campus 3,  
09126 Chemnitz, Germany  
e-mail: office@coinn.de

Bharat Bhushan  
Nanoprobe Laboratory for Bio- & Nanotechnology and Biomimetics (NLB2)  
The Ohio State University  
201 W. 19th Avenue  
Columbus, OH 43210-1142, USA  
e-mail: bhushan.2@osu.edu

Editorial Board

D. Acharjee, IASTM, Kolkata, India  
G.G. Adams, Northeastern University, Boston, MA, USA  
W. Benecke, Fraunhofer Institute for Silicon Technology, Itzehoe, Germany  
J. Chang, National Tsing Hua University, Hsinchu, Taiwan  
N.F. de Rooij, Université de Neuchâtel, Neuchâtel, Switzerland  
M. Esashi, Tohoku University, Sendai, Japan  
J. Felba, Wroclaw University of Technology, Wroclaw, Poland  
T. Fukuda, Nagoya University, Nagoya, Japan  
L. Gu, Massachusetts Institute of Technology, Cambridge, MA, USA  
G. Harsanyi, Budapest University of Technology and Economics, Budapest, Hungary  
R.T. Howe, Stanford University, Stanford, CA, USA  
J. Iannacci, Fondazione Bruno Kessler-FBK, Trento, Italy  
G. Jang, Hanyang University, Seoul, Korea  
Jia-Yang Juang, National Taiwan University, Taipei, Taiwan  
W. Kanert, Infineon Technologies AG, Neubiberg, Germany  
A. Khosla, Lab 177, Ontario, Canada  
K.-D. Lang, Fraunhofer Institute for Reliability and Microintegration IZM, Berlin, Germany  
R.A. Lawes, FCGI FIEE FInstP FREng, Micro-Nanotechnology, Oxfordshire, UK  
W. G. Lee, Department of Mechanical Engineering, Kyung Hee University, Yongin, Korea  
J. Lu, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan  
J. Mehner, Chemnitz University of Technology, Department of Microsystems and Precision Engineering, Chemnitz, Germany  
W. J. Meng, Lousiana State University, Baton Rouge, LA, USA  
N. Meyendorf, Fraunhofer IKTS, Dresden und Berlin, Germany  
M. Michel, IBM Deutschland Research, Boeblingen, Germany  
Th. Otto, Fraunhofer ENAS Chemnitz, Germany  
H.M. Ouakad, Sultan Qaboos University, Muscat, Oman  
N.C. Park, Yonsei University, Seoul, Korea  
Y. Ping, Jiangsu University, Zhenjiang, 212013, P.R.China
Eye-to-hand robotic tracking and grabbing based on binocular vision

Yi-Chun Du, Taryudi Taryudi, Ching-Tang Tsai & Ming-Shyan Wang
Your article is protected by copyright and all rights are held exclusively by Springer-Verlag GmbH Germany, part of Springer Nature. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer’s website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".
Eye-to-hand robotic tracking and grabbing based on binocular vision

Yi-Chun Du1 · Taryudi Taryudi2 · Ching-Tang Tsai1 · Ming-Shyan Wang1

Received: 25 February 2019 / Accepted: 6 May 2019 © Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

This paper aims to study eye-to-hand robotic tracking and grabbing based on binocular vision. Two cameras placed on different locations gave different three-dimensional coordinates of the object from a binocular vision. The robot then analyzed the dynamic vision acquired from the stereo cameras and tracked the object in three-dimensional space utilizing continuously adaptive mean shift (CAMSHIFT) algorithm. Afterward, inverse and forward kinematics were implemented to move the robotic arm to the appropriate position so that it can grab the object using the end effector. The inverse kinematics was analyzed by a geometric algorithm to reduce the computational burden. Consequently, the experimental results verified that the eye-to-hand system was able to track and grasp the target successfully.

