Visual Understanding with RGB-D Sensors: An Introduction to the Special Issue

1. INTRODUCTION

In recent years, we have witnessed the rapid growth of research in visual understanding with RGB-D sensors since the release of Micrsoft's Kinect sensor in November 2010. For a long time, researchers have been challenged by many visual understanding problems such as detecting and identifying objects or human activities in real-world situations. Traditional segmentation and tracking algorithms are not always reliable when the environment is cluttered or the illumination changes suddenly. However, the effective combination of depth and RGB data can alleviate the negative effects of environmental changes, thus improving the accuracy of object identification and tracking.

The freely available online Kinect SDKs and pose trackers for environment modeling further encourage novel solutions to traditional visual understanding problems. However, Kinect sensors face a number of specific challenges, such as missing or corrupted depth pixels and inaccurate calibration between depth and RGB cameras. This special issue is specifically dedicated to new algorithms and applications based on the Kinect sensors.

The goals of this special issue are twofold: (1) provide a survey on the progress of visual understanding with RGB-D sensors in the past years, and (2) introduce novel research and discuss new applications with the RGB-D sensors. We believe it serves as a convincing forum for researchers and practitioners to disseminate their latest research on visual understanding with RGB-D sensors.

2. REVIEW PROCESS

This special issue solicited contributions on a wide range of topics including RGB-D data acquisition, RGB-D data understanding, and new applications based on RGB-D data. We received 23 submissions from North America, Europe, and Asia. The majority of the articles were reviewed by at least three experts. The acceptance decision for each article was discussed among the guest editors, and finally, 12 papers were accepted for the special issue.

3. GUIDE TO ACCEPTED ARTICLES

The articles in this special issue provide an excellent sampling of the recent work on the visual understanding with RGB-D sensors. They can be categorized into RGB-D data acquisition (three articles), RGB-D data understanding (five articles), and new applications based on RGB-D data (four articles) as follows.

3.1. RGB-D Data Acquisition

The recent development of RGB-D sensors enables the real-time acquisition of depth maps and their corresponding color images. However, the performance is largely constrained by the limited quality of the depth maps. Take a representative RGB-D sensor, Microsoft Kinect, as an example. It usually contains distance-dependent measurement errors, misalignments between the depth discontinuities and the edges in the color image, various holes due to invalid measurement, and shadows caused by the displacement between the IR projector and camera. Chen et al. propose to integrate adaptive support region selection, reliable depth selection, and color guidance together under an optimization framework for Kinect depth recovery. Their framework inherits and

DOI: http://dx.doi.org/10.1145/2732265

improves upon the idea of guided filtering by incorporating structure information and prior knowledge of the Kinect noise model. Experiments on real Kinect data demonstrate the superior performance in terms of recovery accuracy and visual quality.

The widespread availability of RGB-D sensors promotes the efficiency of 3D video data acquisition. A problem is that a fast movement may easily lead to motion blur due to the limited exposure speed of RGB-D sensors. Gao et al. present an approach to tackle this concern by eliminating the depth error based on time-series analysis. The quality of depth video is significantly improved by means of the approach. Figueroa et al. propose to reconstruct indoor spaces based on 6D RGB-D odometry and KinectFusion. Their method finds the relative camera pose between consecutive RGB-D frames by keypoint extraction and feature matching both on the RGB and depth image planes. The estimated pose is then fed to the highly accurate KinectFusion algorithm, which uses a fast iterative-closest-point algorithm to tune the frame-to-frame relative pose and fuse the depth data into a global implicit surface. The method is evaluated on a benchmark dataset, and the experiments verified that the method outperforms the state-of-the-art RGB-D SLAM systems in terms of accuracy. Their method is furthermore ready to produce a polygon mesh without any postprocessing steps.

3.2. RGB-D Data Understanding

Image understanding aims to extract the semantics from image content, reducing the tremendous volume of images to concise semantic representation that captures the essence of the data. A wealth of research has been devoted to understanding 2D images in recent years. The emergence of RGB-D images poses new opportunities and challenges to this research topic. Zha et al. observe that the multiview features of RGB-D images offer a comprehensive representation of the objects. The exploration of multiview information, essentially the interactions/correlations among them, would be able to improve the robustness of feature learning as well as help derive rewarding features with better effectiveness. They propose a robust multiview feature learning approach to exploit the intrinsic relations among the views. The learned feature is applied to the tasks of object classification and scene categorization in experiments. The experimental results demonstrate that the effectiveness of learned features is remarkably improved.

