Application of Tracking-Learning-Detection for Object Tracking in Stereoscopic Images

Michal Puheim, Marek Bundzel, Peter Sinčák, and Ladislav Madarász

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering
and Informatics, Technical University of Košice, Slovak Republic
{michal.puheim,marek.bundzel,peter.sincak,
ladislav.madarasz}@tuke.sk

Abstract. We use Tracking-Learning-Detection algorithm (TLD) [1]-[3] to localize and track objects in images sensed simultaneously by two parallel cameras in order to determine 3D coordinates of the tracked object. TLD method was chosen for its state-of-art performance and high robustness. TLD stores the object to be tracked as a set of 2D grayscale images that is incrementally built. We have implemented the 3D tracking system into a PC, communicating with the Nao humanoid robot [4][5] equipped with a stereo camera head. Experiments evaluating the accuracy of the 3D tracking system are presented. The robot uses feed-forward control to touch the tracked object. The controller is an artificial neural network trained using the error Back-Propagation algorithm. Experiments evaluating the success rate of the robot touching the object are presented.

Keywords: Tracking-Learning-Detection, TLD, Nao robot, object tracking, stereo-vision, neural network controller.

1 Introduction

Our goal was to develop a robust and reasonably fast (10 fps) system applicable in mobile robotics capable to track real world objects and to determine their 3D coordinates. The approach we have chosen uses two 2D tracking systems running parallel on the images sensed by a stereovision head of a Nao humanoid robot. The 3D tracking system may be used to touch or grasp objects by the robot or to plan actions based on the spatial distribution of the obstacles in the robot's world, etc.

We have applied Tracking-Learning-Detection algorithm (TLD [1]-[3]) to track objects in 2D images sensed by the stereo-vision of the Nao robot [4][5]. Using the information about object positions on the frames taken by the cameras of the stereovision system simultaneously and the parameters of the stereovision system, 3D coordinates of the tracked object are determined. The proposed system was tested in two sets of experiments. Firstly, the accuracy of the distance measurements was evaluated. Secondly, we tested the combination of the proposed system and a robot's hand controller based on the neural network with the goal to move the robot's hand in order to touch the tracked object.

P. Sinčák et al. (eds.), *Emergent Trends in Robotics and Intelligent Systems*,
Advances in Intelligent Systems and Computing 316, DOI: 10.1007/978-3-319-10783-7_35

2 Determining Depth in Stereovision

Let us assume having two identical cameras installed so that their optical axes are parallel as shown in Fig. 1. Let f be the focal distance of the cameras, c the distance between the cameras, and a the distance of cameras from the central optical axis and c = 2a. Let XR be the x-coordinate of the object as seen by the right camera and XL be the x-coordinate of the object as seen by left camera. Using the similarity of triangles we get the following equations:

$$\frac{X_L}{f} = \frac{(x-a)}{z} \quad (1)$$

$$\frac{X_R}{f} = \frac{(x+a)}{z} \quad (2)$$

By merging these equations and the elimination of x we get the formula which can be used to calculate the z coordinate:

$$z = \frac{-2af}{X_L - X_R} \quad (3)$$

This equation enables us to use the difference in projections of the same object to determine the distance of the object from the image plane (i.e. the depth information) assuming that the projections of the object are recognized and localized. This is called “the correspondence problem”. We use the TLD method to solve the correspondence problem so that only the projections of the tracked object are searched for. We do not construct the depth map.

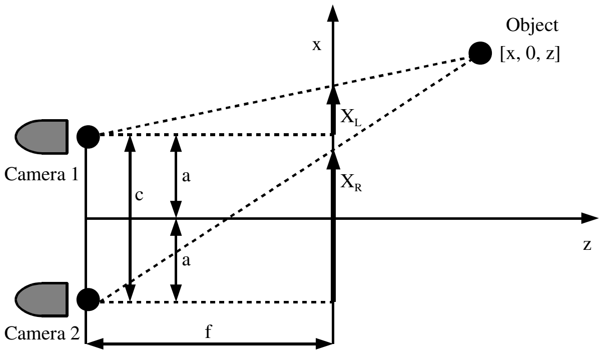

Fig. 1. Estimation of object *z* coordinates using stereo cameras with parallel optical axis. Distance between cameras is $d = 2a$, where a is the distance of the camera from the optical axis. X_R and X_L are coordinates of the object as seen by right and left camera respectively. If f is focal length then z coordinate can be calculated using X_R and X_L.

