RGBD Human Activity Recognition by Multi-Modal Context Fusion

Bingbing Ni
Advanced Digital Sciences Center
Singapore 138632
bingbing.ni@adsc.com.sg

Jiashi Feng
Vision and Learning Center
UC Berkeley, CA
jshfeng@gmail.com

Pierre Moulin
UIUC
IL 61820-5711 USA
moulin@ifp.uiuc.edu

***Abstract*—We propose a novel complex activity recognition and localization framework which effectively fuses information from both grayscale and depth image channels at multiple levels of the video processing pipeline. In the individual visual feature detection level, depth based filters are applied to the detected human/object rectangles to remove false detections. In the next level of interaction modeling, three dimensional spatial and temporal contexts among human subjects or objects are extracted by integrating information from both grayscale and depth images. Depth information is also utilized to distinguish different types of indoor scenes. Finally, a latent structural model is developed to integrate the information from multiple levels of video processing for activity detection. Extensive experiments on a challenging grayscale + depth human activity database which contains complex interactions between human-human, human-object and human-surroundings demonstrate the effectiveness of the proposed multi-level grayscale + depth fusion scheme.**

I. INTRODUCTION

Video based human action recognition and localization is a difficult task due to the individual variations of people in terms of posture, motion, clothing, view angle, camera motion, illumination changes, occlusions, self-occlusions, and the complex and cluttered background. Recognizing more complex human activities that involve human/human, human/object or human/surrounding interactions, often requires a combination of individual human appearance, relative motion, and contextual information. However, the performance of this individual visual feature + context model heavily relies on the quality of the visual feature detection and the accuracy of the contextual information encoding. On the one hand, using state-of-the-art techniques in computer vision, very accurate detection of salient image or motion features such as human key poses, spatial-temporal interest points, motion trajectories etc., is not achievable, due to the complex background. Current methods usually model the pairwise contextual information between individual visual features (*e.g.*, two human key pose detections) based on their spatial and temporal displacement and relative velocity, however, as these quantities are measured in 2D using conventional video cameras, ambiguity could arise and therefore the context encoding could be imprecise. For example, human subjects with large displacement in the depth direction could be spatially very close in the 2D image due to perspective projection.

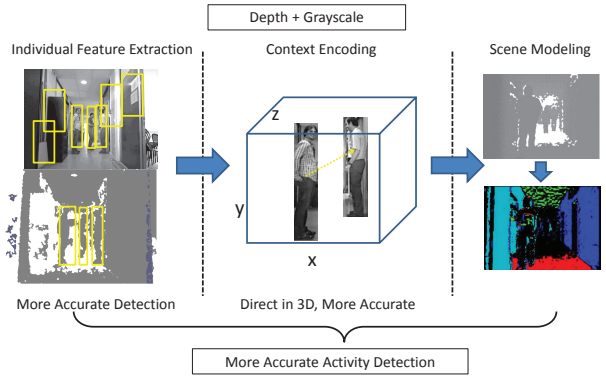

Fig. 1. Motivation of the proposed multi-level depth and image fusion framework for activity detection.

The recent emergence of depth sensors (*e.g.*, Microsoft Kinect) provides information about the 3D structure of the scene as well as the 3D motion of the subjects/objects in the scene. A comprehensive review on the applications of Kinect can be found in [1]. Intuitively, utilization of depth images can alleviate the above-mentioned problems (*i.e.*, inaccurate individual visual feature detection and context modeling) and eventually benefit activity detection. First, the depth image provides layered information about the scene, therefore detection of the salient foreground visual features, *e.g.*, human key poses, could be more reliable, even in the presence of severe background clutters. Second, as the registered depth + image directly provides the 3D coordinates of each scene point, accurate modeling of 3D spatial and temporal relationship for representing various interaction becomes feasible. In addition, the depth map contains very rich 3D structural information about the scene, therefore the scene type can be represented more accurately by using depth map than by using conventional image. Note that the global scene contextual information is also an important cue for activity recognition.

Motivated by these observations, we propose a novel framework which effectively integrates depth information with the conventional grayscale image at different levels of the processing pipeline including: 1) individual feature extraction; 2) pairwise contextual information encoding; and 3) global scene representation. We show that using depth information in these processing stages can boost the performance of complex activity recognition and localization (*i.e.*, activities contain various human/human, human/object, or human/surrounding interactions). The motivation for this work is illustrated in Figure 1. A full version of our work was published in [2].

