RT Lions Team Description Paper

Tobias Gerlach, Tobias Oschwald, and Prof. Dr. M. Rätsch, Felix Ostertag, Thomas Weber, Patrik Huber, Jörn Hoffarth, Robert Hülsmann, Marc Katzenmaier, Philipp Kopp, Sascha Braun, Michelle Foo, Jan Engling, Michael Litz, Sarah Roericht, Tobias Huschitt, Carsten Gedrat, Robin Keinert, Atmaraaj Gopal

RT Lions
Reutlingen University
Alteburgstraße 150
72762 Reutlingen, Germany
teammail@rt-lions.de
http://www.rt-lions.de

Abstract. Our team's main research line is image processing. We use it for a wide range of applications, eg. face analysis, face detection, object detection and tracking in many forms. Last year, we were able to make a lot of progress with the convolutional neural networks. Our main research was done on "Face Models for Face Recognition" and "Active Closer Inspection System".

1 Introduction

This paper describes the RoboCup@Home team, RT Lions, of Reutlingen University, Germany, for the participation in RoboCup@Home 2017 in Nagoya, Japan. RT Lions is founded in 2009, and since then, our team has carried out numerous projects and researches in various fields.
Since the team has done a wide range of researches, it's no easy task to break it down to only the most important ones.
Nevertheless, the researches done are explained here, focusing on the most recent work.

1.1 The Team

We are the RT Lions. A German Robocup Team from Reutlingen University.
Our team consists primarily of students. Some of us for scientific research, and others just for fun and experience in our spare time.
Since we are varied, from students of the first semester of the Bachelor's Degree upto PhD, we have quite a diverse team in terms of experience and expertise.
We are more than 5 PhDs, 6 research assistents and about 15 students.
Nevertheless, every one of us have tasks of our own, considering both ability and interest.
Our main platform is a MetraLabs Scitos, named "Leonie".

We are happy to name companies like Cognitec, Bosch and Daimler, our industrial partners. Furthermore, we have several industrial contracts and funding programs.
So far we have published more than 50 journals and conference publications.
We have also accumulated a decent amount of experience from competition participations and victories. We were World Champion 2009 in Graz, Austria, German Master in 2009, Vice World Champion 2010 in Singapore, Iran Master 2011, 1st place Informatics Inside 2014 and 2015 and finally 1st Place Freebots League at XVI Portuguese Robotics Open 2016 (Branganca).

2 Research

2.1 Face Models for Face Recognition

Face Modelling for Pose Normalization In this project, pose normalization is reviewed as a method to enhance the preprocessing part of a face recognition system. The orientation of an object in a three-dimensional space can be described by using the Euler angles roll, pitch and yaw. A frontal image of a face, like described in 1.1, has no rotation components relative to the camera coordinate system. Per convention, the Euler Angles in this normalized case are zero. Pose Normalization aims to reconstruct such frontal representations even with non-frontal images by the use of algorithms. One approach for obtaining a pose normalized, frontal face image is through fitting a 3D face model to the 2D input image. In the fitting process, the orientation of the input image's face is applied to the model and, depending on the method, also an optimization of shape and texture characteristics. The next step is setting the roll-, pitch- and yaw-angle

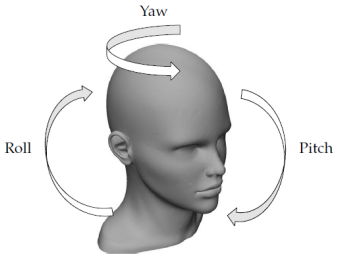

Figure 3.1: Euler angles: roll, pitch, and yaw

of this 3D model head (see figure 3.1) to zero by turning the 3D model to the normalized position. The fitted model, which is now frontal, can then be rendered back to a 2D image, which can be used by the face recognition engine [1] [2].

3D Morphable Model The 3D morphable model (3DMM) is a face modelling technique that can be used for quickly generating 3D models of faces. It was first proposed by Thomas Vetter and Volker Blanz at the University of Tübingen [3]. Various publications havebeen made in this field, describing new methods for

obtaining a 3D morphable model, although the basic concepts remain similar [4]. The 3DMM consists of three-dimensional meshes of real face scans that have been registered to a reference mesh and a texture map. The initial version of this model represents the average face out of the real faces that have been used for the generation of it. Deviations of this reference mesh, novel unique faces, can be calculated by changing the shape and texture parameters of the model [8]. Mathematically, the resulting face is a linear combination of the registered real faces. Parameters of a 3DMM include the model parameters for shape, α and texture, β, and a set of projection parameters, ρ. The projection parameters "include 3D rotations and translations, and the focal length of a virtual camera".

