Robust Perception of an Interaction Partner Using Depth Information

Salah Saleh, Anne Kickton, Jochen Hirth and Karsten Berns
Robotics Research Lab, Dept. of Computer Science
University of Kaiserslautern
Kaiserslautern, Germany
Email: {saleh,a_kickto,hirth,berns}@cs.uni-kl.de

***Abstract*—Social interactive robots require sophisticated perception abilities to behave and interact in a natural way. The proper perception of their human interaction partners plays a crucial role. The reduction of the false positive rate for human detection is very important for increasing the natural interaction abilities. This paper presents a combined method using RGB data as well as depth information to find humans in the robot's surrounding. To track a person over time a Kalman filter is applied, which also reduces the processing time. Furthermore, a head pose estimation on the basis of Support Vector Machines is integrated, which can be used to perceive nonverbal expressive cues like nodding. The proposed method is tested in various experiments.**

***Keywords*-social robots; perception; human-robot interaction;**

I. Introduction

Inspired by an ever older growing population in recent years the development of service and assistance systems for applications like elderly care, nursery, or entertainment and edutainment gained an enormous importance in the field of robotics. Therefore, they need special interaction capabilities that enable socially acceptable behavior. The University of Kaiserslautern is developing a humanoid robot ROMAN for investigating social human-robot interaction, Figure 1. Its control architecture is realized based on psychological principles to enable intrinsic motivation to fulfill specific tasks as well as emotional reactions depending on the progress of the current interaction [1]. To test and evaluate the behavior of the robot, it has been enabled to handle various interaction scenarios.

Interactive robots need the ability to realize complex combinations of speech, gestures, facial expressions, and body poses. One of the main aspects for the realization of human-robot interaction scenarios is a stable perception of the human interaction partners. The challenging task of the perception system is to analyze and understand sensors data. One of the most important sensors to fulfill this task is the vision system, which provides a large data stream. The interactive robot has to analyze this stream of data to understand the surrounding environment as well as the interaction partner. Starting with face detection, the present work proposes head pose detection and a tracking using RGB and depth information. There are many challenges in finding faces in a stream of images due to the variation in scale and orientation, pose, facial expressions, and light conditions [2]. Most of the research works focused on using two-dimensional images. Because of the limitations of the 2D features to describe the reality, the faces detected using only RGB images do not always represent real faces.

Figure 1: The humanoid robot ROMAN of the University of Kaiserslautern playing a question-answer game.

In recent years, most researchers focused on face detection using depth information. *Niese et al.* [3] have applied a model-based matching for the study of facial features and the description of their dynamic changes in image sequences. Their face detection is based on color driven clustering of 3D points derived from stereo image sequences. The face detection and normalization method consists of six steps. In the first step they calibrate the stereo camera. In the second step they generate a surface model of the individual face from an active stereo scan and save it in the database. In the third step, they process sequentially the stereo image sequence with a passive stereo algorithm. The fourth step localizes and post processes the face with help of 3D and color information. In the fifth step, they determine the face pose on the basis of an Iterative Closest Point algorithm (ICP) that finds the matching between the calculated surface model from the database and the post-processed data from stereo data. The last step creates a synthetic image of the face in frontal pose and standardized size. *Burgin et al.* [4] believe that face in video stream can be made more efficient by tracking the detected face. This fact minimizes the search space in each frame by searching only a local neighborhood around faces found in previous frames. They also utilize the depth of each pixel to calculate the possible size of the face centered on the pixel. They restrict their search for the faces of specific size in each depth. *Lu Xia et al.* [5] proposed

ISBN: 978-1-61208-250-9

a method for human detection using depth information by Kinect. Their method is a model based approach. It detects humans using a 2D head contour model and a 3D head surface model. They proposed a segmentation scheme to discriminate the human from the background. They used a two-stage head detection process. The first stage explores the boundary information to locate the candidate region that may contain a human. In this stage they used 2D chamfer distance matching. This matching scheme scans the whole image to give the possible regions that may contain a human. These regions are then examined using a 3D head model using the relational depth information of the array for verification. The parameters of the head are then extracted from the depth array and used to build a 3D head model. The second stage matches the 3D model against all the detected regions to make a final estimation. They have also developed a region growing algorithm to find the entire body of the human.