Abbreviations

\( f \) Focal length of the camera
\( s \) Scale factor in image
\( f_x, f_y \) Focal length in X and Y directions of the image plane
\( b \) Distance between two cameras
\( d \) Disparity
\( P \) The coordinates of the end effector
\( P_i = (x_i, y_i) \) An image plane point
\( c_x, c_y \) Center of the image
\( P_{i+1} \) Homogeneous transformation matrix
\( n \) Link length
\( u, v \) Image plane coordinate
\( d_i \) Link distance
\( \theta_n \) Joint angle
\( \alpha_i \) Link twist
\( (x_c, y_c) \) Center of mass
\( M(x) \) Mean shift value of the vector \( x \)
\( M_{00} \) Zeroth-order momentum
\( M_{10}, M_{01} \) First-order momentum
\( I(x, y) \) Color histogram value
\( X'--Y'--Z' \) Wrist coordinate system
\( X--Y--Z \) Base coordinate system
\( X''--Y''--Z'' \) End effector coordinate frame with respect to base frame
\( P(X_p, Y_p, Z_p) \) World coordinate
\( R \) An orthonormal rotation matrix (3 by 3)
\( T \) Translation matrix (3 by 1)

1 Introduction

The computer vision issue is divided into object recognition and target tracking. Binocular vision is based on the fact that the left and right cameras are placed on the same plane with the center of the lens under the same baseline. Binocular stereo vision is different from planar vision in that it uses two digital cameras to calculate the depth of field information given by the object through pinhole imaging. The depth of the target object is the distance from the camera to the object and can be measured from the pair of left and right images.

Optical flow is used to identify objects (Kale et al. 2015) and provide information about its motion even when there is no quantization parameter to calculate. The motion vector estimation method gives the positions of the object in the continuous pictures to help identify it. In one research, a Gaussian mixture model algorithm was adopted to identify infants (Aljuaid and Mohamad 2013), wherein

1 Department of Electrical Engineering, Southern Taiwan University of Science and Technology, Tainan, Taiwan
2 Faculty of Engineering, Universitas Negeri Jakarta, Jakarta, Indonesia
the characteristics of the baby’s head, body, and feet were used to identify and track the current state of the baby in time. Furthermore, a visual attention model and an extended Kalman filter (Zhou 2014) were proposed for unmanned aerial vehicle (UAV) filming. The algorithm was able to quickly capture and recognize object images and track them, while the filter was adept at removing the noise of the object images to make the identification and tracking more stable. Another system studied by Lee et al. was able to instantly identify the target in the entire work area and provide it with a three-dimensional model (Lee et al. 2005). In the study made by Hung and Li, the scale-invariant feature transform (SIFT) method was used for recognizing objects in complex environments (Hung and Li 2007). Its visual system was established for three purposes: identifying objects, tracking targets, overcoming complex environments, and finally, completing the grabbing task through a 6-DOF robotic arm.

The mean shift method is a color tracking algorithm that refers to the mean shift value (average partial migration) of the current target which is a set of vectors. It is an iterative local optimization approach which aims to find the peak of a probability function by climbing in the positive direction of the density gradient. It has been applied to object tracking problems and has produced many impressive results in several previous pieces of research. However, one drawback of the mean shift algorithm is that it does not give a good estimation of the bandwidth change. In the study made by Cheng, a set of functions with different weight distributions along the distance was added as an improved method to make the weight near the center greater (Cheng 1995). The continuously adaptive mean shift (CAMSHIFT) tracking algorithm is an improved algorithm because of its continuous adjustment. The CAMSHIFT is mainly composed of HSV (hue, saturation, value) color space, back projection, mean shift, and the region of interest (ROI) of the changeable search area. In Hu et al., a CAMSHIFT based tracking method was used to extract the player trajectories from broadcast sports videos. These were then mapped to the real-world court coordinates according to the camera calibration (Hu et al. 2011). Xiu and Ba proposed an improved CAMSHIFT tracking method based on a multi-feature fusion that included a gauss weight function to select the ROI, intercept the ROI from the background, and execute the back projection mapping of the target area for independent tracking, thus eliminating the interference of the background to the object (Xiu and Ba 2016). The CAMSHIFT method and the Haar-like features combined with a Kalman filter were proposed to cope with the problem in object tracking due to skin-color disturbances and missed tracking due to occlusion (Yu et al. 2016).

