Using Depth Information to Improve Face Detection

Walker Burgin[*]
walkerburgin@wustl.edu

Caroline Pantofaru[†]
pantofaru@willowgarage.com

William D. Smart[*][†]
wds@cse.wustl.edu

[*]Washington University
One Brookings Drive
St. Louis, MO 63130
United States

[†]Willow Garage
68 Willow Road
Menlo Park, CA 94025
United States

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: [3D/Stereo Scene Analysis]; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*

General Terms

Algorithms

1. INTRODUCTION

Face detection is one of the most thoroughly-explored questions in computer vision, with applications ranging from consumer cameras to human-computer interaction [2]. It is also a vital skill for any robot designed to autonomously interact with humans. The faster and more reliable we can make face detection, the better our interaction with the people owning these faces will be.

The computer vision community addresses face detection from a monocular image- or video-centric perspective. Most algorithms are designed to detect faces using one or more camera images, without additional sensor information or context. A mobile robot, however, usually has multiple sensors in addition to its cameras, including sensors that can provide depth information. Laser range-finders, stereo camera pairs, the new Kinect sensor [4], and the SwissRanger camera [3] can all provide depth information of varying types and qualities. Additionally, a mobile robot often knows the exact positions of its sensors and, hence, can calculate the 3d positions of the objects that it senses.

All of these additional cues can be used to make face detection more efficient and, potentially, more accurate. In this paper we describe an extension to the classic Viola-Jones face-detection algorithm [8] that considers depth and situational information when searching for faces in an image. This work is similar to that of Dixon *et al.* [1], but extends it by performing a more rigorous performance analysis on a large data set [5].

2. THE COMPUTER VISION APPROACH

The Viola-Jones algorithm [8] involves exhaustively searching an entire image for faces, with multiple scales explored at each pixel. Given no other knowledge about the scene, an exhaustive search is reasonable, and the Viola-Jones algorithm is efficient. If video data are available, the basic algorithm can be made more efficient by tracking the detected faces, and searching only a local neighborhood around faces found in previous frames. Extracting more contextual information, however, is difficult from monocular imagery.

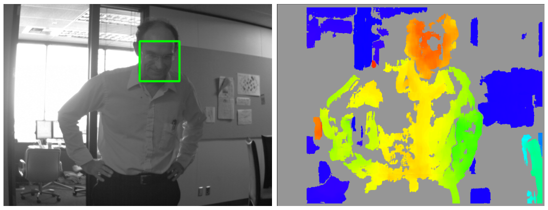

Figure 1: Image with face detection from [4] (left); Corresponding stereo-generated depth image (right, false-color). Grey pixels lack depth information.

3. DEPTH CUES

Many mobile robots are equipped with sensors that provide a depth image which can be calibrated to align with a monocular camera image, as shown in figure 1. Given the depth of a pixel, the plausible size of a face centered at that pixel can be calculated, and the face detector can be restricted to only look for faces of that size. Given a realistic range of human head sizes, depth information allows us to calculate the corresponding size in pixels at any point in the image. We can then restrict the scales at which we search for faces, potentially saving a lot of computation.

Distance thresholding can also be used to improve the efficiency of face detection. Areas of the image corresponding to points further than a given distance are likely to be blurry, or to contain too few face pixels for reliable detection. Since we cannot detect faces in these areas, even if they are present, there is no point in trying, and we can avoid running the face detector over these image areas.

Interestingly, denser stereo images, such as those generated by structured light systems [7] might not be better for speeding up face detection than sparser ones. Traditional stereo cameras provide sparse information. They rely on passively detecting texture in the images and, hence, are unable to give depth information in locations without texture, as in figure 1. Since the Viola-Jones algorithm uses image features to detect faces, it is unlikely to work in areas of the image with no texture. Traditional stereo camera systems will identify these areas (by not detecting texture), while more advanced, active systems will not (since they project their own texture on the world). This means that depth images generated from passive approaches are likely

HRI'11, March 6–9, 2011, Lausanne, Switzerland.
ACM 978-1-4503-0561-7/11/03.

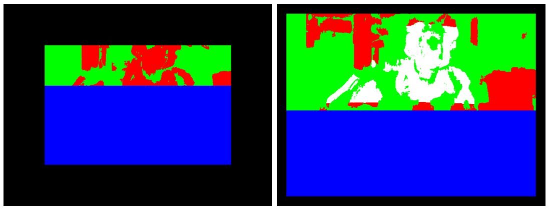

Figure 2: False-color image of pixels tested for faces with face window size of 197 pixels (left), and 46 pixels (right). The face detector is not run at: black (too close to margin); blue (too low); green (no depth information); red (wrong scale). The face detector is run at the white pixels.

to allow us to more aggressively prune the search space of the face-detection algorithm.