An overview of different approaches for head pose estimation can be found in [6]. These approaches can be categorized depending on the type of data they used. Most research focused on using 2D images, while some of them focused on using depth information.

Seemann et al. [7] have proposed a neural network-based system using depth information. They have used a variation of color-based face detection techniques in addition to the depth information for fast and reliable face detection. They have used a three layers feed-forward neural network trained by back-propagation algorithm on a depth information obtained by the stereo camera.

Breitenstein et al. [8] have presented a real-time algorithm to estimate the 3D pose of a previously unseen face from a single range image. They have generated an average 3D face model from the mean of an eigenvalue decomposition of laser scans of 138 adults. They start with finding the nose tip and its orientation to roughly estimate the head pose. They used a 3D shape signature which is computed for each pixel to identify noses. The 3D positions and mean orientations of the nose candidate pixels form a set of head pose hypotheses. They used an error function in evaluation the alignment of reference pose range images and the input range image to estimate the pose. This process is performed in parallel on the GPU.

Fanelli et al. [9] have proposed a real time head estimation with random regression forests using depth data. Their training examples consist of range images of faces annotated with 3D nose location and head rotation angles. They assumed that the head has been already detected in an image. They constructed each tree of the forest from a set of fixed size patches randomly sampled from the training examples. Each patch consists of the extracted visual features associated with the pose parameters of that patch.

This paper presents an approach using depth information to overcome the problem of inaccuracy in human perception. The depth information is used in addition to RGB images to verify the detected faces – e.g. it can be distinguished from a photo. This improves the detection process by decreasing the false positive rate. The Kalman filter is used for face tracking as well as for decreasing the process time of the face detection. It predicts the position of the face in the next frame to reduce the search space. The goal of the face detection and tracking process is to follow the human face and record its movements to determine the persons emotion state.

The remainder of this paper is organized as follows: In Section II some of related concepts will be presented. Section III provides more detailed information on the implementation of the proposed system. Some experiments will be presented in Section IV. In Section V a conclusion and future work will be discussed.

II. CONCEPT

In addition to its cameras, a humanoid robot usually has multiple sensors especially those that provide depth information. These additional sensors can be used to make the detection of human beings, and by doing so also the interaction, more efficient and more accurate. This paper presents an approach using depth information in addition to the RGB images to perceive the interaction partner by detecting his/her face position, orientation, and movements.

A. Face Detection

One of the essential skills of the robot interacting with humans is face detection. In computer vision terms, the face detection task is not easy even though that humans can do it effortlessly [10]. The goal of face detection is to determine whether or not there are any faces in a given image, and if present, returns the location and size of each face [2]. It is a very challenging task for the robots, and has been one of the most studied research topics in the past few decades. The difficulty associated with face detection can be attributed to many variations in scale, location, orientation, pose, facial expression, lighting conditions, occlusions, etc. A faster and more reliable face detection process is a basic condition for proper human-robot interaction.

The face detection field has made significant progress in the last 20 years. One of the most important works in this field is the novel work of Viola and Jones [11]. Most of the face detection approaches have focused on the use of two-dimensional images without any additional information [2][10]. The additional sensors found in most robots such as depth information sensor can provide additional information that the robot can use in efficient and accurate face detection [4]. Depth information has several advantages over 2D intensity images. It is simple representations of 3D information in addition to its robustness to the change in color and illumination [5].

The present paper uses the RGB image and depth information provided by Kinect sensor. The detection process is accomplished in two stages. The first stage uses Haar cascade classifier to detect faces (candidate faces) in 2D image. The second stage takes the candidate faces detected by the first stage and uses depth information to verify if they are really faces or not.

