J. Vis. Commun. Image R. 25 (2014) 719–726

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Recognizing object manipulation activities using depth and visual cues ☆

Haowei Liu [a,*], Matthai Philipose [b], Martin Pettersson [a], Ming-Ting Sun [a]

[a] University of Washington, Seattle, USA
[b] Intel Labs Seattle, Seattle, USA

ARTICLE INFO

Article history:
Available online 29 March 2013

Keywords:
Activity recognition
Action recognition
Joint object and action recognition
HMM
Depth camera
Temporal action recognition
Temporal smoothing
Boost

ABSTRACT

We propose a framework, consisting of several algorithms to recognize human activities that involve manipulating objects. Our proposed algorithm identifies objects being manipulated and models high-level tasks being performed accordingly. Realistic settings for such tasks pose several problems for computer vision, including sporadic occlusion by subjects, non-frontal poses, and objects with few local features. We show how size and segmentation information derived from depth data can address these challenges using simple and fast techniques. In particular, we show how to robustly and without supervision find the manipulating hand, properly detect/recognize objects and properly use the temporal information to fill in the gaps between sporadically detected objects, all through careful inclusion of depth cues. We evaluate our approach on a challenging dataset of 12 kitchen tasks that involve 24 objects performed by 2 subjects. The entire framework yields 82%/84% precision (74%/83%recall) for task/object recognition. Our techniques outperform the state-of-the-art significantly in activity/object recognition.

1. Introduction

Many day-to-day human tasks involve the manipulation of objects. Medication regimens for elders, maintenance sequences for factory workers, experimental protocols for laboratory technicians and recipes for home cooks are useful applications that people have tried to develop. Details to be monitored include the identity of the task being performed (e.g., medicating), the object being manipulated (e.g., pill box, cup) and the manner in which it is handled (e.g., unscrewing the lid, picking and placing the medicine bottle). A restricted but broadly useful setting for such tasks, adopted in most preceding work, is a fixed indoor work area such as a counter, table or floor. In this paper, we propose an approach aimed at inferring details of manipulation-based tasks in such settings, with particular emphasis on addressing the challenges posed by realistic deployment scenarios.

Work over the past decade, e.g. [1] has shown that jointly reasoning about the task being performed, the identity of object being manipulated and the hand actions being performed on the object can significantly improve the accuracy of all three. Such reasoning requires recognition of the objects being manipulated and the (3-D) pose of the hand manipulating them. However, day-to-day indoor settings pose substantial challenges to solving these problems, including variability in pose and attributes of users and objects used, severe sporadic occlusion by users and the environment, a scarcity of perspective cues to judge depth, heavy clutter in the environment and working surfaces, fast motions and variable lighting.

One way in which recent approaches have handled these challenges is to identify scenarios that avoid some of them. For instance, users are required to face the camera such that the manipulated object is always visible, objects are required to have abundant SIFT features, or work surfaces are structured so that objects occlude each other minimally. However, many natural manipulation scenarios exhibit all the above challenges routinely. In fact, the interaction of objects with hands can have a mutually confusing effect in that hand occlusion and motion may make object recognition hard, and handled objects may make hand/arm detection difficult.

Kinect [2] has been a wide success since its launch yet most of its applications are on gesture control for gaming. In this work, we examine how to use the depth images generated by Kinect, to address the above problems. The use of depth derived from a non-visual channel has several advantages. First, a rapid depth-based segmentation of the scene can usually identify users holding objects regardless of the vagaries of lighting, clothing variability, and clutter. Second, the depth image yields object size and shape as cues that can be used for recognition even if motion blur and lack of key points make local features less informative. Third, distances between different entities such as hands and the environment are directly measurable, making it easier to detect and characterize hand-environment interactions.

☆ The work was done when Haowei Liu and Martin Pettersson were with the University of Washington, Seattle, WA 98195, USA. Matthai Philipose was then with Intel Labs Seattle, Seattle, WA 98195, USA.

* Corresponding author.
E-mail addresses: hwliu@uw.edu (H. Liu), matthai.philipose@gmail.com (M. Philipose), marpett@uw.edu (M. Pettersson), mts@uw.edu (M.-T. Sun).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.03.015

In what follows, we propose a framework and algorithms that make comprehensive use of depth cues to supplement visual cues in various aspects of understanding manipulation. The main contributions of the paper are:

- A novel algorithm based on local spatial statistics to parse a 3D-cloud representing the human body into torso and arms even under heavy occlusion.
- An algorithm to use the location of the arms to locate the object being held, and to incorporate size-based features of the object into a suitable object recognizer.
- A temporal super-segmentation algorithm based on combined depth and visual cues that identifies long stretches of frames where the object being manipulated is highly unlikely to change. Using these stretches as the units of object and activity recognition improves both significantly.