Fitting The morphable model can be used to reconstruct a 3D representation from a 2D image through fitting. Fitting is the process of adapting the 3D morphable model in such a way, that the difference to a 2D input image is minimal. Given an input image, I and a set of facial landmarks, an initial guess of the model's parameters is calculated. Then, a cost function iteratively minimizes the difference between the modeled image, I_m and the original input image, I by adapting the shape parameters, α and β, and the projection parameters, ρ for each model vertice, k:

$$\alpha, \beta, \rho = \arg\min_{\alpha,\beta,\rho} \sum_{\forall k} ||I_m(k; \alpha, \beta, \rho) - I(x_k, y_k)||^2$$

This method is called analysis by synthesis, because in each iteration, a new fitting is produced and compared with the original image until the difference between model and original input is sufficiently small. After fitting, various conditions can be changed, for example, the illumination and pose of the face model [4].

Rendering The process of generating a two-dimensional representation from a three-dimensional model is called rendering. Because most face recognition engines are optimized for the use of 2D images, the pose normalized 3D morphable model needs to be transformed back to a new 2D image in order to be used for face recognition purposes. Then, a perspective transformation maps the vertices and the texture components of the 3D model to the 2D image plane, to produce the rendered image.

2.2 Active Closer Inspection

Face recognition plays an increasingly important role for many applications. For example, surveillance system, consumer marketing (discovery of potential customer and advertising). For these systems, the analysis of mimic, age, gender and movement of face actions are necessary. Therefore, face analysis is one of the critical issues for these applications. High resolution images are required for good quality face analysis. Furthermore, it is important that this camera system can

cover a range as wide as possible with the least movement of the camera. Since we don't have a camera with wide viewing angles and a very high resolution, a system with two specific cameras and a pan-tilt-unit (PTU) is necessary. A wide-angle camera should scan the surrounding for faces. When it finds a face, the PTU will pan and tilt a zoom camera to the target. For recognition, it may be essential to zoom in to the face [5] [6] [7] [8].

Tracking Approaches

Single Camera Tracking The simplest tracking system requires only one camera. All necessary tasks (detection, tracking and recognition) are executable with this camera. To track a larger area the camera is mounted on a pan tilt unit. For a more accurate recognition of faces with a large distance to the camera system we use a zoom camera. So we can zoom in on distant faces for better results for face recognition. The advantage is, after face detection, we only need a simple control which checks the eccentricity of the face in WaC stream and adjusts the angles of the PTU. The disadvantage is that the movements of the PTU are necessary to include all areas around Leonie.

Dual Camera Tracking To compensate the disadvantage of the single camera tracking, we need a second special camera. This camera should cover a large area. For that a wide angle camera is useful. It's mount statically and is independent of the pan tilt zoom camera movements. So we can scan a huge tracking area without PTU movements, but the resolution of the wide angle cameras is insufficient to recognize faces credibly. Therefore, we have to convert the pixel-applied position information of the static wide angle camera to pan and tilt angles for the PTU and a zoom level for the zoom camera. The conversion must be very accurate because this tracking system comprises no possibility to double-check and readjust the result on the PTZ camera. For example, if the conversion to angles for PTU is incorrect and the detected face on WaC is not centered in the PTZ camera stream, then face recognition is impossible. The most error-prone component of calibration the system for conversion is the distance between human and camera system. But if the conversion is correct, we can use the high-resolution of PTZ camera for face recognition.

Combined Dual Cam Tracking This tracking system combines the advantages of both systems above for effectual results. We use two cameras in the same build-up as dual camera tracking. For face initialization, we need dual cam tracking to move PTU at the right position, for further tasks after detection and conversion, we use the features of single camera tracking. Therefore we can track faces even more accurately in comparison to the simple dual cam tracking. So this system needs two tracking: Once on the WaC for the time between the first face detection on WaC and recovering on PTZC and a second tracking on PTZC after recovering. While tracking a face on PTZC, we need the tracking on WaC

to detect other faces in the surroundings. The disadvantages in relation to the other systems sare that this system needs the most programming expense and process resources.

