Person Tracking and 3D Model-based Face Analysis for Robust Human-Robot Interaction

Introduction

Recent advances in robotics research and the latest industrial trends, mainly termed as fourth industrial revolution or factory of the future, show that human-robot interaction will play a major role in our future lives. While robots evolve from isolated manufacturing machines to collaborative or assisting co-workers in industrial and domestic settings, the need for robust machine-interaction infrastructures grows – for verbal and non-verbal interaction alike. Apart from gestures, the human face is the most important element in non-verbal interaction, providing information such as identity, gender and emotion. While the task of analysing a face is considered to be solved for scenarios where head pose, lighting and overall image quality are controlled, uncontrolled conditions are still a great challenge.

In this report, we present our research activities which aim to minimise the impact of a subject's head pose on the performance of a face analysis framework. We illustrate our robot platform that involves the combination of an active camera positioning system, adaptive object tracking and a fast 3D model based pose normalisation for face alignment.

1 Robot Platform

We conduct our research on a SCITOS G5 mobile robot platform, a system that is widely being used in commercial and industrial applications. Apart from being able to navigate autonomously, it offers a Linux-based computing platform that is powerful enough to perform complex tasks in computer vision and pattern recognition. The human machine interface of the robot consists of a touch-based display, a speech synthesis system and a head with moveable, human-like eyes.

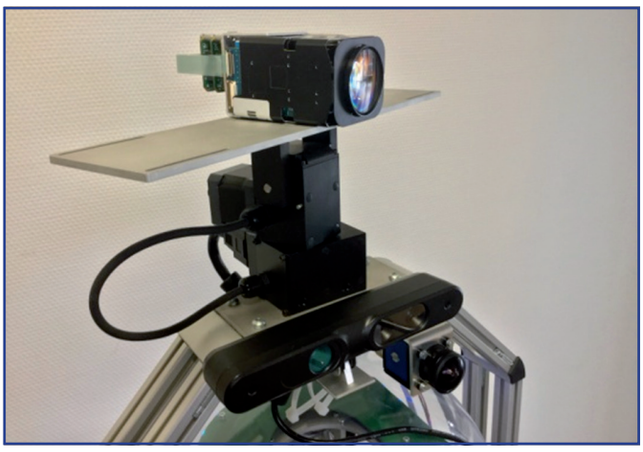

Figure 1: Pan-tilt-unit mounted on robot with automatic zoom lens camera.

We extended the system with a Directed Perception D46-17 pan-tilt-unit (PTU) that is mounted on top of the robot and a Sony FCB EV 7500 camera with a 4.3-129.0mm automatic zoom lens attached (see figure 1). This wide focal length range allows us to get a more detailed view on subjects at various distances to perform accurate face analysis.

2 Software Framework

2.1 Tracking

When a person interacts with the robot, it is essential to know where surrounding people are located within the scene and to be able to track them continuously. The tracking process is triggered and initialised by the face detection provided with OpenCV. We then apply an adaptive method that uses a particle filter to track position and size of the target and estimates the target motion using an optical flow based prediction model. Furthermore, a SVM-based learning strategy is implemented to calculate the particle weights and to prevent bad updates while staying adaptive to changes in object pose [1]. We do not only use the tracking for object detection but also to determine position deviations in order to keep the camera's focus on the subject by moving the pan-tilt unit.

2.2 3D Morphable Face Model Fitting

Some frames acquired by the tracking framework can have a face image that is challenging to analyse due to pose variations. The main issues of a non-frontal head pose (relative to the camera plane) are self-occlusion of facial attributes and the deformation of geometric relations – both being features that are crucial for a face recognition algorithm. In order to align the 3D orientation of a face that was captured in a 2D image frame, we fit a 3D Morphable Model onto the face. Such a model is in essence a vector space representation of faces (shape and texture) constructed out of real 3D scans and can be used to model continuous transitions (i.e. morphing) between different novel faces by adapting the model parameters [2]. The optimisation problem of adapting these parameters to a 2D image of a face is called fitting.

The complexity and dimensionality of the fitting's cost function depends on the amount and the type of image information that is being used, as well as the capabilities of the 3D Morphable Model itself. Existing algorithms that perform fitting on various aspects like shape, colour (albedo), lighting and camera parameters by performing not only shape-to-landmark fitting but also shape from shading often result in non-linear cost functions [3]. Minimising such a non-linear cost function is a task that runs in the order of minutes on a modern computer or robot and is therefore not suitable for real-time applications.

A less complex class of fitting algorithms only uses facial landmarks like the eye positions or mouth corners to fit shape and camera parameters, and directly extracts image information (pixel values) to model the face texture. The resulting cost function of this approach is linear and can therefore be solved in a matter of milliseconds. Computation time is a limiting factor for our robot application which operates on live video streams. For this reason, we focus on the fast fitting of a 3D Morphable Model based on landmarks.

2.3 Landmark Fitting

Similar to Aldrian and Smith [4], we obtain a linear solution by decomposing the problem into two steps which can be iterated and interleaved. First, a camera projection matrix is estimated using known 3D-2D correspondences between the given landmarks y_i found in the 2D image and the 2D projections of the 3D model $y_{m2D,i}$. In the current framework, we use automatically detected landmarks found by a cascaded regression based approach similar to Feng et al. [5]. In the next step, 3D shape parameters α_s can be estimated using the known camera projection matrix. A shape vector α_s is computed by minimising the reprojection error of the landmarks:

$$\arg\min_{\alpha_s} \sum_{i=1}^{3N} (y_{m2D,i} - y_i)^2 + \lambda \|\alpha_s\|^2 ,$$

where N is the number of landmarks and λ is a weighting parameter for the regularisation term that is needed to only allow plausible shapes. Subsequently, the camera matrix can then be re-estimated using the now obtained identity specific face shape, instead of using only the initial mean face of the model. Both the camera and shape estimation steps are iterated for a few times until a sufficiently small residual is reached.

