Image and Vision Computing 32 (2014) 870–883

Contents lists available at ScienceDirect

Image and Vision Computing

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

Estimating layout of cluttered indoor scenes using trajectory-based priors

Muhammad Shoaib *, Michael Ying Yang, Bodo Rosenhahn, Joern Ostermann

Institute for Information processing (TNT), Leibniz University Hanover, Appelstr. 9A, 30167 Hannover, Germany

ARTICLE INFO

Article history:
Received 23 September 2013
Received in revised form 2 April 2014
Accepted 21 July 2014
Available online 27 July 2014

Keywords:
Scene segmentation
Trajectory
Scene layout
Semantic context
Conditional random field

ABSTRACT

Given a surveillance video of a moving person, we present a novel method of estimating layout of a cluttered indoor scene. We propose an idea that trajectories of a moving person can be used to generate features to segment an indoor scene into different areas of interest. We assume a static uncalibrated camera. Using pixel-level color and perspective cues of the scene, each pixel is assigned to a particular class either a sitting place, the ground floor, or the static background areas like walls and ceiling. The pixel-level cues are locally integrated along global topological order of classes, such as sitting objects and background areas are above ground floor into a conditional random field by an ordering constraint. The proposed method yields very accurate segmentation results on challenging real world scenes. We focus on videos with people walking in the scene and show the effectiveness of our approach through quantitative and qualitative results. The proposed estimation method shows better estimation results as compared to the state of the art scene layout estimation methods. We are able to correctly segment 90.3% of background, 89.4% of sitting areas and 74.7% of the ground floor.

1. Introduction

Estimating layout or structure of an indoor scene is important for many tasks, such as activity analysis [1], robot navigation [2,3], scene understanding [4,5] and object placement [6–8]. Specifically for the analysis of elderly activity, scene layout provides a semantic context knowledge that is necessary for long-term observation. With the help of scene context, we can localize a person and monitor his daily behavior. Semantic context also benefits the unusual event prediction, such as fall detection [9,10]. Lying on the sofa has a different interpretation from lying on the floor. With semantic context information usual lying on sofa can be taken as usual activity.

Keeping these important aspects and applications in mind, different mechanisms have been proposed for indoor scene layout estimation. State of the art methods use either spatial image features [11–13] or trajectory-based temporal information [1,5] for this purpose. However, both types of features have their own problems and either of the features cannot be recommended individually to estimate the layout of indoor scene. A major challenge for spatial image-based feature techniques arises from the fact that most indoor scenes are cluttered by a lot of furniture and decorations [14]. They often obscure the geometric structure of the scene, and also occlude boundaries between walls and the floor. Appearances and layouts of clutters can vary drastically across different indoor scenes, so it is extremely difficult (if not impossible) to model them consistently. Similarly trajectory-based techniques normally cluster the trajectory data and model only the paths [5,15]. They do not take care of the clutter or resting places in the scene. The modeling of resting areas [1] has been done using stop points of a trajectory inside a resting place. This mechanism cannot be reliably used in case of noise in the trajectory data. A normal stop outside a resting area might be taken as a resting place.

Trajectory data either can be directly used or it can be used for action recognition. Action information in turn can be used for different purposes like for detecting action consequences and classifying videos of manipulation action according to action consequences [16]. Similarly another approach encodes the essential changes in visual scenery in a condensed way such that a robot can recognize and learn a manipulation without prior object knowledge [17].

Though image features and trajectory data are not self-sufficient for reliable scene layout estimation but they can be used together to achieve reliable indoor scene layout estimation. In this work, we propose a mechanism which learns the scene semantic context model using image segmentation mechanism in an unsupervised way. We do not use trajectories directly for scene layout estimation rather our segmentation mechanism uses both spatial image and trajectory-based features. We are also able to model the resting or sitting places in the scene. An overview of the approach is as follows. We assume a static and uncalibrated surveillance camera in the scene. Given a moving person in the scene; we first model the trajectory of moving person using a set of key-points on his silhouette. We identify or cluster the regions corresponding to feet locations of moving person as floor. Given a potential floor area, we define the relative height of each point relative

* Corresponding author.
E-mail address: shoaib@tnt.uni-hannover.de (M. Shoaib).

http://dx.doi.org/10.1016/j.imavis.2014.07.003
0262-8856/

to the floor. Similarly using lines and trajectories we define the orientation of each point in the scene. We now incorporate height, orientation and color information into a Conditional Random Field (CRF) to define relationship between different points in the scene. Fig. 5 gives an overview of the CRF-based image segmentation for unsupervised scene layout estimation procedure. A graphcut-based inference algorithm is run on our CRF to define the final scene segmentation or layout.