3 Tracking-Learning-Detection

Tracking-Learning-Detection (TLD) method [1]-[3] is designed for long-term tracking of arbitrary objects in unconstrained environments. One of the advantages of the system is that it does not need to separate an offline learning stage. To initialize the tracking, the target object is delimited by a bounding box in the initial image. Further learning of the alternative appearances of the object is performed during the run-time, i.e. the longer the algorithm runs, the better it should be able to recognize the target object. TLD system is composed of three basic components:

- Tracker – a short term tracker based on the Lucas-Kanade method [6], which is used to track the given object and to generate examples for the learning of the detector.
- Detector – has the form of a randomized forest [7], enables incremental updates of its decision boundary and real-time sequential evaluations during run-time [2]. Runs independently from the tracker.
- Learning algorithm – so called “P-N Learning” [1] which uses trackers to generate positive (P) and also negative (N) examples that are further used in order to improve the model of the detector.

The object is supposed to be tracked by a tracker component and simultaneously learned in order to build a detector that is able to re-detect the object once the tracker fails. The detector is built upon the information from the first frame as well as the information provided by the tracker. Both components make errors. The stability of the system is achieved by mutual cancellation of these errors. The learned detector enables reinitialization of the tracker whenever a previously observed appearance reoccurs [2].

4 3D Tracking System

The proposed system applies the TLD method on the stereo images sensed by the Nao robot. The robot is able to track an arbitrary selected object and also determines the position of the target object in the three-dimensional space, see Fig. 2.

The operator delimits the target object manually by selecting its bounding box. The initialized TLD is duplicated so that the images from the stereovision camera are processed individually. We have implemented a method to synchronize the object models of the two TLD systems in order to prevent excessive difference between their object models.

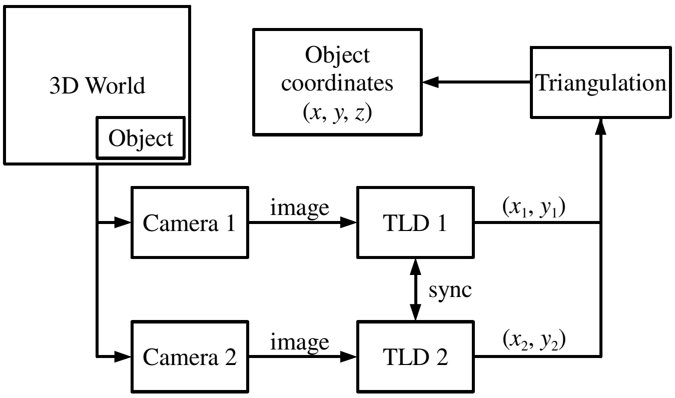

Fig. 2. The simplified block diagram of the proposed stereo-vision system employing the TLD method for the object tracking. The target object is simultaneously captured by two cameras. The images are used as the input to the two synchronized TLD systems producing the 2D coordinates of the target object in both images. The pair of 2D coordinates are used to calculate 3D coordinates of the target object using the triangulation method.

Each of the parallel TLD systems produces four outputs defining the bounding box of the object on the camera image, see Fig. 3.

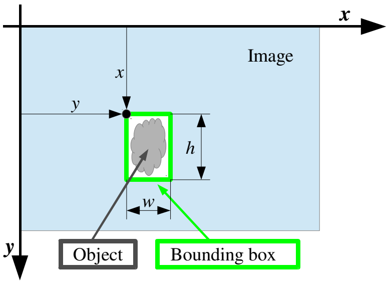

Fig. 3. Output of the TLD system is the bounding box of the tracked object. This bounding box is defined by four values (x, y, w, h) which are horizontal and vertical coordinates of the upper left corner of the box, width and height of the box.

Using these values we can calculate the centres (xc, yc) of the bounding boxes for both TLD systems. Let (xL, yL) and (xR, yR) be the centre coordinates of the object on the image captured on the left and right camera respectively, given in pixels. The 3D coordinates of the target object are calculated as:

$$z = \frac{cf}{x_L - x_R} \tag{4}$$

$$x = \frac{c}{2} + z\frac{\left(x_L - \frac{w_i}{2}\right)}{f} \tag{5}$$

$$y = -z\frac{\left(y_L - \frac{h_i}{2}\right)}{f} \tag{6}$$

where wi is the image width (in pixels), hi is the image height (in pixels), f is the focal distance of the cameras (in pixels), and c is the distance between cameras (in meters). The distance d of the target object from the cameras is the length of the vector (x, y, z):

$$d = \sqrt{x^2 + y^2 + z^2}\ . \tag{7}$$

5 Robot Arm Controller

The task to test the applicability of the proposed 3D tracking system was to touch the target object with the humanoid's hand. We have implemented a neural network-based feed-forward controller of the robots arm. The polar 3D coordinates of the target object and orientation of the robots head (i.e. the neck joints angles) represent the controllers inputs, see Fig. 4.