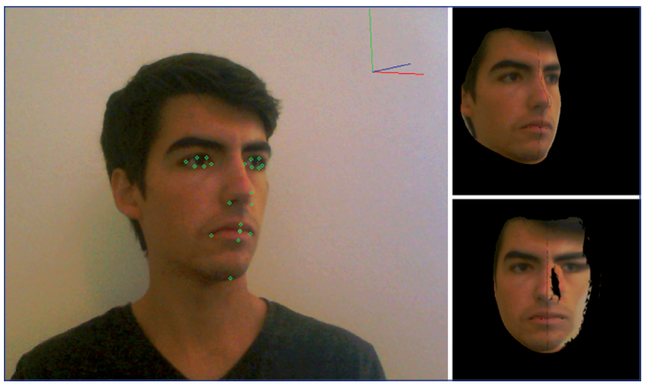

Figure 2: The workflow of our 3D Morphable Model fitting approach: Detection of facial landmarks, shape fitting and texture extraction, and pose normalisation.

2.4 Pose Normalisation for Face Analysis

After the shape fitting process, the texture from the input image is mapped onto the 3D model. The now obtained 3D model can then be moved around freely and can be aligned to a frontal representation for face analysis. Research in face analysis and our own experiments have shown that such a pose normalisation improves further image processing steps significantly [6]. Figure 2 shows example results of the current pose normalisation application from a live video stream. Although the self-occluded parts (especially on the right side of the nose) are still not visible in the frontal representation, the positions and geometric relations of facial features resemble those found in an image of a frontal face. It is therefore comprehensible that identification algorithms, which in general work on frontal images, are likely to perform better on the pose normalised image.

Figure 3: Schematic overview of the framework for human-robot interaction on our SCITOS G5 platform. The landmark-based 3D Morphable Model fitter acts as a pre-processing method for various face analysis tasks.

Applications in Human-Robot Interaction

On the robot, we use the frontal renderings to estimate the gender, age, emotion or identity of the human in conversations with the robot. These attributes of the human are used by the dialog manager of the robot to react most suitable to the person and the specific situation (see figure 3). Possible use cases for a robot of this kind are industrial mobile robotics (see figure 4), shop or museum tour guides or a surveillance system (e.g. in a supermarket). In the emerging market of ambient assisted living, a human-computer interaction system can further provide safety and supervision for elderly people. In all these cases, the robot has to interact naturally with technically unskilled people. It is therefore convenient to be able to analyse people independent of their pose, without requiring them to look at the robot directly. Identification, age and gender estimation on a frontal pose normalised rendering offer the possibility to make the robot's behaviour dependent on the interacting persons.

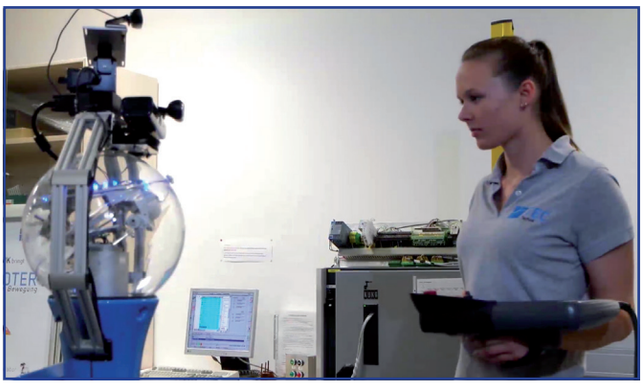

Figure 4: Direct interaction between our robot platform and a co-worker.

Summary

We presented our robot framework and our efforts to make face analysis more robust towards self-occlusion caused by head pose. By using a lightweight linear fitting algorithm, we are able to obtain 3D models of human faces in real-time. The combination of adaptive tracking and 3D face modelling for the analysis of human faces is used as a basis for further research on human-machine interaction on our SCITOS robot platform.

We want to thank the entire RT-Lions team at Reutlingen University for their support. Our face model fitting library as well as a low-resolution shape-only model are available at https://github.com/patrikhuber/eos.

References

[1] P. Poschmann, P. Huber, M. Rätsch, J. Kittler, and H.-J. Böhme, Fusion of tracking techniques to enhance adaptive real-time tracking of arbitrary objects, Conference on Intelligent Human Computer Interaction (IHCI), 2014.

[2] T. Vetter and V. Blanz, A Morphable Model for the synthesis of 3D faces, Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187–194, 1999.

[3] S. Romdhani and T. Vetter, Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 986–993, 2005.

[4] O. Aldrian and W. A. Smith, Inverse rendering of faces with a 3D Morphable Model, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 5, pp. 1080–1093, 2013.

[5] Z.-H. Feng, P. Huber, J. Kittler, W. Christmas and X.-J. Wu, Random cascaded-regression copse for robust facial landmark detection, IEEE Signal Processing Letters, volume 22, pp. 76–80, 2015.

[6] R. van Rootseler, L. Spreeuwers and R. Veldhuis, Using 3D Morphable Models for face recognition in video, 33rd WIC Symposium on Information Theory in the Benelux, 2012.

Autoren

Philipp Kopp
Reutlingen University
philipp.kopp@student.reutlingen-university.de

Michael Grupp
Technische Universität München
michael.grupp@tum.de

Patrik Huber
University of Surrey, United Kingdom
p.huber@surrey.ac.uk

Prof. Dr. rer. nat. Matthias Rätsch
Reutlingen University
matthias.raetsch@reutlingen-university.de

Hochschule Reutlingen
Alteburgstraße 150
72762 Reutlingen

www.reutlingen-university.de