1.1. Contributions

The key contribution of this work are as follows:

(1) Indoor scene layout estimation using both trajectory data of a moving person and image features. The estimation process is fully automatic and unsupervised. We do not use any training data. No assumptions are considered about the structure of the scene.
(2) Efficient estimation of the scene structure in the presence of scene clutter. We classify scene clutter as either sitting areas or scene background. Modeling resting areas as a separate class improves overall scene layout estimation process.
(3) We show that using line segments instead of voting-based straight lines we can obtain better orientation map or surface normal. Improvement in orientation map improves overall scene layout estimation by providing correct orientations for resting places like sofa and bed.
(4) Experiments are performed on a new dataset of scene videos with moving person along with publicly available videos of the indoor scenes and better segmentation results are achieved. We show using quantitative and qualitative results that by combining trajectory information of moving persons with image attributes, we can obtain an accurate indoor scene layout, superior to geometric methods. We will make the data and segmentation results publicly available for comparison.

The rest of the paper is organized as follows. Related work and contributions of this paper are described in Section 2. The proposed scene layout estimation mechanism is elaborated in Section 3. Image segmentation mechanism used for scene layout estimation is explained in Section 4. The performance of the approaches are evaluated in Section 5. The proposed approach is compared with other approaches in this section followed by a conclusion in Section 6.

2. Related work

Layout of indoor scenes has been estimated in literature mainly using single-image segmentation [11–13,18–20]. Such techniques use the image attributes like line segments, geometric context and color to define 3D structure or geometry of the scene. However, recognition of scene structure using only spatial image features is challenging. In [21,12] different features are learned from an image database and then these learned features are used to train a classifier to segment a scene into different layers like ceiling, walls, and clutter. These features like color context are not discriminative enough for different classes. Due to color similarity a part of a wall might be detected as scene clutter or vice versa. Another set of approaches tries finding volumetric structures inside the scene to define different objects in the single images [22,11,23,19]. They also find cubic objects like beds in the scene image. They have high dependencies on straight lines in the scenes. In home environment it is difficult to detect all straights lines on objects due to cluttered scene and occlusions. Such approaches fail to detect objects like bed and sofa if they do not have enough straight lines and cubic constraint is not fulfilled.

Some other techniques use supplementary information to compensate the shortcomings of spatial image features. In the recent years features generated from laser range data [24], or Kinect-based 3D data [25–27] have been used for scene layout estimation. Similarly different areas in a scene are marked as suitable for sitting using 3D data by their ability to support a sitting action [28]. Structure from camera motion has also been used in the layout estimation of the indoor scenes [29].

Tracking information has also been used in literature to couple different actions with certain regions in the scene. Resting areas are modeled as a Gaussian mixture using minimum description length [1]. The image points where a person stopped in the scene are clustered as sitting or resting areas. Simply using tracking information is not sufficient enough for scene layout estimation while a person stopping outside a resting area might be taken as a resting area.

Some other trajectory-based approaches [30,31,23,28] detect different areas in the scene by object interactions. A chair or a sofa for example is detected when someone sits in a particular scene area. Motion or person interactions alone are not reliable enough for scene layout estimation. Motion-based feature like speed is affected by the errors in moving object detection mechanism. Similarly user interaction-based methods are dependent on user detection and posture classification. Any error in these two modules will propagate in scene layout estimation process.

In this work, we build on these efforts and take one step further to jointly segment scene and sitting places. We combine trajectory information of the moving persons along image attributes like color and perspective cues to segment indoor scene into activity areas like ground floor, inactivity areas and sitting places like bed, table, sofa and the remaining image area as background. We assume that sitting places are higher than floor and have orientation similar to floor. As objects like tables can also be used for sitting and poses similar attributes: i.e. they have surface orientation similar to a bed and are higher than floor, we also consider them as sitting or resting places.

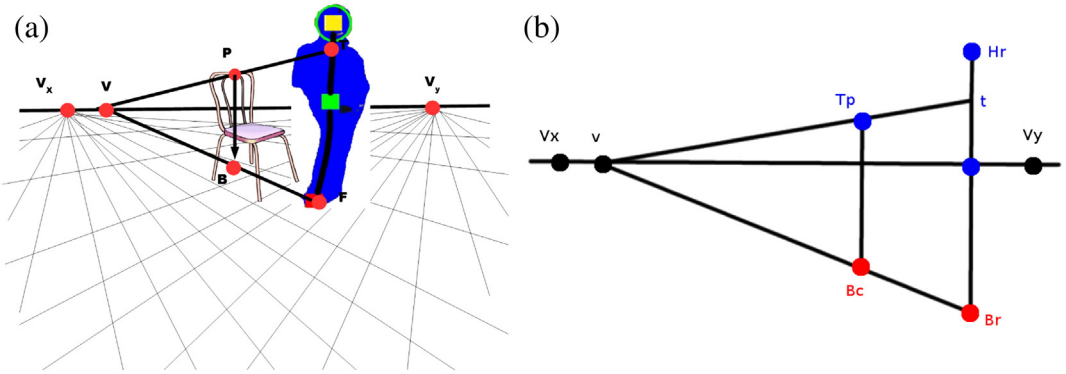

Fig. 1. Height computation (a) process for finding the floor of a point p (b) process to compute the height of the point p.