B. Head Pose Estimation

In addition to speech, people have the ability of interacting nonverbally using different aspects. One of these nonverbal aspects is the human head movement. Humans have the ability of interpreting these movements quickly and effortlessly, while it is regarded as a difficult challenge in computer systems and robotics. Detecting the human head movement requires estimating the head pose (position and orientation) over the time. In a computer vision context, the process of detecting the pose of a human head from digital imagery is called head pose estimation. In order to build a robust human-robot interaction system, a robust and reliable head pose estimation algorithm is needed.

In addition to the sensitivity to illumination, the head pose estimation in 2D images suffer from the lack of features due to occlusion (in some poses) [8]. The depth information is used in the last years and it displayed encouraging results. The present paper uses the depth information in head pose estimation. Three Support Vector Machines (SVMs) for regression are trained to detect the pose angles (pitch, yaw and roll). After detecting the face position, the pose estimator applies the depth features of the face to the SVMs to estimate the pose.

C. Face Tracking

The face tracking is another important aspect of human-robot interaction in order for designing a robust interaction. In this paper, the Kalman filter [12] is used for face tracking. The kalman filter also used to speed up the face detection process and to give more confidence to the pose estimation. The prediction of the face position, size and orientation in the next frame enable us to search within a specific area of the frame. This reduces the detection time tremendously.

The Kalman filter is an algorithm which uses a series of measurements observed over time and produces estimates of unknown variable. The motion state of the face can be formulated by the following state model:

$$x_k = Fx_{k-1} + Bu_k + w_k \tag{1}$$

Where

- x_k is the state vector in time k.
- F is the state transition which is applied to the previous state x_{k-1}.
- B is the control input model which is applied to control vector u_k.
- w_k is the process noise.

At the time k, a measurement z_k of the state x_k is calculated according to the following model:

$$z_k = H_k x_k + v_k \tag{2}$$

Where

- z_k is the observation (measurement) vector.
- H_k is the observation model which maps the true state space into the observed space.
- v_k is the observation noise.

Based on the above two models, the state vector along with its covariance matrix can be updated to predict the next position and orientation of the face.

III. Implementation

The perception of human using 2D images encounters serious problems because its sensitivity to illumination and shadow. The recent 3D acquisition systems could help overcoming these problems. Schmitz and Berns [13] have suggested a communication partner model. They have used auditory and visual perception system. The present paper proposes a robust visual perception system for human-robot interaction using depth information. The system uses depth information in addition to the standard 2D images in perceiving the partner. The system uses Kalman filter to track the face position and orientation to keep eye contact as well as reducing the calculation time. An overview of the system modules is depicted in Figure 2.

In training phase, a large database of 15K, 640x480 range images of faces for 20 persons has been used [9].

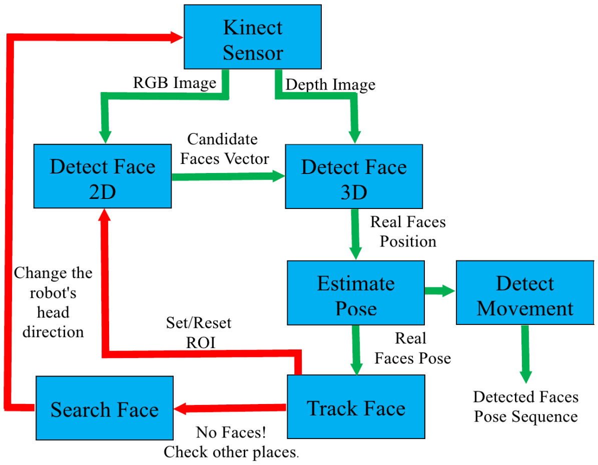


Figure 2: The proposed system block diagram.