To summarize, depth cues can be used to constrain the search for faces by (1) constraining the face width at each pixel to realistic sizes; (2) thresholding the distance faces can be from the camera; and (3) only searching for faces in locations with texture and depth.

4. CONTEXTUAL CUES

Robots often have strong contextual information that can be used to constrain face locations in the camera image. The robot knows the direction its cameras are pointing in. When combined with depth information this determines the relative real-world position of points in the camera image, meaning that areas of the image unlikely to contain faces can be ignored. For example, there are unlikely to be faces close to the floor plane or ceiling, at least in normal offices.

We can also perform task-based filtering. For example, if the robot is driving down a hallway performing a delivery tasks, it can safely ignore faces far from its planned path, since they are likely to be irrelevant to its mission. Additionally, people in a hallway are generally standing, allowing us to further restrict the search space.

Finally, the robot can also keep track of its position in the world, allowing historical information about where people are likely to be to inform and prioritize face-detection.

5. PRELIMINARY RESULTS

We have implemented a filtered face-detector, using depth information to limit the areas in the image where we attempt to find faces. Figure 2 shows the areas of a typical image ignored by our algorithm. Notice that we determine which pixels to consider at each of the scales used by the Viola-Jones algorithm. We then run the face-detector at the appropriate locations and scales.

Figure 5 shows the results of our initial experiments, for three data sets drawn from the Moving People, Moving Platform dataset [5]. We compare our algorithm to the standard Viola-Jones implementation available in the ROS robot software architecture [6]. For all of the sets, our approach is faster than the standard algorithm. For two of the sets, it is considerably faster. For the first data set, the speed improvement is less marked. We believe that this is due to the nature of the images. In this set blank walls take up large portions of many images, causing the standard Viola-Jones face detector to terminate early in the cascade, and making it almost as efficient as our filtered algorithm.

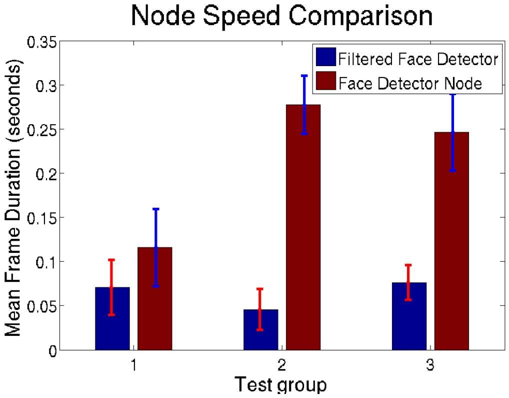

Figure 3: Mean per-frame times and standard deviations for our filtered face detector (blue) and the standard ROS [5] face-detector node (red) for three test groups.

6. CONCLUSIONS

The information available to a robot through a variety of sensors and contextual awareness is rich and unique. In this paper, we have argued that depth and context can improve frontal face detection, in turn improving the ability of robots to interact with humans, and supported this claim with encouraging preliminary experimental results. As future work, we will attempt to apply the same concepts to the much more difficult problem of detecting faces in profile, further expanding the population with which a robot can interact.

7. REFERENCES

[1] M. Dixon, F. Heckel, R. Pless, and W. D. Smart. Faster and more accurate face detection on mobile robots using geometric constraints. In *IEEE/RSJ International Conference on Robots and Systems (IROS 2007)*, pages 1041–1046, 2007.

[2] E. Hjelmås and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83:236–274, 2001.

[3] MESA Imaging. SwissRanger sensor. `http://www.mesa-imaging.ch`.

[4] Microsoft Corporation. The Kinect sensor. `http://www.xbox.com/en-US/kinect`.

[5] C. Pantofaru. The Moving People, Moving Platform Dataset. `http://bags.willowgarage.com/downloads/people_dataset.html`.

[6] M. Quigley, K. Conley, B. Gerkey, T. B. Faust, Josh an d Foote, J. Leibs, R. Wheeler, and A. Y. Ng. ROS: An open-source robto operating system. In *ICRA 2009 Workshop on Open Source Software in Robotics*, Kobe, Japan, 2009.

[7] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition (CVPR 2003)*, volume 1, pages 195–202, 2003.

[8] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.