AUDIO-VISUAL PERCEPTION OF HUMANS FOR A HUMANOID ROBOT

Kai Nickel, Hazim K. Ekenel, Michael Voit, Rainer Stiefelhagen

Interactive Systems Labs
Universität Karlsruhe (TH)

ABSTRACT

We present a combined system that enables a humanoid robot to sense humans using different modalities. The user is localized by means of a joint audio-visual person tracker. His face is identified using a novel local-appearance based approach which is robust against partial occlusion. In order to determine the user's focus and object of interest respectively, we estimate headpose using a neural-network classifier and recognize pointing gestures based on hand motion.

1. INTRODUCTION

In order to build human-friendly, human-centered robots, the perception of the user(s), their locations, identities, activities, and communicative cues are an essential necessity. Such perception is necessary to understand what users do, what they want from the robot, and to generate an appropriate response or proactive behavior of a robot. The location of the user and his body parts is of course also important for efficient and safe interaction between the robot and a user.

In this paper we present our recent work on building a audio-visual perceptual system that allows a humanoid robot to perceive humans, their locations, identity and visual communication modalities, such as pointing gestures and head orientation, in the surrounding of the robot. This work complements our lab's work on speech recognition and dialogue systems for humanoid robots [1], and provides among other things the users' visual communication cues for multimodal human-computer dialogue [2].

The body of related work covering the individual methods used in this integrated system exceeds the scope of this overview paper. There is, however, work on integrated systems aiming at similar targets: For example Lang et al. [3] have demonstrated a perception system for a mobile robot that localizes the user by means of frontal face detection, acoustic source localization and leg detection with a laser range finder. Furthermore, the user's face is identified using the Eigenface [4] method. Okuno et al. [5] describe an audio-visual tracking system for an upper-torso humanoid equipped with a stereo camera and 2 pairs of microphones. Based on face detection and acoustic source localization, the user is tracked and motor commands for following the user with the robot head are generated.

The remainder of this paper is organized as follows: Section 2 presents an overview of the proposed perceptual system and presents the individual perceptual components – audio-visual person tracking, pointing gesture recognition, head pose estimation, face recognition – in some detail. In Section 3, we then present experimental results of the individual components. We conclude the paper in Section 4.

2. PROPOSED METHODS

Fig. 1 shows an overview of the components employed for human sensing. The available sensors on the robot head are: a fixed baseline stereo camera and 6 omni-directional microphones. The camera and microphone input streams are jointly processed by an audio-visual person tracker that fuses both modalities in a probabilistic framework. The output of the person tracker, a hypothesis about the user's head position, is then passed on to the following components: a face identification module based on local DCT features, a head-pose estimator using ANNs, and a 3d-blob tracker for hand tracking. The hand positions are in turn processed by an HMM-based pointing gesture recognizer. The pieces of information gathered by this perception chain are: the location of the user, his identity and head orientation, and the pointing direction in case of a detected gesture. The following sections describe each of the individual components in more detail.

2.1. Audio-visual person tracking

Localizing its user is a basic capability for a robot interacting with humans. Succeeding perception modules like person identification, gesture recognition and head-pose estimation require the location of the user in world and image coordinates to perform their calculations.

The audio-visual person tracking module employed it this system is a further development of the tracker presented in [6]. This section summarizes the algorithm and its extensions. For a more detailed description we refer to the above mentioned publication.