A camera and a laser range finder with a background subtraction technique are usually used for object tracking (Long and Yang 1990; Itoh et al. 2006). The robotic arm grabs the target after identifying the object. The general derivation of kinematics for the robot is divided into the following two methods. The first method uses the relative relationship between the reference coordinates and the end effector. Then, the inverse kinematics is derived by multiplying the inverse matrices (Long and Yang 1990). The second method is a relatively simplified derivation process obtained by fixing the position of the end effector, the relative direction of the end effector, and the reference coordinate. This method minimizes complexity and offers a unique solution because both the end effector and the reference coordinate maintain a fixed direction in the derivation process.

2 Stereoscopic analysis

Binocular vision is based on the fact that the left and right cameras are placed on the same plane with the center of the lens under the same baseline. Its main advantage is that the depth of the target object, i.e., the distance from the camera to the target object, can be measured from the pair of images. In Fig. 1 (Taryudi and Wang 2017), the world coordinates \( P(X_p, Y_p, Z_p) \) were assumed to be the coordinates of the target object and were respectively projected on the left eye image and the right eye image to form points \( P_1 = (x_1, y_1) \) and \( P_2 = (x_2, y_2) \), wherein the distance between the centers of the binocular camera is \( b \) and the camera focal length is \( f \).

The following relative equations were obtained from the relationship of similar triangles in Fig. 2 (Taryudi and Wang 2017).

\[
\frac{X}{x_1} = \frac{Z}{f} \tag{1}
\]

\[
\frac{Y}{y_1} = \frac{Z}{f} \tag{2}
\]

\[
\frac{b - X}{x_2} = \frac{Z}{f} \tag{3}
\]

The definition of \( b \) leads to the equation

\[
b = \frac{Z}{f} x_1 + \frac{Z}{f} x_2 \tag{4}
\]

In addition, the binocular vision has the same \( Z \) coordinate values,

\[
Z = \frac{b \cdot f}{x_1 + x_2} \tag{5}
\]

The disparity \( d = x_1 + x_2 \) is the difference in the values of the \( x \) coordinate in Figs. 1 and 2.
The object imaging principle of the camera can be described by a three-dimensional coordinate matrix composed of the internal parameter matrix of the camera itself and the external parameter matrix. The conversion equation is as follows (Opencv 2019):

\[
\begin{bmatrix}
    f_x & 0 & c_x & \vline & r_{11} & r_{12} & r_{13} & t_1 \\
    0 & f_y & c_y & \vline & r_{21} & r_{22} & r_{23} & t_2 \\
    0 & 0 & 1 & \vline & r_{31} & r_{32} & r_{33} & t_3
\end{bmatrix}
\]

where \( s \) is the scale factor, and \( u \) and \( v \) stand for the pixel coordinates of the object projected onto the camera. \( \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \end{bmatrix} \) is the internal parameter matrix of the camera. \( f_x \) and \( f_y \) are the focal lengths in the X and Y directions of the image plane. \( c_x \) and \( c_y \) are the reference points which are ideally the center of the image. \( \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \) is the external parameter matrix of the camera, where \( R \) is expressed as the angular amount of the rotation of the X, Y, and Z axes. \( T \) is the column vector for the translation of the object coordinates to the XYZ coordinates of the camera. Finally \((X, Y, Z)\) represent the position of the object in the world coordinate system. The distance between the object and the camera can be obtained by matching the parameters obtained above.

3 Color tracking algorithm

For \( N \) sample points \( x = [x_i], i = 1, \ldots, N \) in a given d-dimension space \( R^d \), the mean shift value of the vector \( x \) is defined as

\[
Ms(x) = \frac{1}{k} \sum_{i=k} (x_i - x)
\]