Object-level knowledge mining from a large set of cluttered scenes poses significant challenges. Zhang et al. assert that two such challenges are how to initiate model learning with the least human supervision and how to encode the structural knowledge. They propose a model learning method that starts from a single labeled object from each category and mines further model knowledge from a number of informally captured, clutter scenes. The robust 3D shapes from RGB-D images are used to reduce the model bias due to less supervision. They believe in this way, the trained category models are able to detect and recognize objects in RGB images, and furthermore, they can be transferred to guide model learning for a new category where depth information is missed. Experiments show that the performance of the method is comparable to fully supervised learning methods. Huang et al. propose an algorithm to segment moving objects for telepresence systems. They claim that approaches based on depth sensors usually suffer from mis-segmentations on the object boundary due to inaccurate and unstable estimation of depth data. However, their adaptive multicue decision fusion algorithm can accurately and robustly segment moving objects in real time. Well-designed experiments conducted on benchmark dataset demonstrate their claims.

Another active research topic is to understand human action—that is, human action recognition. It shows great potential by using 3D depth data. Zhu et al. focus on finding some complementary features and combine them to improve the recognition accuracy. They study different fusion schemes comprehensively, using diverse features for action characterization in depth videos. Experiments on four challenging depth

action datasets show that their fusion scheme is able to improve accuracy significantly and outperform the state-of-the-art approaches. Spurlock et al. propose to use RGB-D sensors to acquire annotated training data for human pose estimation from 2D images. Their approach is able to run on any gesture-based games, which differentiates it from others. Extensive experiments validate the effectiveness of the approach.

3.3. New Applications Based on RGB-D Data

Sign language recognition transcribes sign language into text accurately and efficiently. The current challenges for this research topic are how to capture the optimal features efficiently from signers and how to model different signs and correctly classify them. Therefore, the presence of depth information could greatly benefit vision-based sign language recognition. Sun et al. propose to first assign a binary latent variable to each frame in training for indicating its discriminative capability, and then develop a latent support vector machine model to classify the signs and localize the discriminative and representative frames in each video. The experiments conducted on a benchmark American Sign Language dataset demonstrate the effectiveness of the method. Tang et al. propose a two-stage sign language recognition system. The first stage is to detect and track the movement of hands, while the second stage is to automatically learn features from hand posture images by Deep Neural Networks. The achieved real-time recognition accuracy of this system is as high as 98.12%.

Drowsy driving is one of the major reasons for fatal traffic accidents. Zhang et al. present a real-time system that utilizes RGB-D sensors to automatically detect driver fatigue for alerts. They first present a real-time 3D head pose estimation and design a scheme to predict eye states based on Weber Local Binary Pattern. They believe the combination of the two visual cues can reduce uncertainties and resolve ambiguities. The system is deployed in inside-car environments, and the all-day test verified its effectiveness and robustness. Kyan et al. present a framework for the real-time capture, assessment, and visualization of ballet dance movements as performed by a student in an instructional and virtual reality setting. In the system, movement data is facilitated by skeletal joint tracking using RGB-D sensor, while instruction and performance evaluation is provided in the form of 3D visualizations and feedback through a CAVE virtual environment. The proposed framework is based on the unsupervised parsing of ballet dance movement into a structured posture space using the spherical self-organizing map. The evaluation of the recognition accuracy and the virtual feedback functionality of the systems is performed to verify its effectiveness and showcase its industrial potentials.

ACKNOWLEDGMENTS

We thank all the reviewers for their valuable comments that ensure the high quality of this special issue, and all the contributing authors for their interesting and innovative work. We would also like to thank Editor-in-Chief Professor Qiang Yang for providing guidance and the assistant to EIC Dr. Weike Pan for his support, help, and patience throughout the process.

Richang Hong
Hefei University of Technology Hefei, Anhui, China
hongrc@hfut.edu.cn

Shuicheng Yan
National University of Singapore, Singapore
eleyans@nus.edu.sg

Zhengyou Zhang
Microsoft Research
zhang@microsoft.com

Guest Editors