A. Face Detection

The face detected by Haar cascade classifier does not always represent a real face. This is because that this method uses 2D features that can be found in any variations of intensities (colors) that may similar to a human face. Also it detects each face in a picture as a real face what represent a problem at the level of human-robot interaction. This problem can be overcome using depth information. The faces detected using cascade classifier can be checked again to determine if they are really faces or not. This can improve the detection process by decreasing the false positive rate.

In this paper, the 2D detection module uses multi-appearance cascade classifier to detect the face. Then the 3D detection module verifies the detected faces and decides which of these candidate faces are real faces and which are not. In this context, the detected face using the cascade classifier referred to as *candidate face*. The input of this module is a list of candidate faces with their positions. Its output is a list of detected faces with less false positive rate. The verification process examines each of the candidate faces. The corresponding depth information of each candidate face is used rather than the RGB information.

The candidate face is simply rejected if it doesn't meet one of the following criteria:

1) The relation between the depth (the distance from the face to the camera) and the face size must be governed by the following equation as presented by [5].

$$h = p_1.z^3 + p_2.z^2 + p_3.z + p_4 \tag{3}$$

Where

- z is the depth of the center of the face,
- h is the size of the face,
- $p_1 = -1.3835 \times 10^{-9}$
- $p_2 = 1.8435 \times 10^{-5}$
- $p_3 = -0.091403$
- $p_4 = 189.38$

If the candidate face doesn't satisfy the above formula then the face will be rejected and considered as not real.

2) The candidate face must have a reasonable depth to be a real face. The low depth or high depth face may represent a picture of a face or a color variation in the image. The face depth is calculated as the difference between the maximum and minimum distances after removing the outliers.

$$facedepth = max\_depth - min\_depth \tag{4}$$

If the depth of a candidate face is less than a predefined threshold or greater than a predefined threshold then the face is rejected and considered as not real.

3) The face must not be included in another face. If there is such case, then the face nearest to the criterion 1 is regarded and the other is omitted.

B. Head Pose Estimation

Head pose estimation is an essential skill that is needed in human behavior analysis. Estimating a head pose using depth information has shown very reasonable results. The proposed work uses depth information in head pose estimation.

Three SVMs for regression [14] are used as head pose estimators for the three angles *pitch*, *yaw* and *roll*. These SVMs are trained to find the pose angles from a set of depth features. These features are calculated after normalizing the depth information of the face. By dividing the face into set of rows and columns and omitting the borders, these features are the average depths of the regions lie on these rows and columns.

After calculating the depth features of the face, the Principal Component Analysis (PCA) is performed to find the most important features. This process reduces the problem dimensionality and speeds up the training of SVMs. Then the SVMs are trained using the resulting features from the PCA.

When a new face is applied, the depth features of the face are calculated and the most important features are derived using the eigenvectors to be supplied to the SVMs.

C. Face Tracking

Human face tracking is important aspect in order of natural interaction. A Kalman filter is used for human face tracking. It also decreases the process time of face detection. It predicts the position and orientation of the face in the next frame to reduce the search space. It uses twelve variables as state vector components. The six variables (x, y, z, pitch, yaw and roll) represent the face position in 3D space and three orientation angles. The other six variables are the velocities (vx, vy, vz, vpitch, vyaw and vroll). The (x,y) point is used to determine the position of the face in the next frame whereas the z value (depth) is used to determine the estimated face size according to equation 3. Depending on the face position and size in the next frame, a Region Of Interest (ROI) is determined to reduce the search space. The ROI will be searched to detect the face in the next frame rather than the whole frame. This will reduce detection time tremendously. The position of the ROI is centered on the position of the face in the next frame. While the size of the ROI is calculated by multiplying the face size by some *factor* to ensure that the whole face is located within the ROI.

To find a suitable face size factor for the tracking algorithm, empirical studies have been conducted. Figure 3 shows the relation between the number of frames per second and the face size factor for each depth. Figure 4 shows the optimal face size factor for each depth. The linear equation we get from the experiments is:

$$factor = 0.00053 \times depth + 0.8667 \tag{5}$$

Algorithm 1 shows the process of detection, pose estimating and tracking faces.