There are \( k \) group sample points in the rectangular space \( S \) in Eq. (8). Furthermore, \( Ms(x) \) is equal to the sum of the offsets of the \( k \) group sample points divided by \( k \), i.e. the mean value. This algorithm can be regarded as an automatically adjusted gradient rising peak search method which aims to converge at the peak of the rising peak.
Figures 3, 4 and 5 show the implementation of the mean shift algorithm (Tsai 2005). The blue circle area in Fig. 3 is the region of interest (ROI); the red hollow point is the centroid; and the yellow arrow is the mean shift vector. The new mean shift vector shown in Fig. 4 was obtained by taking the above-mentioned mean shift vector end point as the center and substituting it into Eq. (8). After repeating the above steps, the mean shift algorithm converged to the place of the largest probability density. The final result is shown in Fig. 5. In the research made by Fukunaga and Hostetler, the color probability was added to the mean shift algorithm by inputting a set of colors, the location of the highest and most stable convergence point, as well as the concentrated area of this color distribution to the mean shift (Fukunaga and Hostetler 1975). This is applied to image tracking.

The mean shift algorithm searches the target object mainly by means of a probability density distribution; therefore, the color probability distribution function is needed to assist the analysis. In this paper, the back projection histogram was used to calculate the color probability distribution. The back projection histogram was used to circle the area in the picture, and this area is called the ROI. It first converted the circled color format to HSV, which reduced the amount of calculation and also improved the influence of light. Then, the color probability distribution for the circled area was calculated. The graph was added up for those with the same color, and the accumulated color in the histogram became the peak basis that the mean shift algorithm pursued.

The conversion from RGB to HSV is given by

\[
H = \begin{cases} H1, & \text{if } B \leq G \\ 360^\circ - H1, & \text{if } B > G \end{cases}
\]

\[
H1 = \cos^{-1}\left(\frac{0.5[(R - G) + (R - B)]}{\sqrt{(R - G)^2 + (R - B)(G - B)}}\right)
\]

\[
S = \frac{\text{Max}(R, G, B) - \text{Min}(R, G, B)}{\text{Max}(R, G, B)}
\]

\[
V = \frac{\text{Max}(R, G, B)}{255}
\]

The disadvantage of the mean shift algorithm is that its ROI is a fixed value. Therefore, when the target object approaches the lens, the object in the image becomes larger, and the effect of the fixed ROI is small. However, when the target object is far away from the lens, the object in the image becomes smaller. The proportion of the smaller objects in the ROI makes tracking unstable and causes misjudgment. The CAMSHIFT tracking algorithm modified from mean shift has the ability to continuously adjust the ROI values as the target becomes larger or smaller in the image to achieve the appropriate ROI to complete the continuous tracking of the object.

Figure 6 shows the CAMSHIFT flowchart. The detailed steps are described as follows:

1. Load the image and encircle the region of interest.
2. Convert the Hue image in the ROI to a color histogram and set the ROI to the initial value of the mean shift window.
3. Calculate the color range ratio of the color histogram.
4. Calculate the center of mass \((x_c, y_c)\) for the specified color in the mean shift window.
5. Move the center of the mean shift window to the center of the color probability density distribution to update the color probability density distribution of the ROI.
6. Repeat steps 4 and 5 until the mean shift window converges steadily.
7. When the convergence is reached, adjust the size of the mean shift window to obtain the appropriate ROI. In the study conducted by Bradski, the ellipse was used to find the best ellipse in the mean shift window that can cover the object area (Bradski 1998). The projections of the long and short axes on the X- and Y-axis are the result of the adjustment.
8. Record the best elliptical center position, size, and other information in the mean shift window to provide the next image initialization.
9. Load the next image and repeat steps 2 through 8 with the updated ROI.

The center of mass \((x_c, y_c)\) is calculated as

\[
x_c = \frac{M_{10}}{M_{00}}; \quad y_c = \frac{M_{01}}{M_{00}}
\]  

where \(M_{00} = \sum_x \sum_y I(x,y)\) is the zeroth-order momentum, and \(M_{10} = \sum_x \sum_y xI(x,y)\) and \(M_{01} = \sum_x \sum_y yI(x,y)\) are the first-order momentums, respectively.

\[
M_{00} = \sum_x \sum_y I(x,y) \quad (14)
\]

\[
M_{10} = \sum_x \sum_y xI(x,y) \quad (15)
\]

\[
M_{01} = \sum_x \sum_y yI(x,y) \quad (16)
\]

