Int J Comput Vis (2016) 118:113–114
DOI 10.1007/s11263-016-0915-4

Guest Editorial: Human Activity Understanding from 2D and 3D Data

Junsong Yuan[1] · Wanqing Li[2] · Zhengyou Zhang[3] · David Fleet[4] · Jamie Shotton[5]

Published online: 30 May 2016

Automatic analysis of human motion has become one of the most active research topics in computer vision due to both the scientific challenges of the problem and the wide range of applications. Such applications include intelligent video surveillance; human–computer interfaces; intelligent humanoid robots; gaming; diagnosis, assessment, and treatment of musculoskeletal disorders; sports analysis; realistic synthesis and animation of human motion; and monitoring of elderly and disabled people at home. Extensive studies have been conducted in the past decade using 2D visual information captured by single or multiple cameras. However, the problem is far from being robustly solved, especially for viewpoint independent recognition of diverse human actions and activities in a real environment.

Recent advances in 3D depth cameras using structured light or time-of-flight sensors, 3D information recovery from 2D images/videos, and the availability of portable human motion capture devices have made 3D data readily available. This 3D data is nurturing a potential breakthrough in human activity recognition. The release of Microsoft's Kinect sensors, ASUS's Xtion Pro Live sensors, and Intel's RealSense cameras have provided a commercially viable hardware platform to capture 3D data in real-time. This special issue seeks high quality and original research on human activity understanding using such 3D input, as well as that using more traditional 2D input. The goal of this special issue is two-fold: (1) to advocate and promote research in human activity recognition using 2D and 3D data; and (2) to present novel human activity understanding techniques applicable to diverse applications.

This special issue is composed of eight papers, each of which went through a rigorous review process with at least 3 reviewers. We summarize these eight papers below:

Kernelized Multiview Projection for Robust Action Recognition by Shao et al. (doi:10.1007/s11263-015-0861-6) presents a practical algorithm to fuse multiple different features for action recognition. The algorithm is based on a linear multi-view embedding method using kernel matrices from different views. Using an alternate optimization via relaxation, a near-optimal projection and weights for each view are learned. Experiments demonstrate significant improvements over single view methods and other feature fusion methods.

Exploiting Privileged Information from Web Data for Action and Event Recognition by Li et al. (doi:10.1007/s11263-015-0862-5) proposes a new learning method to train robust classifiers for action and event recognition by using web videos as freely available training data. The proposed multi-instance learning methods not only take advantage of the additional textual descriptions of training web videos as privileged information, but also explicitly cope with noisy labels. New domain adaptation methods are also proposed to deal with the situation when the training and test videos come from different data distributions.

Fusion R features and Local Features with Context-aware Kernels for Action Recognition by Yuan et al. (doi:10.1007/s11263-015-0867-0) presents a new feature that captures the global spatio-temporal distribution of interest points. The feature is extracted through 3D *R* transform and fused with the widely used bag-of-visual-words feature extracted locally at the spatio-temporal interest points. To improve the robustness, a context-aware kernel approach is developed for

✉ Junsong Yuan
JSYUAN@ntu.edu.sg

1 Nanyang Technological University, Singapore, Singapore

2 University of Wollongong, Wollongong, Australia

3 Microsoft Research, Redmond, WA, USA

4 University of Toronto, Toronto, Canada

5 Microsoft Research, Cambridge, UK

similarity measurement. The paper demonstrates the value of the R feature and the context-aware kernel in action recognition through experiments on the UCF Sports, UCF Films and Hollywood2 datasets.

Capturing Hands in Action using Discriminative Salient Points and Physics Simulation by Tzionas et al. (doi:10.1007/s11263-016-0895-4) works on hand motion capture in interaction scenarios, where hands interact with other hands or objects. The proposed method combines a generative model with discriminatively trained salient points to achieve a low tracking error. It also incorporates collision detection and physics simulation to achieve physically plausible estimates even in case of occlusions and missing visual data. The method can work well for monocular RGB-D sequences as well as setups with multiple synchronized RGB cameras.

Gaze Estimation in the 3D Space Using RGB-D sensors. Towards Head-Pose and User Invariance by Mora and Odobez (doi:10.1007/s11263-015-0863-4) studies the problem of 3D gaze estimation within a 3D environment from remote sensors, which is valuable for human–human and human–robot interactions. It leverages the depth data of RGB-D cameras to perform an accurate head pose tracking, acquires head pose invariance through a 3D rectification process that renders head pose dependent eye images into a canonical viewpoint, and computes the line-of-sight in 3D space. To address the low resolution of the eye image, an appearance-based gaze estimation paradigm is applied. They demonstrate good performance through extensive gaze estimation experiments on a public dataset as well as a gaze coding task applied to job interviews in a natural setting.

Multi-modal RGB-Depth-Thermal Human Body Segmentation by Palmero et al. (doi:10.1007/s11263-016-0901-x) addresses the problem of human body segmentation from multi-modal visual cues. Human body detection and segmentation plays an important role in automatic human behavior analysis. A new RGB-Depth-Thermal dataset is provided in this work, along with a multi-modal segmentation baseline. Several modalities are registered using a calibration device and a registration algorithm. They also report solid results on the new dataset.

A Hierarchical Video Description for Complex Activity Understanding by Liu et al. (doi:10.1007/s11263-016-0897-2) describes a latent discriminative structural model to automatically detect a complex activity, atomic actions and the temporal structure of atomic actions. The associated model learning method is semi-supervised and requires that the training video samples be partially annotated with atomic actions. The paper demonstrates how the model is used to extract rich and hierarchical description of activities from videos.

A Deep Structured Model with Radius-Margin Bound for 3D Human Activity Recognition by Lin et al. (doi:10.1007/s11263-015-0876-z) extends the convolutional neural network (CNN) for action recognition from RGB-D data. The idea consists of integrating latent temporal structure with CNNs to decompose an activity video into sub-activity segments, and training several CNNs, one CNN per segment, to learn the spatiotemporal features. To improve generalization on small training data sets, the paper adopts radius-margin bound regularization in the classification. The model is evaluated on several public available RGB-D datasets.

These eight papers cover a diverse range of human activity understanding from 2D and 3D data. They share new state-of-the-art ideas and technology, and will benefit researchers and engineers who work in this area. With 2D and 3D visual sensors becoming cheaper and more capable, we believe human activity understanding will continue to be a fertile area for growth given its many exciting applications.