II. Method Overview

Our basic idea is to integrate the depth information into multi-level processing pipeline for detecting human activities that involve human/human, human/object or human/surrounding interactions. Figure 2 illustrates our proposed processing pipeline for activity recognition and localization. Depth information at various processing stages is utilized and integrated in the following way. In the visual feature extraction step, we detect human key pose and object of interest in every input frame of the grayscale image sequence and the corresponding depth maps provide further constraints to filter out false detections. These human key pose and object detections are afterwards spatially and temporally matched throughout frames into *tracklets* by applying the motion constraints in both grayscale and depth channel. As a byproduct, invalid detections without sufficient temporal durations are further filtered out at this stage. In the next stage, we model the 3D spatiotemporal interaction/contextual attributes using combined grayscale and depth information. In the third stage, depth information is utilized for classifying the indoor scenes into different scene categories. Finally, the obtained spatial-temporal interaction attributes, key pose attributes of the tracklets and the scene classification results are integrated by a latent structural SVM for discriminatively recognizing and localizing activities.

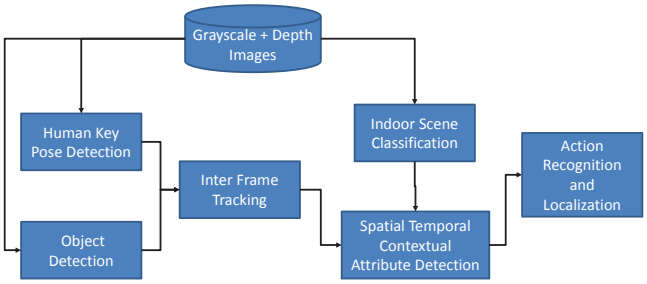

Fig. 2. Overview of the proposed action detection framework.

III. Experiments

We use the HARL 2012 competition dataset [4] for an action recognition and localization experiment. For detailed information about the dataset and the evaluation protocol please refer to http://liris.cnrs.fr/harl2012/evaluation.html.

We first show that using the depth induced constraints, human subject/object detection can be made more accurate. To demonstrate this, we randomly choose 1000 human subject detection results with groundtruth manual labels by directly applying the HOG-SVM detector without depth constraints from the testing video sequences. We then set to zero the detection scores for those samples that violate any of the two depth constraints developed in this work. The comparison of the ROC curves with and without depth constraints is illustrated in Figure 3. We note that using the depth constraints, large portion of the false detections is removed.

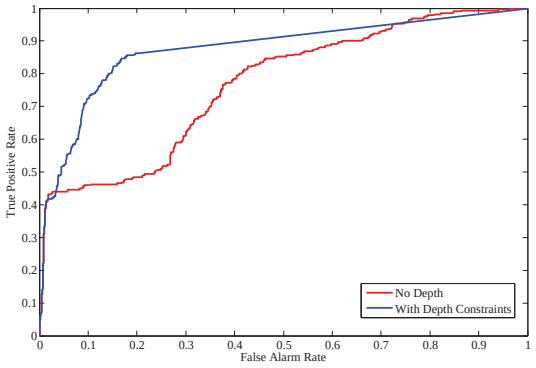

Fig. 3. ROC curves for human subject detection results. Red: without the depth constraints; Blue: using the depth constraints.

TABLE I
Comparisons of performance (F-scores for action recognition and localization) with the state-of-the-art methods.

Measure	I_{sr}	I_{ps}	I_{rt}	I_{pt}	I_{all}
Yuan et. al [5]	0.214	0.367	0.225	0.358	0.291
Wang et. al [3]	0.192	0.308	0.245	0.316	0.265
Ours	0.317	0.448	0.295	0.430	**0.372**

We then compare the performance with the state-of-the-art action detection and localization methods including: 1) The method proposed in [5], where spatiotemporal interest points (STIP) are extracted from representing actions and subvolume mutual information maximization is used to effectively search the best activity volume, *i.e.*, localization; and 2) The part-based action recognition method proposed in [3]. For both methods, the best parameters are empirically tuned based on the training data. The comparisons are shown in Table I. We note that the proposed method greatly outperforms the previous art.

References

[1] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *T-SMC-B*, 2013.

[2] B. Ni, Y. Pei, P. Moulin, and S. Yan. Multilevel depth and image fusion for human activity detection. *T-SMC-B*, 43(5):1383–1394, 2013.

[3] Y. Wang and G. Mori. Hidden part models for human action recognition: Probabilistic versus max margin. *T-PAMI*, 33(7):1310–1323, 2011.

[4] C. Wolf, J. Mille, L. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandrea, C. Bichot, C. Garcia, and B. Sankur. The LIRIS human activities dataset and the ICPR 2012 human activities recognition and localization competition. *Technical Report RR-LIRIS-2012-004, LIRIS Laboratory*, 2012.

[5] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. *T-PAMI*, 33(9):1728–1743, 2011.