J. Vis. Commun. Image R. 38 (2016) 82–99

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Depth-based detection with region comparison features ☆

Ruud Mattheij *, Kim Groeneveld, Eric Postma, H. Jaap van den Herik

Tilburg Center for Cognition and Communication (TiCC), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

ARTICLE INFO

Article history:
Received 15 May 2015
Accepted 16 February 2016
Available online 23 February 2016

Keywords:
Face detection
Person detection
Depth data
Haar-like features
Random forest classifier
Integral image representation
Region comparion
Kinect

ABSTRACT

Most object detection approaches proposed over the years rely on visual features that help to segregate objects from their backgrounds. For instance, segregation may be facilitated by depth features because they provide direct access to the third dimension. Such access enables accurate object-background segregation. Although they provide a rich source of information, depth images are sensitive to background noise. This paper addresses the issue of handling background noise for accurate foreground–background segregation. It presents and evaluates the Region Comparison (RC) features for fast and accurate body part detection. RC features are depth features inspired by the well-known Viola–Jones detector. Their performances are compared to the recently proposed Pixel Comparison (PC) features, which were designed for fast and accurate object detection from Kinect-generated depth images. The results of our evaluation reveal that RC features outperform PC features in detection accuracy and computational efficiency. From these results we may conclude that RC features are to be preferred over PC features to achieve accurate and fast object detection in noisy depth images.

1. Introduction

In the last few years, the automatic detection of objects from digital video and image sources has gained considerable attention within the field of image analysis and understanding [1–3]. Many object detection approaches focus on two-dimensional visual features [4–6] in order to segregate objects from their backgrounds. Well-known visual features for object detection are the Haar-like features [7] as proposed by Viola and Jones [8,9].

Despite the widespread and successful use of two-dimensional (2D) visual features in visual detection tasks, they have some limitations. Their main limitation is that they typically respond to local visual transitions, without being sensitive to the larger spatial context [10]. As a consequence, they are sensitive to factors that may influence scene properties locally, such as illumination conditions [11,12]. Bright lights, for example, may cause shadows (i.e., non-object contours) in the image. Local 2D visual features will respond to the contours of the shadows in the same way as to the contours of other, real objects.

Typical situations in which 2D visual features fail, are those where variations in the third dimension (depth) lead to shape deformations. In general, the failures are caused by object pose variations [1,13]. A wide variety of methods attempts to overcome these sensitivities. The most frequently applied methods focus on extracting context-sensitive features (see, e.g., [4]). Although such approaches improve classification performance, they tend to be costly in terms of computational resources [13,14].

1.1. From 2D features to 3D features

To overcome the limitations of 2D features, researchers have added a third dimension, yielding 3D features (which combine 2D spatial and 1D depth information) [15–19]. Depth cues will then provide contextual information for a scene, thereby facilitating image segmentation [20–23]. Indeed, visual objects such as faces or persons are much easier to distinguish in a 3D space than from a 2D image [24,25]. In recent years, the use of depth cues became feasible by the development of affordable depth sensors, such as Microsoft Kinect [26].

1.2. Capturing depth with Microsoft Kinect

The Microsoft Kinect device generates its depth images by (1) illuminating a spatial area with the Kinect's infrared laser, and (2) triangulating the corresponding depth using an infrared sensor [27]. The infrared laser passes through a diffraction grating and is thus able to create a grid of infrared dots. Given the known spatial distance between the Kinect's infrared laser and sensor, the process of matching the dots observed in an image (where the dots

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: R.J.H.Mattheij@tilburguniversity.edu (R. Mattheij), K.Groeneveld@tilburguniversity.edu (K. Groeneveld), E.O.Postma@tilburguniversity.edu (E. Postma), h.j.vandenherik@law.leidenuniv.nl (H.J. van den Herik).

http://dx.doi.org/10.1016/j.jvcir.2016.02.008
1047-3203/

are projected using the pattern from the diffraction grating) allows for effective depth triangulation. The resulting depth images have a resolution of 640×480 pixels. The pixel values of the depth images encode for the distance between an object and the Kinect device. A large depth value indicates a large distance between the object and the Kinect device, while a small depth value encodes for a small distance. Figs. 4–6 show several examples of depth images that are created with a Kinect device.

1.3. Shotton's pixel comparison features

Using the Microsoft Kinect, Shotton et al. [28–30] proposed a depth-based object detection algorithm that is able to classify individual pixel locations from single depth images as belonging to faces, body joints and body parts. To classify the pixel locations, Shotton et al. select a subset of random pixel locations from each depth image. For each pixel location P from the subset, the depth difference is computed by comparing the depth values at two randomly chosen offset locations Q and R. The offset locations are defined by the radius and angle with respect to P. The radius is defined to be inversely proportional to the depth value of P. A small depth value results in a larger radius for offset locations P and Q, and vice versa. This way, a scale-invariant measure of depth is obtained. A single depth comparison between locations Q and R provides only a weak indication of the depth difference in a spatial area. Repeating the measurements for other random locations around point P, however, provides a fair description of the depth difference in an area around the location of point P. For the sake of readability, we refer to these pixel-based depth comparison features as the Pixel Comparison (PC) features.