Proposed Tracking System Approach The original goal was to build these approaches one upon the other. Given that the single cam tracking had been realized in previous project stages, the dual cam tracking would have been the next step. However, during work on this tracking type, it became evident that the Combined Dual Cam Tracking was the easier and more reliable of the two. This is mainly due to the fact, that, to get depth information from the dual camera video streams, the desired object needs to be present in both images. However since the PTU mounted camera may not yet be oriented in the right direction, a rough estimate must be used for the distance. Given that the accuracy of such an estimation is rather bad, the zoom factor of the PTU mounted camera would have to be set somewhat lower to guarantee the persons face is inside its field of view. At this point, further adjustment of the zoom level could also be simply done by relatively comparing the tracking box size to that of the image. This also makes the recalculation of the relative camera positions, as well as the internal camera parameters, which are needed for depth estimation, unnecessary.

3 Hardware

Scitos G5 For Leonie, we used the robot Scitos G5 developed by MetraLabs Robotics. It combines industrial and research properties in one robot. Means that it has the advantages of an industrial robot, such as robustness and longevity, with the mobility and flexibility of a research robot. With its weight of 60 kg it can move at a speed up to 1.4 m/s and handles payloads up to 50 kg without any difficulties. It contains one lithium iron phosphate battery, which allows it to operate up to 20 hours of normal usage.

Next Unit of Computing We installed two Mini PC Intel NUC Kit NUC7i7BNH on Leonie. With them, we practice the Divide and Conquer design pattern, by capsuling programs from the Brain to the NUCs. We use two because of the rather large number of programmers working on Leonie and thus, to allow flexibility in terms of preferred operating system.

Pan-Tilt Unit To achieve a large range of visibility with suffiecient details, we use two different cameras and a pan-tilt-unit (PTU). A wide-angle camera scans the surrounding for faces. If it finds a face, the PTU will pan and tilt a zoom camera to the target.

Leap Motion To achieve advanced communication for people with handicapped speech, we are working with a stereo IR camera, for the purpose of hand gestures.

4 Software

4.1 B.R.A.I.N.

Best Reutlingen Artifficial Intelligence Network is our central control module. It combines Deterministic Finite Automation (DFA) with Fuzzy Logic . It acts as an Interface to all sensors and actors.

4.2 Emofani

Our animated face was designed and rigged in Blender and Blend-Tree in Unity3D, by a former team member. It simulates human like behaviour: breathing, blinking, mouth, speach animation, gaze, random eye movements and mico movements. While servo turns the head, eyes and face move back. The programm is not only used in several industrial projects, but also open source (https://github.com/steffenwittig/emofani) [9] [10] [11].

4.3 Hand Gesture Detection

We have two different projects in this topic, which are both at use on Leonie. The first one involves the Leap Motion API and is used for simple detections on Leonies front.
The second one was programmed by using caffe, a deep learning framework, for more complex detection from Leonies head.

4.4 Speaker Detection

This programm was based on Prime Sense. It uses a MS Kinnect 2, for detection of the angle of all bodies and its microphone array for detection of the angle towards a source of noise. In combination of both features it has a great confidence in detecting people in its view.

4.5 Speech Recognition

Our speech recognition module consists of 3 parts: The Google Speech API, the Stanford Parser and a Decision Maker. The Google Speech API is used to convert the input audio to text. Powered by Machine Learning, this Speech API is one of the best speech recognition technology available online. With the help of the Stanford Parser, we are able to tag the words in the incoming string. The tagged string is then passed onto our self-programmed Decision Maker, which picks out important keywords and decides what the robot should do. Such as carry out a command, answer a question or just interact with the person. For answering general knowledge questions, we integrated the Wolfram Alpha API in our answering questions program.

4.6 Navigation, Localization, Mapping, WaypointVisitor

To know its environment, Leonie uses its two laser sensors located on its front and back and its odometer. These are used to detect obstacles in its path as it moves autonomously towards its goal. The goal setting is done on the MiraCenter Software, which provides control and status of Leonie's sensors and odometer.

The mapping is done with the software or optionally, when needed, also with image editing software, like GIMP. The uniqueness of the mapping implementation in MiraCenter is the multi maps structure. The mapping is done with data from 4 separate .png map files: "static", "nogo", "speed" and "one way". Each map type has its own function which enables flexibility in controlling Leonie's movements with the maps, without needing to execute special commands.