In our scenario, the task of person tracking is to localize the user in front of the robot. The features used for tracking are foreground segmentation, face and body detection as well as color models on the video side. On the audio side, we

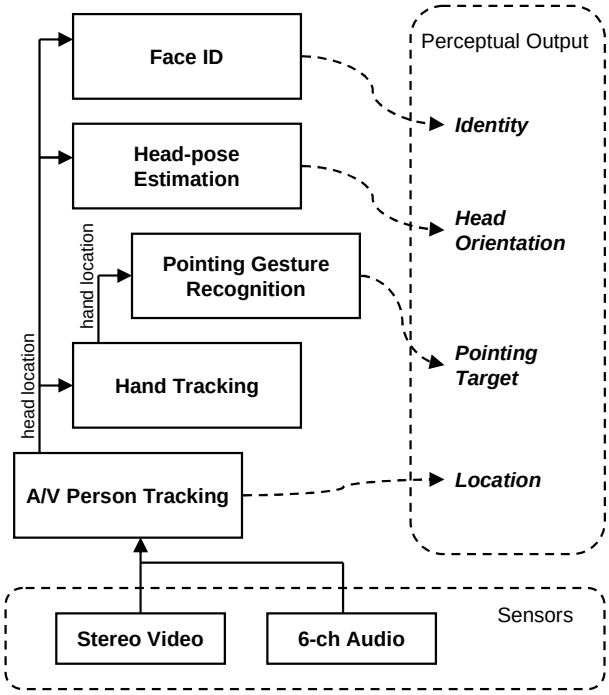

Fig. 1. System overview.

localize the user by evaluating the time-delay-of-arrival of the user's speech with respect to the multiple microphones on the robot's head.

We fuse both modalities in a joint particle filter framework. Particle filters [7] represent a generally unknown probability density function by a set of m random samples $s_{1..m}$. Each of these particles is a vector in state space and is associated with an individual weight π_i. The evolution of the particle set is a two-stage process which is guided by an observation model and a motion model. The state space is given by $s_i = (x, y, z)$, i.e. each particle is a hypothesis about the 3d-position of the user's head centroid.

Using the features described in following subsections, we can calculate a weight for each particle by combining the normalized probabilities of the visual observation V_t and the acoustical observation A_t using an adaptive weighting factor α:

$$\pi_i = \alpha \cdot p(A_t|s_i) + (1 - \alpha) \cdot p(V_t|s_i) \quad (1)$$

The weighting factor α is adjusted dynamically according to an acoustic confidence measure (e.g. signal energy). For the motion model we chose simple Gaussian diffusion, i.e. a 0th order motion model. The weighted mean of the particle set is the final hypothesis about the user's location.

2.1.1. Video features

For each particle, the visual observation model calculates three scores that are made up of the following visual cues. These scores are combined to generate a pseudo-probability $p(V_t|s_i)$ as required by equation 1. Fig. 2 shows an overview of the feature maps.

The first cue is based on foreground segmentation. Each pixel's mean intensity is stored in the background model and continuously updated with a learning factor λ. The foreground map is made up of the absolute differences between the input image and the learned background model. By choosing λ close to 1 we achieve quick model adaptation that allows for occasional camera movements. To evaluate a particle with respect to the foreground map, a person model at the particle's position is projected to the image. The foreground score is then calculated by accumulating the foreground pixels covered by this model. The model consists of three cuboids for head, torso and legs. The projection is approximated by three orthogonal rectangles in the image plane, such that the accumulation of the pixel values can be computed efficiently by means of the integral image [7].

The face detection algorithm proposed by [8] and extended by [9] is known to be both robust and fast: it uses Haar-like features that can be efficiently computed by means of the integral image, thus being invariant to scale variations. Typically, a variable-size search window is repeatedly shifted over the image, and overlapping detections are combined to a single detection. Exhaustively searching a $W \times W$ image region for a $F \times F$ sized face while incrementing the face size n times by the scale factor s requires $\sum_{i=0}^{n-1} \left(W - F \cdot s^i\right)^2$ cascade runs. This is an issue for real-time processing.

In the proposed particle filter framework however, it is not necessary to scan the image exhaustively: the places to search are directly given by the particle set. For each particle, a head-sized cuboid is projected to the image plane, and the bounding box of the projection defines the search window that is to be classified. Thus, the evaluation of a particle takes only a single run of the cascade. Particles producing detector hits are given high scores, particles without detector hits are scored low. In addition to face detection we use a detector trained on upper bodies. The detector hits are incorporated in the particles' scores using the same method as for face detection.