4 Motion control of the robotic arm

Forward kinematics refers to the use of the angles of rotation of the axes of the robotic arm in deriving the position of the working point in the three-dimensional space and the end effector attitude vector. Inverse kinematics calculates the rotation angle of each axis by using the working point coordinates in the three-dimensional space and the end effector attitude vector. The linkage coordinate system of the robotic arm is defined according to the Denavit–Hartenberg (D–H) rule. A 6-DOF robotic arm RV-3SD (Mitsubishi 2010) was used in the paper. Table 1 and Fig. 7 were used in the derivation of the forward kinematics of the robot to find the absolute coordinate position of the end effector coordinate system relative to the robotic arm after completion of the movement, i.e., the operation shown in Eq. (17), where \(P\) is the coordinates of the end effector (Tsai 2005).

\[
P = \begin{bmatrix} u_x & v_x & w_x & d_x \\ u_y & v_y & w_y & d_y \\ u_z & v_z & w_z & d_z \end{bmatrix}
\]  

\[(17)\]

where

\[
\begin{bmatrix} c\theta_1 & 0 & -s\theta_1 & A_1c\theta_1 \\ s\theta_1 & 0 & c\theta_1 & A_1s\theta_1 \\ 0 & -1 & 0 & d_1 \\ 0 & 0 & 0 & 1 \end{bmatrix}
\]

\[
\begin{bmatrix} c\theta_2 & -s\theta_2 & 0 & A_2c\theta_2 \\ s\theta_2 & c\theta_2 & 0 & A_2s\theta_2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
\]  

\[(18)\]

\[
\begin{bmatrix} c\theta_3 & 0 & s\theta_3 & A_3c\theta_3 \\ s\theta_3 & 0 & -c\theta_3 & A_3s\theta_3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
\]
Since the robotic arm reaches a certain position in space, the rotation angle of each axis on the arm is required. Inverse kinematics gives the position of the end effector of the robotic arm and estimates the angle of each linkage required for the motion (Jiang and Huang 2006). Figure 8 shows the situation where the end effector is mounted on the arm. X–Y–Z is the base coordinate system. The X₀–Y₀–Z₀ coordinate is defined as the wrist coordinate system in which Z₀ is in the DE direction, X₀ is perpendicular to the ground, and E is the origin of the wrist coordinate system. The mutually perpendicular X″, Y″, and Z″ vectors define the local coordinate system of the end effector relative to the base coordinates and are also called the pose vectors, where Z″ represents the direction of the end effector and X″ shows the direction of the end effector. The end effector is mounted in the direction of EP, and X₀ or Y₀ shows the rotating offsets of the end effector. The respective arm lengths are represented by OA = L₁₀, AB = L₁₁, BC = L₂, CD = L₃₀, and DE = L₃₀. Figure 9 shows the point with coordinates P(px, py, pz) that can be projected onto the XY plane as

\[ \theta_1 = \tan^{-1}(py/px) - \sin^{-1}(Ly/L) \approx \tan^{-1}(py/px) \tag{21} \]

where L = \sqrt{px^2 + py^2} and generally Ly << L. Through trigonometric functions, Fig. 10 shows the angles θ₂ and θ₃ as

Since the robotic arm reaches a certain position in space, the rotation angle of each axis on the arm is required. Inverse kinematics gives the position of the end effector of the robotic arm.
\[ \theta_2 = \begin{cases} \pi/2 - (\angle CBE + \angle EBC), & \text{if elbow up} \\ -\theta_2, & \text{if elbow down} \end{cases} \]

\[ \theta_3 = \pi - \angle BCE + \angle E \]  

where

\[ \angle CBE = \cos^{-1} \left( \frac{r_1^2 + P^2 + Q^2 - r_2^2}{2} \times \frac{\sqrt{P^2 + Q^2}}{r_1} \right) \]  

\[ \angle EBC = \tan^{-1}(Q/P) \]  

\[ \angle BCE = \cos^{-1} \left( \frac{r_1^2 + r_2^2 - (P^2 + Q^2)}{2} \times \frac{r_2}{r_1} \right) \]  

\[ \angle E = \tan^{-1}(L_30/L_31) \]

\[ r_2 = L_3 = \sqrt{L_{30}^2 + L_{31}^2} \]

\[ r_1 = L_2 \]  

r1 and r2 are the distance between the second and third axes and the distance between the third axis and the fifth axis, respectively.

