C2M — Coordinated Conscious Machines

Kyungnam Kim and Arturo Flores
HRL Laboratories, LLC.
3011 Malibu Canyon Rd.
Malibu, California 90265

Abstract

In the Coordinated Conscious Machines (C2M) project, we are developing robots that interact with humans and other robots to form a team to perform tasks with heterogeneous and complementary abilities. At IJCAI 2009, our robotic perception and control techniques developed under C2M are demonstrated using three different robotic platforms - a mini ground vehicle (R/C car) that automatically follows another car, a humanoid version of the Lego Mindstorm NXT that follows a ball automatically, and a 23” tall humanoid showing human-robot interaction capabilities such as audio source localization, visually-guided grasping, and object detection/localization.

Introduction

The central theme of the C2M project consists of 1) observing and imitating the behaviors of another similar entity, 2) anticipating from the environment and other entities the needs for future actions, and 3) complementing other people or agents in the environment based on what was anticipated.

Example tasks that a C2M robot is expected to perform include following humans through buildings, imitating stealthy behavior, manipulating tools and weapons, learning complex assembly by observation, and moving as a team to provide cover or surveillance. The premise is that developing along these central themes will produce the basic functionalities that will alleviate the meticulous programming required in state-of-the-art robotic systems to adapt a system to perform new behaviors robustly in unstructured environments. Research is motivated by the needs of coordinated and cognitive machines like robots that learn complex assembly/disassembly tasks from example, cars that avoid accidents by anticipating actions of pedestrians and other drivers, micro UAVs that navigate unfamiliar indoor environments.

Our robotic perception and control techniques developed under C2M are demonstrated using three different robotic platforms. First, a mini autonomous ground vehicle (R/C car) automatically follows another car. This involves robust visual tracking on unstable wireless videos and wireless PC-based car control. Second, we show that a humanoid version of the Lego Mindstorm NXT can follow a ball automatically. This is possible by autonomous ball tracker and control algorithm embedded in RoboRealm (Rob see website). Third, a more advanced humanoid, named NAO (NAO see website), demonstrates audio source localization using a microphone array, visually-guided grasping of an object with a fiducial marker, and object detection/localization for faces and any objects trained for classification.

The remaining sections are presented with visual tracking, audio source localization, object detection, and finally conclusion and future work.

Visual Tracking

Visual tracking has been thoroughly studied by computer vision researchers and is widely used in many different application areas such as video surveillance, bio-medical image analysis, gaming, and robotics. There are many different ways to perform object tracking from video, from simple color blob tracking to sophisticated probabilistic methods like particle filtering (Arulampalam et al. 2002) or mean-shift (Comaniciu and Ramesh 2003). The capability of following an independently moving object is important to un-manned autonomous vehicles and humanoids for their human-robot interaction/coordination.

We implemented and embedded visual tracking modules into two robotic platforms - an R/C car following another car and a humanoid version of Lego NXT Mindstorm following a small ball.

Fig.1 shows the humanoid version of Lego Mindstorm NXT automatically following a colored ball. It has a USB camera as its head for visual inputs. Object following by visual tracking is possible by a ball tracker and a control algorithm, all embedded in RoboRealm (Rob see website) which has the programming interfaces for USB cameras and Lego Mindstorm NXT. The robot is wirelessly controlled through Bluetooth. The humanoid can move forward, backward, left/right by changing the speed/turning direction of the motor in each leg. The ball tracker uses a simple color blob tracking method based on color thresholding and connected component analysis. The centroid/size of the biggest blob detected is chosen for the ball location/size. Given the ball size and location (within the image) measured by the ball tracker, the proper control command is sent to the robot so it can keep the ball in the center of the image (moving...
Audio source localization

Audio source localization is the ability to identify the location or direction of a detected sound using an array of multiple microphones. The microphones are placed such that they can capture the difference between the time when sound from a single source reaches the near microphone and when it reaches the far microphone. This is called ‘interaural time difference’ (ITD). The level (amplitude) difference can also be used for sound localization and it is called ‘interaural level difference’ (ILD). The hearing capability on a mobile robot is useful because it is omnidirectional and can complement vision to help localize a person or interesting event for human-robot interaction.

In order to match a human auditory system with high-level complexity, it is not necessary to limit robots to a human-like auditory system using only two microphones. Our humanoid platform, NAO, has four microphones placed on the left, right, front, and back sides of its head. More microphones can increase the resolution of audio source detection and the robustness by filtering out noise. We developed an audio source localization algorithm used to turn NAO’s body to the source of high-level sound, for example “Hey NAO!” or “Look here!” . It measures the source based on ITD and ILD between two of four microphones, those values of which are obtained and updated at a high frequency (say 10-20Hz) by using NAO’s API functions. This sound localization capability was successfully demonstrated in indoor environments with low-level noise. When multiple sources are presented at the same time, it tends to turn to the first one which has the sound level larger than the preset threshold.

Visually-Guided Grasping

Manipulation is an important skill for any robotic system and constitutes a key component for many robotic applications on areas such as medical, industrial, service, and entertainment robotics. One of the most common robotic manipulation tasks is grasping. Most of the daily tasks performed by humans involve some form of grasping action like “Pick up the stapler on the table”. Humans effectively grasp things with their hands mostly with the aid of vision.

