Detecting and Following Humans with a Mobile Robot

Tim Braun\textsuperscript{1}, Kristof Szentpétery\textsuperscript{1}, Karsten Berns\textsuperscript{1}

\textsuperscript{1} University Kaiserslautern, Robotic Systems Group, Kaiserslautern, 67653, Germany

email: braun@informatik.uni-kl.de

Summary

This paper presents a method to detect and follow human subjects with a mobile robotic platform. By combining color vision and a laser range scanner, realtime face detection in color images becomes possible using a highly accurate, modified SVM. Successfully detected humans can then be followed by tracking their legs using the laser scanner.

Introduction

The development of service robots that are able to assist humans during their everyday tasks has become a popular research area over the last few years. Possible application scenarios for such robots include assistance for elderly people and general service robots for public areas or shopping centers. A key requirement for service robots is the ability to detect humans and interact with them in a non-technical, natural fashion.

In this paper, we focus on following a human as an important, basic interaction capability for a service robot equipped with a color camera and a laser scanner. In order to follow a person, both methods for the initial detection of humans and the subsequent tracking need to be implemented. The detection system presented here is based upon a modified Support Vector Machine that is trained to detect faces in the camera images. In order to work in realtime, several techniques are used to restrict the search space as well as speed up the general classification. The tracking system is activated after a successful detection and tracks leg positions in the laser scan. With this approach, the robot can even follow a human who does not constantly face the camera.

The remainder of this paper is organized as follows. After the presentation of related work in the next section, the actual detection approach will be described in detail. Afterwards, several results obtained with our mobile robot MARVIN (see fig. 1) will be presented and the main results summarized.

Related Work

There are several active research groups working on service robots with a focus on natural human-machine interaction. Most of the pursued approaches employ a multi-modal strategy that combines range sensors and color cameras in order to detect legs respectively faces of interaction partners. In many projects, additional hardware is used to improve the interaction capabilities, like for example microphones \cite{4}, pan-tilt units \cite{3}, omnidirectional vision \cite{5} or an artificial face for gesture expression \cite{3}. The robot \textit{BIRON} \cite{4} generates person candidates based on audible clues combined with static, rule-based leg detection from laser-scans and validates them by applying skin-color filtering and a pattern classification algorithm using haar-features to the camera image. \textit{Albert} \cite{3} uses a background subtraction method to dynamically detect moving legs and finds faces using a neural network. Human detection without using laser scanners is done by \textit{PERSES} \cite{5}. Here, an attentional system is realized using sound and motion cues from ultrasonic sensors, while the actual human detection is based on votes from skin color and head contour detection combined with a cascade-correlation neural network.
SVM-Based Face Detection in Realtime

In order to reliably detect faces from color images in the dynamic environment surrounding the mobile robot, a pattern recognition method is needed that can cope well with variable lighting and differing facial poses or expressions. The Support Vector Machine (SVM) has been shown to yield excellent face detection performance even in difficult conditions and thus seems to be a good choice in this context. However, the computational demands of a SVM-based face detection system are high compared to haar-based classifiers or neural-network approaches.

In order to reduce the computational complexity of a SVM-based face classifier, we have developed an approach that speeds up the face detection task using three complementing techniques:

1. A special, highly efficient **Sequential Reduced SVM (SRSVM)** is used for classification of image patches instead of a regular one.

2. Prior to the application of the SRSVM, its search space is reduced to image parts containing **face candidates**. These candidates are determined quickly using skin-color filtering and geometrical constraints.

3. The search space is further reduced by fusing range information from the laser scanner with the captured image. This yields distance information for each face candidate and is used to **restrict the scale** at which to look for faces.

The **SRSVM** has been introduced by Romdhani and Schölkopf [2] as a way to speed up regular SVMs. Key idea behind their approach is to reduce the number of support vectors (SVs) that need to be taken into account for classification. This reduction is achieved by replacing the set of support vectors of a normal SVM with a reduced set that contains a lot less SVs, but still defines approximately the same decision surface. In a second step, the SVM evaluation scheme is modified so that the reduced set can be evaluated sequentially. Consequently, each image patch is first classified using just one SV; additional SVs are only considered if this classification does not yield a sufficiently clear result.

To confine the application of the SRSVM to image areas that are likely to contain faces, **face candidate** areas are extracted from the input image prior to classification. These candidate areas are determined by applying a skin-color filter [1] to the color input image. Further processing by morphological smoothing, binarization and contour extraction yields outlines of potential faces. These outlines are then tested using several geometrical validation rules: 1. The contour bounding box (BB) must have a minimum size of 19x19 pixels. 2. The width of the BB must lie between its height and half of its height. 3. The relative amount of filled pixels in the BB
after binarization must be above 70 percent. Figure 2 shows an example for the face candidate generation.