IV. Experiments

To evaluate the proposed system different experiments have been conducted. These experiments proved reliability within distance range 1-4 meters.

Algorithm 1 Face Detection, Head Pose Estimation and Tracking

```
Initialize Kalman Filter;
Set the Region Of Interest (ROI) as the whole frame ;
for each frame from the sensor (RGB and Depth) do
    Detect Face in the ROI using Cascade Classifier (RGB);
    Get a list of candidate faces positions (x,y) and sizes;
    Verify the detected faces using depth information (z);
    Get a list of the real faces positions and the depth (x,y,z);
    if there is real faces then
        Report the existence of a human face at the position (x,y,z);
        Estimate the head pose;
        Record the head pose information;
        Use the current position (x,y,z) and the current pose (pitch,yaw,roll) as the state variables in Kalman filter;
        Predict the next position of the face (x′,y′,z′) and the next pose (pitch′,yaw′,roll′);
        Get the face size in the next frame from the value of z′ using equation 3;
        Set the position of the ROI position (x′,y′) in the next frame;
        Calculate the face size factor from equation 5;
        Set the size of the ROI as $(facesize * facesizefactor)$;
    else
        Move the robot’s head to search for a face depending on the last position of the face and the motion direction;
        Set the ROI as the whole frame;
    end if
end for
```

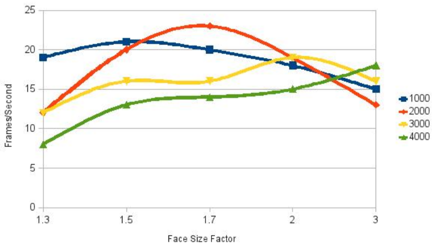


Figure 3: The relation between the frames per second and the face size factor.

Experiment 1

In order to assess the proposed face detector, a real-time stream of data from Kinect sensor with resolution of 640x480 has been examined. The experiment has shown that the false positive rate of the proposed method is lower than using only 2D data as depicted in Figure 5. The cascade classifier scans the entire image searching about faces with many scales as a first stage. The second stage does not need to rescan the entire image. It only examines the detected faces from the previous stage. Consequently, the time needed to examine faces in the second stage is not noticeable. The detected faces were in a rate of 5 frames per second. The experiment has also shown that the Haar cascade classifier influences the perception system especially in the bad lighting conditions.

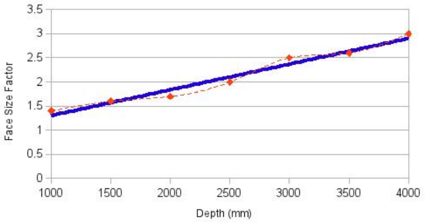
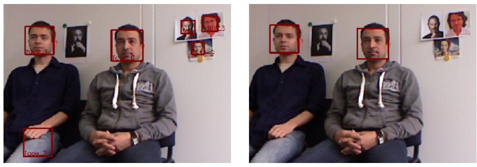


Figure 4: The optimal face size factor for each depth.



(a) candidate faces = 6 (b) real faces = 2

Figure 5: 2D vs. 3D face detection

Experiment 2

This experiment is to compare the proposed head pose estimation with other work. ETH Face Pose Range Image Data Set [8] has been used for comparison. It contains over 10K range images of 20 peoples in many different poses. The head pose ranges cover about ±90° yaw and ±45° pitch rotations. The parameters of the three SVMs are different. Some of them are set by conducting some experiments. The experiments have shown precise and fast responses. Table I compares the results of the proposed head pose estimator with two other methods. It shows the mean and standard deviation of the pose estimation errors in addition to the percentage of correctly classified poses based on 10° threshold.