There are two advantages of classifying individual pixel locations rather than image regions (e.g., by means of a sliding window): (1) the selection process allows for the detection of partially occluded objects, and (2) the classification process reduces the time required to process an entire depth image. Thus, using the PC features makes their detector computationally efficient. In addition to these qualities, the detector works directly on the raw input depth data, i.e., without an image preprocessing stage to reduce noise in the data [31]. The combination of efficient depth-comparison features and the raw input depth image results in a high detection speed, which allows for real-time operation.

The detection speed, however, comes at the cost of accuracy. The classification accuracy is hampered by two limitations [26,32,33]: (1) the limited quality of the depth images generated by the Kinect device, and (2) the limited resolution of the depth images. The first limitation arises from the triangulation sensor as used by the Kinect device. Depending on the image geometry, parts of a scene may not be illuminated by the sensor's laser, i.e., the grid of infrared dots. These parts are therefore not captured by the infrared sensor, which results in empty regions in the depth image [32]. The second limitation is due to the point density of the Kinect device's sensor. Using its laser and depth sensor, the Kinect device generates a point cloud of triangulated depth measurements. The dimensions of the spatial area that are covered by the point cloud increase quadratically with distance from the Kinect device. Hence, the resolution of the depth images generated by the Kinect device decreases with the distance [32]. These two limitations result in noisy depth measurements. It calls for feature computation methods that are able to efficiently deal with the noisy nature of depth images.

1.4. Improving object detection in depth images

Shotton et al. suggested that a larger computational budget may allow for the design of "potentially more powerful features based on, for example, depth integrals over regions, curvature, or more complex local descriptors" [29]. Alternatively, studies seeking to improve object detection in depth images [34] can opt to use a larger computational budget to refine the input depth data itself, e.g., by including an advanced depth image filter and refinement techniques [35–38].

This paper proposes an improvement of Shotton et al.'s Pixel Comparison (PC) features by introducing advanced region-based descriptors, that do **not** require an increased computational budget: the Region Comparison (RC) features. Inspired by the work by Viola and Jones [8], Haar-like region features [7,39] are combined with the integral image representation [39] to detect transitions in adjacent regions of depth images. The RC features provide an indication of the direction and the extent of depth transitions in an area of a depth image by averaging over regions, i.e., large groups of pixels. The additional computational cost to calculate the surfaces of the regions is negligible when integral images are employed [40,41]. It is, however, unclear to what extent RC features enable fast and accurate body part detection in noisy depth images. To assess to what extent the RC features enable fast and effective body part detection in noisy depth images, we first define the *region comparison detector* which incorporates our RC features. Then, we compare its performance to a detector that deploys Shotton et al.'s PC features: the *pixel comparison detector*. In a comparative evaluation of the RC and PC features, both associated detectors are trained and evaluated on three quite different and challenging object detection experiments: two face detection tasks (with smoothed background and non-smoothed background) and a person detection task. There are two evaluation criteria. The first evaluation criterion is the classification performance, which is defined as the average per-class segmentation accuracy. The second evaluation criterion is computational efficiency, which is defined in terms of the time required to process an entire depth image. A shorter processing time therefore corresponds to a higher computational efficiency. It is assessed to ensure that improvements in accuracy do not lead to insurmountable computational costs that prohibit real-time operation. We consider the RC features superior to the PC features only when the detector incorporating the RC features outperforms the detector featuring the PC features on evaluation criterion 1 and performs equally well or better on evaluation criterion 2.

1.5. Related work

Our approach for improved detection accuracy in depth data deals effectively with background noise, without requiring additional computational power. It relates to several contributions in the fields of image refinement, computer vision and image understanding. In what follows, the related work is discussed, and – where appropriate – we describe how the work discussed inspired our research.

First, several approaches aiming to counteract background noise in depth data include advanced depth image filter and refinement techniques [35–37]. Although image refinement is likely to improve the quality of the input depth data, it comes at the cost of computational power. This may influence the detection time negatively. An interesting approach was presented by Fanello et al. [38] in the form of their 'filter forests'. Using location-dependent adaptive filters, their approach can be used to refine the quality of depth images. Such filters are computationally demanding and therefore not suitable for our goals. Inspired by their approach, our RC features incorporate a more straightforward – and computationally less demanding – way of filtering noisy depth images.

Second, Nanni et al. [42] aim to detect human faces by applying the well-known Viola–Jones detector [8] to visual (RGB – Red Green Blue) images. Aligned depth images are then used to validate