We developed a vision-based grasping module used in our NAO humanoid with two arms/hands. It uses visual information to identify and locate the object and moves NAO’s arm/hand to the location for grasping. Currently, our grasping module allows NAO to grasp an object with easily detectable colored marker presented within the camera view at a certain depth zone. NAO has two cameras on its head, but they do not constitute a stereo pair. The demonstrated grasping assumes the object and the robot body do not move while grasping. The sequence of the grasping steps are: (1) Detect the object to be grasped using the color blob tracker, (2) Move the end-effector from the initial position to the object position, (3) While moving, fully open the hand and close when the destination position has been reached.
Figure 3: Visually guided object grasping. Our humanoid, NAO, visually detects the blue end of the stick and moves its arm to the object location, and then grasps it.

The NAO arm motion module uses a Jacobian approach to move the end-effector of a specific chain to the destination point in a cartesian space. The cartesian space needs to be properly mapped to the visual space to find the object location in terms of the robot frame.

Fig.3 shows our NAO humanoid grasping the end of the stick with a blue fiducial marker. For future work, we plan to estimate the depth of the object, planning the approaching angle depending on the shape of the object, and picking up an object on the ground by using all joints in the humanoid body.

Object Detection

Face and Car Detection

Face detection has been widely studied and is a crucial component for effective human-robot interaction. We added the face detection capability to our NAO humanoid. We used an object detection module included in OpenCV (Ope see website), which is based on a cascade of boosted classifiers working with haar-like features (Viola and Jones 2002). The classifier is trained with a few hundred of sample views of a particular object such as a face.

The NAO face detector we developed successfully detected multiple frontal faces simultaneously at 0.5-8m, with the image resolution of 320x240 at the speed of around 3-5 fps. With the car classifier, it detects a frontal view of car. We used an R/C car since NAO is just 23 inch tall. The car detector performs worse than the face detector, probably because faces have more unique visual features than cars.

Object Localization using SIFT and RANSAC

Automatic detection and localization of general objects is a fundamental capability to a robot that can assist humans with basic tasks. As an example, a human points to an object and tells the robot to retrieve it. The robot then uses computer vision and machine learning techniques to find the object, navigate towards it, and retrieve it.

We use the Scale Invariant Feature Transform (SIFT) (Lowe 1999) and Random Sample Consensus (RANSAC) (Fischler and Bolles 1981) to localize a target object in a scene. We assume we have images for a small set of target objects. Given images of the target object and a scene containing the target object, the target object is localized via SIFT and RANSAC. First, SIFT descriptors are computed for both images. Matching descriptors are found using the algorithm suggested by Lowe (Lowe 2004) to reject ambiguous matches. RANSAC is then used to simultaneously discard outliers and compute the homography between the target object image to the scene image. Finally, a bounding box can be drawn using the homography.

Sample results can be seen in Fig.5. Qualitatively, the results for these images are good and adequate for a robot to localize the object and perform an action with it, such as retrieving it. This approach works well when you know the target objects in advance and have good multi-perspective views of the targets. This approach would not work well when the target objects are not known in advance since retrieving images using tools such as Google image search can return many false matches. As future work, we plan to implement a more robust method such as Multiple Instance Learning (MIL) (Dietterich et al. 1997) or bag-of-features (BOF) (Nowak et al. 2006).

Conclusion and Future work

We developed and demonstrated a few perception/control modules for our robotic platforms - a PC-controlled R/C car with wireless video transmission, a humanoid version of Lego Mindstorm NXT, and a more advanced humanod NAO. The capabilities of audio source localization, visually-guided grasping, face detection, and object detection/localization are very important for natural and effective human-robot interaction.

We will further develop additional perception/control methods for human pose estimation and complicated mobile manipulation like picking up from the ground. Finally, we will complete human-robot interaction challenges like the one as follows:

- **Interaction/attention**: A commander says “Hey! NAO” with a loud voice to draw its attention.
- **Audio source localization** using microphone array: NAO turns to the commander.
- **Human pose estimation**: The commander asks “Go pick up that rubber duck (or any other object)” by pointing at the duck with his/her hand.

15
Figure 5: Sample object localization results for images from (a) original SIFT paper, (b) SRVC 2008 dataset, and (c) HRL. Top row shows matching SIFT descriptors for target object and scene. Middle row shows matching SIFT descriptors with outliers removed after RANSAC. Bottom row shows bounding box on target object in scene.

- **Image data collection for learning**: NAO doesn’t know what ‘rubber duck’ is, so it does Google/Yahoo/MSN Image Search and collects matching images.
- **Training classifier**: Given the candidate images collected, NAO trains a classifier from them or just saves them for matching.
- **Bio-inspired preprocessing using visual saliency**: NAO goes the area of the commander’s pointing and generates a visual saliency map.
- **Object recognition**: Performs object recognition and localizes the objects within an area constrained by hot candidate regions in the saliency map.
- **Manipulation/control**: NAO picks the rubber duck up and comes back to the commander with it.

**References**