We evaluate our approach on a realistic dataset with 12 activities, 24 objects and 2 subjects with various lighting conditions and poses, to analyze the value and show the effectiveness of the above techniques.

2. Related work

A number of researchers [1,3–6] over the past decade have shown the utility and feasibility of understanding object manipulation. Their emphasis has been on demonstrating the value of inferring objects, actions, and activities together. We instead focus on improving the performance of lower-level detectors for these quantities by using depth cues.

Understanding manipulation typically requires tracking arms and detecting/recognizing objects being manipulated. Past systems have made various simplifying assumptions to reduce the demands on these capabilities. Moore et al. [6] and Gupta et al. [3] do not attempt to detect and recognize objects when they are in the hand, but rather rely on detecting the point of contact with plainly visible objects on the work surface. We believe that given natural occlusion and clutter, relying solely on on-surface detection of objects is insufficient. Like Abhinav et al. [7] and Wu et al. [5], we therefore detect and recognize objects as they are being used. Unlike Abhinav et al. [7], however, we do not assume that we have a frontal view of the task space, or that the manipulation happens so that the object being manipulated is mostly visible to the camera. We assume sporadic visibility is the norm and use a temporal super-segmentation technique derived from our depth data to cope. Unlike Wu et al. [5], who rely on SIFT features to avoid having to segment out the precise outlines of the manipulated object, we exploit the shape, texture, size, and other region-based cues for object recognition. Many common objects do not have useful SIFT-style features due to their coloring, material, and motion. We use our depth-based segmentation to determine the region representing the manipulated object, including the size of the region.

To find the pose, most systems typically perform a multi-scale, multi-orientation search for candidate body parts followed by imposing constraints on the part connectivity [7–9]. [2] is by far the most successful pose estimation system in terms of speed and performance. However, it does not model the scenarios when users are holding objects. Also, in our application, we only need arm locations instead of detailed pose information. Another closely related work is [10], where the authors use the skeleton information to recognize human actions. Although some of the target actions involve object interaction, they can be recognized using motion information (e.g. tennis swing versus golf swing). Also, it assumes that the skeleton information is always available. However, for some applications, such as table/counter top object manipulation, the skeleton information might be incorrect because the lower body is occluded by the table or the head is not visible due to camera perspectives (See Fig. 2).

Given a 3-D point cloud of the user, we show how to use local spatial statistics to determine whether each patch on the user belongs to torso or limb. Our local, bottom-up technique does not rely on having a view of the head and shoulder as a starting point for parsing the 3-D cloud as these body parts are often not visible in manipulation footage. On the other hand, we do not provide full upper-body kinematic modeling (we simply identify arms and torso) or enforce consistency globally across parts.

3. System design

Fig. 1 shows the overall structure of our system. Given an incoming 640×480 frame of depth values, we detect the likely human(s) in the frame, parse them into torso and arms, identify the hands and extract pixels corresponding to the handled object, and classify the object based on the visual cues from these pixels combined with size cues. We use the position of the hand and the overall structure of the scene to determine when the hand is close to environmental surfaces. When combined with pixel information on whether the handled object has changed, we are able to conservatively infer whether an object has been picked up. We segment the incoming stream at these predicted pickup events, and classify entire segments based on which object we think is being used in that segment (we currently assume one object is manipulated at a time). Given a stream of segments annotated with object use, we use a Hidden Markov Model (HMM) to infer activity over time. The following sections provide more detail.

3.1. Detecting people and arms: background subtraction and local spatial statistics

Given an incoming depth-map of the activity scene as in Fig. 2, we seek to identify the points representing the person performing

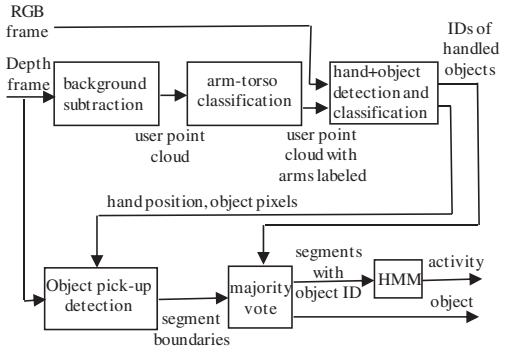
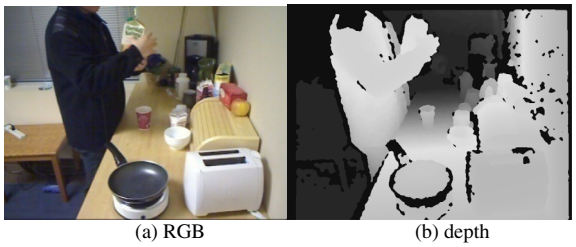


Fig. 1. High-level system design.



(a) RGB (b) depth

Fig. 2. Synchronized depth and RGB images. Image (b) is a depth map of (a), with lighter shades closer to the camera.