In Fig. 11, \( X''_w, Y''_w, \) and \( Z''_w \) are the pose vectors seen in the wrist coordinates. The coordinate rotation can be used to convert \( X'', Y'', \) and \( Z'' \) into the posture vectors.

Fig. 9 Determination of \( \theta_1 \)

Fig. 10 Determination of \( \theta_2 \) and \( \theta_3 \)

Fig. 11 Determination of \( \theta_4 \) and \( \theta_5 \)

Fig. 12 a, b Situations of \( \theta_5 \)

Fig. 13 The target of a blue glue bottle cap
$X'_w, Y'_w, \text{ and } Z'_w$ in the wrist coordinates ($X'–Y'–Z'$). $Z''_w$ is in the same direction as the $EP$ vector. For $Z''_w = (u, v, w)$, 
\[ \theta_4 = \tan^{-1}(v/u) \]

Referring Fig. 12a, b, the angle of $\theta_5$ can be determined by

\[ \theta_5 = \begin{cases} 
-\theta_{5-1}, & \text{if } \theta_{5-2} > \pi/2 \\
\theta_{5-1}, & \text{if } \theta_{5-2} \leq \pi/2 
\end{cases} \]

where

\[ \theta_{5-1} = \cos^{-1}((Z''_w) \cdot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T/||Z''_w||) \]
\[ \theta_{5-2} = \cos^{-1}(\text{Rotz}(\theta_4) \ast \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T \cdot (Z''_w)/||Z''_w||) \]

The sixth axis controls the face of the end effector, so its angle is a fixed value.
5 Experimental results

The robotic arm used in this paper was the RV-3SD from Mitsubishi Electric Corporation, and the two cameras were the Logitech HD network camera C310. The 3-finger adaptive robot gripper (Robotiq 2015) shown in Fig. 13 was adopted in the experiment due to its versatility and flexibility. It could pick up objects of various shapes.

After obtaining the plane coordinates of the target object through the cameras, the user interface analyzed the information to find the depth information of the target object and the tracking frame for tracking. The coordinate of the object based on the camera was converted into the position of the object with respect to the robotic arm and then transmitted to the robotic arm for tracking. When asked to decide whether to end tracking, if the answer was yes, the tracking was stopped and the robot arm was commanded to grasp the body. If no, the screen and the robotic arm continued to track the object.

The camera was calibrated by using an image that can be used as a reference which is a $9 \times 7$ checkerboard pattern with a grid size of $40 \times 40$ (mm$^2$), as shown in Fig. 14. The left and right cameras were able to simultaneously capture the complete checkerboard and adjust its direction when placed in front of the cameras. The
resolution of the image frame was 1280 × 720 pixels. The internal parameters of the camera could be obtained by calculating the image via the StereoCalibrate function (OpenCV 2019). The internal parameters of the left and right cameras are as follows,

\[
\begin{bmatrix}
808.4347 & 0 & 309.2988 \\
0 & 808.5470 & 256.7041 \\
0 & 0 & 1 \\
806.9295 & 0 & 311.0859 \\
0 & 806.5466 & 258.0227 \\
0 & 0 & 1
\end{bmatrix}
\]  \tag{35}

The distance measurement of the binocular vision target mainly measured the distance of the target at different positions and verified this data with the actual distance. In the experimental method, a glue bottle cap was placed on a marked point on the table. The distance between each point was 4 cm. Figure 15 shows the distance data measured at each position. The measured data shown in Fig. 15 are roughly the same as the actual distances, however, errors occurred because light and shadow changes at different distances affected the ROI in the CAMSHIFT, resulting in the offset of the center point.