Table I: Comparison of the proposed head pose estimation method with [8] and [9]

	Yaw error mean/stdev	Pitch error mean/stdev	Pose Estimation Accuracy
The proposed method	3.7/3.4°	2.6/3.0°	94.9%
Fanelli et al.	5.7/15.2°	5.1/4.9°	90.4%
Breitenstein et al.	6.1/10.3°	4.2/3.9°	80.8%

Experiment 3

This experiment demonstrate the use of Kalman filter in human tracking. It reduces the processing time tremendously. Figure 6 shows the processing time with and without tracking. It has shown a processing speed average of 17 frames per second compared to 5 frames per second without tracking. Using the depth in the state variables of Kalman filter overcomes the problem of tracking two persons when they cross each other. Also tracking the head pose increases the estimation confidence.

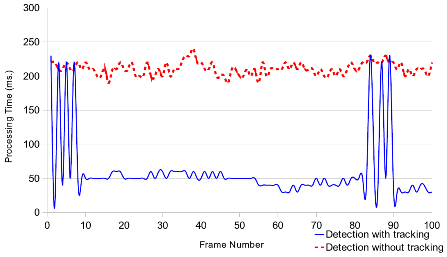


Figure 6: Processing time

V. Conclusion and Future Work

Social interactive robot require stable method for detecting interaction partners. For natural interactive behavior it is important to minimize the false positive rate of these algorithms. The paper presented an approach using a combination of RGB images and depth information to improve the human detection. Besides the pure perception the presented approach also provides a possibility to determine the pose of the head using the 3D information. For tracking a person's face over time a Kalman filter has been added, which also reduces the processing time of the proposed algorithm. This also enables to detect nonverbal cues like shaking the head or nodding. The developed approach has been evaluated in an interactive game scenario.

The next steps in this context will be to include further experiments to evaluate the quality of the developed system and its application for human-robot interaction. Furthermore, the system should be extended in a way that facial expressions as well as body postures of the interaction partners can be evaluated to gather information on the interaction partners emotional state. Other nonverbal expressions like conciously performed gestures should be also recognized in order to improve the robot's interactive capabilities.

References

[1] J. Hirth, "Towards socially interactive robots – designing an emotion-based control architecture," Ph.D. dissertation, Department of Computer Science, University of Kaiserslautern, April 16 2012.

[2] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, Jan 2002, pp. 34–58.

[3] R. Niese, A. Al-Hamadi, and B. Michaelis, "A novel method for 3d face detection and normalization," *Journal of Multimedia*, vol. 2, no. 5, 2007.

[4] W. Burgin, C. Pantofaruy, and W. Smart, "Using depth information to improve face detection," in *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, march 2011, pp. 119 –120.

[5] L. Xia, C.-C. Chen, and J. Aggarwal, "Human detection using depth information by kinect," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, june 2011, pp. 15 –22.

[6] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 607 –626, April 2009.

[7] E. Seemann, K. Nickel, and R. Stiefelhagen, "Head pose estimation using stereo vision for human-robot interaction," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, may 2004, pp. 626 – 631.

[8] M. Breitenstein, D. Kuettel, T. Weise, L. van Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008, pp. 1 –8.

[9] G. Fanelli, J. Gall, and L. V. Gool, "Real time head pose estimation with random regression forests," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, june 2011, pp. 617 –624.

[10] E. Hjelmas and B. K. Low, "Face detection: A survey," in *Computer Vision and Image Understanding*, vol. 83, no. 3, Sept. 2001, pp. 236–274.

[11] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[12] G. Welch and G. Bishop, *An Introduction to the Kalman Filter*, University of North Carolina Chapel Hill, Chapel Hill (NC), USA, July 2006.

[13] N. Schmitz and K. Berns, "Perception systems for naturally interacting humanoid robots," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Atlanta, GA, USA, July 31 - August 3 2011, pp. 425–432.

[14] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, aug 2004.

ISBN: 978-1-61208-250-9