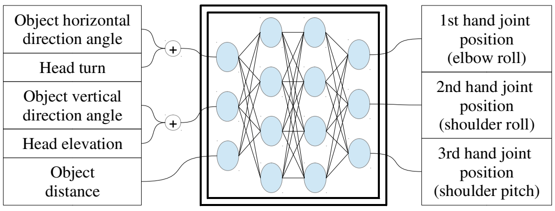
Object horizontal direction angle
Head turn
Object vertical direction angle
Head elevation
Object distance
+
+
1st hand joint position (elbow roll)
2nd hand joint position (shoulder roll)
3rd hand joint position (shoulder pitch)

Fig. 4. Controller of the arm motion of the robot based on the feed-forward neural network. Inputs are the information about head turn and elevation and the information about the object position obtained from the stereo-vision TLD system. Outputs are the positions of the hand joints which can be used in order to move the hand to the object location.

The synaptic weights are optimized by the error Back-Propagation algorithm. Training data is generated by setting the 3D tracking system to follow the hand of the humanoid in order to create enough samples of corresponding input and output data. The neural network represents a transformation of the measured coordinates of the target object and the robot’s arm joints angles. Feedback control is to be implemented to improve performance.

6 Experiments

To test the applicability of the proposed system we have performed two sets of experiments. The first evaluates the accuracy of the object distance measurement. In this experiment we place the target object 20 times in front of the robot in a variable distance ranging from 0.4 to 2.0 meters. At each measurement we compared the distance computed by the stereo-vision system with the distance measured by physical means. Results are shown in Table 1.

In the second experiment we addressed the testing of the proposed hand controller in order to determine the ability to touch the tracked object. Again, we put the target object 20 times in front of the robot in various positions but within reach of the humanoid robot. We have counted the number of times the robot successfully touched the object. Results are shown in Table 2.

Table 1. Results of the distance measurement experiment

	Real distance (m)	Measured distance (m)	Difference (m)
1.	0.5348	0.3873	0.1475
2.	0.6132	0.4553	0.1475
3.	0.6972	0.6403	0.0569
4.	0.7849	0.6778	0.1071
5.	0.8752	0.7223	0.1529
6.	0.9675	1.0049	–0.0374
7.	1.0142	0.8145	0.1997
8.	1.0611	1.0870	–0.0259
9.	1.1559	1.0040	0.1519
10.	1.2514	1.1738	0.0776
11.	1.2994	1.0040	0.2954
12.	1.3476	1.2771	0.0705
13.	1.4443	1.2771	0.1672
14.	1.4929	1.0851	0.4078
15.	1.5414	1.5327	0.0087
16.	1.6389	1.6927	–0.0538
17.	1.7367	1.3960	0.3407
18.	1.8346	1.5327	0.3019
19.	1.9329	1.6927	0.2402
20.	2.0313	1.6927	0.3386
	Average absolute measurement error (m):		**0.1670**

The results of the distance measurement experiment have shown that the proposed system is not suitable for precise distance measurement. Considerable inaccuracies mainly at greater distances of the target object are caused by the nature of the sensing system (small spacing of the cameras and lower resolution used) and by inaccurate localization of the object by the TLD systems. The possible solution to this is the implementation of corrective procedures improving the localization based on epipolar geometry.

The experiment with the hand controller has shown that the closer to the robot the 3D tracking system is, the more accurate it is to enable the touching or grasping of the target object. This can be further improved by the implementation of feedback control.

As it is, the proposed object tracking system is unable to tell more about the target object besides its location. This would be a problem if manipulation with the objects of different sizes or shapes is desired. To solve this problem a different object tracking method could be used (see [8]) or we can try to track various parts of the tracked objects separately to determine their orientation additionally.