The target, a blue glue bottle cap, is shown in Fig. 13. The color histogram in Fig. 16 and the back projection histogram in Fig. 17 were obtained after encircling the ROI and calculating the color histogram. The horizontal axis in Fig. 16 represents the pixel value of its color, and the vertical axis stands for the total cumulative number of pixels of each pixel value in the frame (ROI). The tracking process generated by the above-mentioned CAMSHIFT tracking algorithm is shown in Fig. 18. Figure 19 displays the tracking process of the green glue bottle cap.

After encircling the object in the screen and marking the tracking frame, the three-dimensional coordinates of the object in space were obtained through stereoscopic analysis. These coordinates were transmitted to the robotic arm for tracking and grasping. Figure 20 shows the entire object grabbing tracking process. In Fig. 20a, the initial
motion for the tracked object was selected, then, the robotic arm moved to the initial object position in Fig. 20b. Figure 20c–e depict the process of object tracking by the robotic arm. Figure 20f–h show the robotic arm confirming the object, grabbing it downwards, and lifting it. Based on Table 1, the related parameters are listed in Table 2. In addition, the corresponding joint angles of the robotic arm in Fig. 20 are given in Table 3. The differences between the actual and real angle values were found to be very small.

In this experiment, three processes were pursued to achieve object tracking and grabbing, image capture, as well as CAMSHIFT algorithm and kinematics for tracking and grabbing. In image capture using stereo camera, the time required was only 0.001546 s. Furthermore, the execution time of the CAMSHIFT algorithm and kinematics was less than 1 s. As a result, 1-Hz sampling rate was used in the experiment.

Fig. 20 The entire tracking process to grab the object

<table>
<thead>
<tr>
<th>Table 2 Robot parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>L10</td>
</tr>
<tr>
<td>L11</td>
</tr>
</tbody>
</table>


**Table 3** Joint angles of the robotic arm in Fig. 19

<table>
<thead>
<tr>
<th>Joint angle</th>
<th>Figure 19</th>
<th>(a)</th>
<th>(d)</th>
<th>(f)</th>
<th>(h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual</td>
<td>0°</td>
<td>$-79^\circ$</td>
<td>$-62.75^\circ$</td>
<td>$-62.75^\circ$</td>
<td></td>
</tr>
<tr>
<td>Real</td>
<td>0°</td>
<td>$-78.99^\circ$</td>
<td>$-62.74^\circ$</td>
<td>$-62.74^\circ$</td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual</td>
<td>$-0.25^\circ$</td>
<td>6.55°</td>
<td>60.40°</td>
<td>55.68°</td>
<td></td>
</tr>
<tr>
<td>Real</td>
<td>$-0.182^\circ$</td>
<td>6.549°</td>
<td>60.40°</td>
<td>55.67°</td>
<td></td>
</tr>
<tr>
<td>$\theta_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual</td>
<td>128.59°</td>
<td>143.85°</td>
<td>75.66°</td>
<td>72.89°</td>
<td></td>
</tr>
<tr>
<td>Real</td>
<td>128.23°</td>
<td>143.86°</td>
<td>75.65°</td>
<td>72.92°</td>
<td></td>
</tr>
<tr>
<td>$\theta_5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual</td>
<td>51.08°</td>
<td>29.66°</td>
<td>44°</td>
<td>51.47°</td>
<td></td>
</tr>
<tr>
<td>Real</td>
<td>51.86°</td>
<td>29.63°</td>
<td>43.95°</td>
<td>51.37°</td>
<td></td>
</tr>
</tbody>
</table>

**6 Conclusions**

In this paper, the proposed system obtained the position of the target object and tracked it by using two cameras, the binocular vision principle, and the CAMSHIFT tracking algorithm. After the final position of the target was transmitted to the robotic arm and converted into the values of the corresponding coordinate system, the robotic arm was able to successfully move and grab the target using an end effector.

**Acknowledgements** The authors would like to express their appreciation for the financial support bestowed by the following: Ministry of Science and Technology, Taiwan, contract no. MOST 107-2622-E-218-008-CC2 and Ministry of Education, Higher Education Sprout Project.

**References**


Mitsubishi (2010) RV-3SD/3SDJ/3SDB/3SDBJ Series standard specifications manual


**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.