Combining these two abilities, we have the WaypointVisitor where Leonie can learn multiple coordinates and respective orientations on a built map.

4.7 Follow Me

We use a monocular camera system on our mobile robot platform to track heads. Based on this information the robot platform follows the person. For detecting the heads a Single Shot MultiBox Detector CNN (SSD) [12] was used which was trained for our task. Based on the position of the bounding box a single object tracking based on the combination of a Kalman Filter [13] and global nearest neighbor data assignment [14] was used to follow the selected head. The position and size of the head is used to follow the person based on a proportional control system.

5 Conclusions and Future Work

We are always focused on surpassing our own expectations and expand as far as possible.In the future, we will put a lot of effort in improving our ability to pick and place objects. We have recently acquired a Sawyer by Rethink Robotics for that purpose and we are confident to get amazing results for challenges in the coming years.

6 Robot Leonie Hardware Description

Specification are as follows:

- Base and Torso: Scitos G5 developed by MetraLabs Robotics. Combining industrial and research robot, 1.4 m/s max speed. Handles payloads of up to 50 kg.
- Contains one lithium iron phosphate battery, which allows to operate up to 20 hours of normal usage.
- Two Mini PC Intel NUC Kit NUC7i7BNH
- ACI:
  - Pan Tilt Unit - PTU-D46-17
  - Wide angle Camera - SVS-VISTEK SVCam-ECO
  - Zoom Camera - Sony FCB-EV7500
- Microsoft Kinect 2
- Mikrofon Senheiser MKE 600
- Head: Beamer that project the face of Leonie on a plastic bullet
- Robot dimension: height: 1.90 m, width: 0.7 m, depth: 0.9 m
- Robot weight: 80 kg

7 Robot Leonie Software Description

For our robot we are using the following software:

- Platform: Windows 10 and Ubuntu 14.04
- Navigation, localization and mapping: MiraCenter and CogniDrive
- Speech recognition: Google StT and Sphinx
- Speech generation: Ivona TTS
- Object recognition: own software, based on caffe
- Face recognition: own software, based on caffe
- other generation/recognition own software

References

1. P. Huber, "Bottom-up and top-down face analysis based on 3d face models," *Interaction Design*, p. 1.
2. P. Huber, P. Kopp, B. Christmas, M. Ratsch, and J. Kittler, "Real-time 3d face fitting and texture fusion on in-the-wild videos," *IEEE Signal Processing Letters*, 2016.
3. V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194, ACM Press/Addison-Wesley Publishing Co., 1999.
4. J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas, "3d morphable face models and their applications," in *International Conference on Articulated Motion and Deformable Objects*, pp. 185–206, Springer International Publishing, 2016.
5. 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," *Procedia Computer Science*, vol. 39, pp. 162–165, 2014.
6. P. Kopp, M. Grupp, P. Huber, and M. Rätsch, "Person tracking and 3d model-based face analysis for robust human-robot interaction,"
7. M. a. K. Florian Fritz, Robert Hülsmann, "Active closer inspection," tech. rep., Reutlingen University, 2015.
8. P. Kopp, M. Grupp, P. Poschmann, H.-J. Böhme, and M. Rätsch, "Tracking system with pose-invariant face analysis for human-robot interaction," *Interaction Design*, p. 9, 2015.
9. S. Wittig, U. Kloos, M. Rätsch, J. Mun, J. Kim, D. Won, M. S. Laghari, S. Kaushik, S. Tiwari, C. Agarwal, *et al.*, "Emotion model implementation for parameterized facial animation in human-robot-interaction.," *JCP*, vol. 11, no. 6, pp. 439–445, 2016.
10. S. Wittig, M. Rätsch, and U. Kloos, "Parameterized facial animation for socially interactive robots.," in *Mensch & Computer*, pp. 355–358, 2015.
11. S. Wittig, U. Kloos, and M. Rätsch, "Animation of parameterized facial expressions for collaborative robots," *Interaction Design*, p. 11.
12. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, pp. 21–37, Springer, 2016.
13. G. Welch and G. Bishop, "An introduction to the kalman filter," 1995.
14. P. Konstantinova, A. Udvarev, and T. Semerdjiev, "A study of a target tracking algorithm using global nearest neighbor approach," in *Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech'03)*, pp. 290–295, 2003.