Since the employed SVM can only classify image patches with a fixed size\(^1\), an input image containing faces of unknown sizes must be examined at several scales. To avoid this time consuming multi-scale analysis, the distance between the robot and the depicted human is calculated and used to directly determine the appropriate scale for the face detection. For this, the distance measurements made by the laser scanner built in the robots' trunk are projected onto the image plane with a precalculated projection matrix [6], yielding a distance estimate for each image column. It is assumed that the distances measured at the height of the laser scanner (10 cm above ground) are also valid for all higher points. This is an acceptable assumption for humans who stand straight. Figure 3 shows a captured image where the measured distances are visualized by bars - the longer a bar, the farther away the corresponding image column. In order to assign a distance to each face candidate, all distance measurements having x-coordinates spanned by the candidates BB are clustered by distance. For each cluster, the average distance is calculated separately and the lowest distance is taken as distance estimate for the face candidate.

To actually detect faces in realtime using the collected information, the extracted face candidates are scaled according to their corresponding distance measures. Then, the area occupied by the candidate BB is split into 19x19 subwindows and classified by the SRSVM.

**Human Tracking using a Laser Scanner**

Whenever a new face is detected, it is assumed that the mobile robot has to follow that person. The face detection module cannot be used for tracking, since the human will most likely turn around while walking. Therefore, the tracking is based on the laser scanner and works as follows: After the distances measured by the scanner have been transformed into scan points lying in a robot-centered cartesian coordinate system, the scan points are clustered by proximity and the cluster centers are calculated. Now, whenever a face candidate is accepted as a real face, the scan cluster that provided the distance information for that candidate is selected as the *human cluster* to be followed. In each subsequent update step, the human cluster position is set to the closest scan cluster available within a predefined distance limit.

Given the position of the human cluster in robot-centered coordinates, a tracking behaviour can easily be implemented. The tracking keeps a minimum distance between the robot and the leg cluster, so that the user does not feel annoyed by the robot. At the same time, it is guaranteed that the robot always faces the human with its camera in order to reaffirm itself whenever a face detection is possible.

**Experimental Results**

In order to test the presented approach for detecting and tracking persons, it has been implemented on the mobile robot MARVIN (Fig. 1). The basic SVM needed for the face detection system has been trained on over 20000 face examples and 102000 non-face examples which have been collected from various image databases. The initial training using a gaussian kernel with a width of \(\sigma = 0.04\) resulted in a SVM with 27267 support vectors. This set was subsequently reduced to 100 support vectors that formed the basis of the SRSVM. We have found that an image patch can be classified by the SRSVM with only 4 SVs on average, taking \(12\mu s\) per patch on a 1.6 GHz P4. Compared to a normal SVM, the resulting speed gain is immense. Nevertheless, the detection performance is only slightly degraded.

\(^1\)The image window used in our implementation is 19x19 pixels wide.
The search space reduction achieved by skin-color filtering and the incorporation of range
data is also substantial. While the number of image patches that would have to be examined
each frame without any reduction amounts to about 200000 (6 scales of a 320x240 pixel im-
age), the implemented approach does only test about 700 patches per face candidate. Since
in most frames, at most one or two candidates are present, this is again a big saving in com-
putational time. In total, the face detection part is able to process 3 fps while leaving enough
processing time for the motion and safety systems. Using the presented approach, the robot has
demonstrated its ability to successfully detect and follow a human as planned (see Fig. 4).

![Figure 4: A Human Detection and Tracking Scenario](image)

**Left:** A Person has been classified as candidate but not yet detected; **Middle:** Human cluster position has been set after successful face detection; **Right:** Robot follows person by leg tracking. **1.Row:** Original image with skin-color outlines marked; **2.Row:** Overhead view of the trajectories followed by the person (black) and the robot (red).

**Summary and Conclusion**

In this paper, a method to detect and follow human subjects with a mobile robotic platform has
been presented. It has been shown that by implementing a variety of time-saving strategies,
realtime face detection can be achieved using a Sequential Reduced SVM, making this an in-
teresting and powerful pattern classification method usable in mobile robots. The presented
approach has been validated by implementing it on a real robot. Future work on the project will
encompass more sophisticated leg cluster tracking and face candidate selection. Also, a rigor-
ous comparison with other face detection approaches in terms of speed and accuracy will be
conducted.