Table 2. Results of the hand controller testing experiment

	NN Inputs (object coordinates)			NN Outputs (hand joint positions)			OK?
	d (m)	*v* (rad)	*h* (rad)	s_p (rad)	s_r (rad)	e_r (rad)	
1.	0.134146	0.027178	0.290937	0.057461	0.357181	-0.788319	Y
2.	0.190570	-0.068388	0.589369	-0.172792	0.477867	-0.169463	Y
3.	0.198999	-0.073607	0.765424	-0.184782	0.740034	-0.160568	Y
4.	0.164649	0.103553	0.424750	0.061034	0.431430	-0.588861	Y
5.	0.221017	-0.173743	0.923930	-0.495389	0.778589	-0.039320	Y
6.	0.223769	0.021826	0.735878	-0.166607	0.515333	-0.049318	Y
7.	0.162134	-0.435731	0.842525	-0.614742	1.071152	-0.477502	N
8.	0.194244	-0.387002	1.168614	-0.553836	1.345710	-0.174646	N
9.	0.191741	-0.055399	0.831764	-0.122697	0.975315	-0.333835	Y
10.	0.185708	-0.190491	0.733462	-0.329463	0.733202	-0.207705	Y
11.	0.153799	-0.011105	0.150542	-0.038457	0.190594	-0.393123	Y
12.	0.163303	0.320239	0.386022	0.225971	0.462984	-0.762712	Y
13.	0.185293	0.017649	0.357128	-0.072228	0.261859	-0.185558	Y
14.	0.217610	-0.034607	0.627409	-0.253171	0.383026	-0.042686	Y
15.	0.201207	-0.011139	0.518828	-0.137416	0.347757	-0.096682	Y
16.	0.209144	-0.268592	1.175758	-0.458409	1.302464	-0.116085	Y
17.	0.129244	-0.330636	0.627409	-0.385313	0.940373	-1.098963	N
18.	0.141164	-0.175151	0.122507	-0.166845	0.178159	-0.397670	N
19.	0.193073	-0.401333	0.510929	-0.773710	0.276209	-0.030392	Y
20.	0.193977	0.237270	0.403652	0.181841	0.320561	-0.257261	Y
	Successful touch: **16**			Failed touch: **4**		Percentage: **80 %**	

Each row represents one measurement, distance values d are given in meters and all other values are given in radians. Inputs of the neural network hand controller: d – object distance, v – vertical directional angle of the object, h – horizontal directional angle of the object. Outputs of the neural network hand controller: s_p – shoulder pitch joint, s_r – shoulder roll joint, e_r – elbow roll joint. The last column determines if the touch was successful or not.

7 Conclusion

We have implemented a system processing images from two cameras of a stereovision system mounted on a Nao humanoid robot to estimate 3D coordinates of the target object. The images are captured at approximately 10 fps and individually processed by TLD tracking systems. The applicability of the 3D tracking system was tested and the ability of the humanoid to touch objects tracked in 3D by the proposed system was verified.

Acknowledgment. The work presented in this paper was supported by VEGA, Grant Agency of Ministry of Education and Academy Science of Slovak Republic under Grant No. 1/0298/12 – “Digital control of complex systems with two degrees of freedom.” The work presented in this paper was also supported by KEGA under Grant No. 018TUKE-4/2012 – “Progressive methods of education in the area of control and modelling of complex systems object oriented on aircraft turbo-compressor engines.”

References

1. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. In: Conference on Computer Vision and Pattern Recognition (2010)
2. Kalal, Z., Matas, J., Mikolajczyk, K.: Online learning of robust object detectors during unstable tracking. In: 3rd Online Learning for Computer Vision Workshop 2009. IEEE CS, Kyoto (2009)
3. Kalal, Z., Matas, J., Mikolajczyk, K.: Forward-Backward Error: Automatic Detection of Tracking Failures. In: International Conference on Pattern Recognition, Istanbul, Turkey, August 23-26 (2010)
4. Aldebaran Robotics. Nao Website, `http://www.aldebaran-robotics.com/en/`
5. Aldebaran Robotics. Nao Documentation v1.14.2, `http://www.aldebaran-robotics.com/documentation/index.html`
6. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)
7. Breiman, L.: Random forests. ML 45(1), 5–32 (2001)
8. Yilmaz, A., Javed, O., Shah, M.: Object Tracking: A Survey. ACM Comput. Surv. 38(4), Article 13, 45 (2006)
9. Puheim, M.: Application of TLD for object tracking in stereoscopic images. Diploma thesis. Technical University of Košice. Faculty of Electrical Engineering and Informatics. Košice, 68 pages (2013) (in Slovak)