As the third visual cue we use individual color models for the three body parts of our model: head, torso and legs. The color models are implemented as histograms in the rgb color space. For each body part, a support map is generated by histogram back-projection. Just like for the foreground segmentation cue, the particles are scored by accumulating the support map pixels under the projected model. Again, the integral image can be used to reduce the computational complexity. The color models are updated after each frame using the final tracking hypothesis.

2.1.2. Audio features

Consider a pair of microphones, and let m_1 and m_2 respectively be the microphones' positions. Let x denote the po-

Fig. 2. Video feature maps with the 3-box body model superimposed. From left to right: a) Camera image, b) Foreground segmentation, c) Detector response, d) Color segmentation (head-torso-leg color model projected to the red-green-blue channel respectively).

sition of the speaker in a three dimensional space. Then the *time delay of arrival* (TDOA) between the two microphones can be expressed as

$$T(x) = T(m_1, m_2, x) = \frac{\|x - m_1\| - \|x - m_2\|}{c} \quad (2)$$

where c is the speed of sound.

To estimate the TDOA, a variety of well-known techniques [10, 11] exist. Perhaps the most popular method is the *phase transform* (PHAT), which can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})|} e^{j\omega\tau} \, d\omega \quad (3)$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the signals of the microphone pair.

Normally one would search for the highest peak in the resulting cross correlation to estimate the position. In our particle filter framework, however, we interpret the PHAT as probability density function. As the values returned by the PHAT can be negative, but probability density functions must be strictly nonnegative, we found that setting the negative values of the PHAT to zero yielded acceptable results. The acoustic particle scores are then given by the PHAT values at the hypothesized time-delay position $T(x = s_i)$.

We integrate the scores from all those microphone pairs that are exposed to direct sound given the microphone position with respect to the hypothesized sound source. Thus, we have a multi-target acoustic tracker with an implicit best-microphone-selection strategy.

2.2. Pointing gesture recognition

The modules for pointing gesture recognition and hand tracking used in this system have been described in detail in [12]. The gesture recognition is based on hand motion. Dedicated HMMs are used to model the begin-, peak- and retract-phases of typical pointing gestures. They were trained on 3d hand trajectories of hundreds of sample pointing gestures. Whenever the models fire successively, a gesture is being detected and the peak phase is being identified. Within the peak phase, the line between the head and the pointing hand is used to estimate the pointing direction. If the positions of potential pointing targets are known, the most likely target can be selected by measuring its deviation from the pointing direction.

The underlying hand tracking algorithm is based on a combination of adaptive skin-color classification and dense disparity maps. The skin-colored pixels are associated with their corresponding disparity values and are then spatially clustered. Those clusters are evaluated in order to find an optimal assignment of head and hands to skin-color clusters in each frame. The optimal assignments maximizes three aspects: a) the match between hypothesis and observation in terms of skin-color pixels, b) the naturalness of the hypothesized posture, and c) the smoothness of transition from the preceding to the current frame. After temporal smoothing, the hand trajectories are passed to the gesture detection module described above.

2.3. Head-pose estimation

A person's head orientation provides good indication of a person's focus of attention, i.e. about the objects, area or people in which someone is interested or with which he or she interacts [13, 14]. It is in particular a quite reliable cue to determine whether a person is addressing a robot or someone else [15], which is important to build robots that only respond to users when it is appropriate.

As for head pose estimation, we integrated the basic system that has been presented and described during CLEAR Evaluation Workshop in 2006 [16]. We train one neural network classifier for estimation the camera-relative head orientation of a person standing in front of the robot. Retrieving a region of interest as a coarse hypothesis, in which the tracked head has to reside in, we first align a tight-fitting bounding box around our head candidate with the help of a skin-color segmentation in HSV color space. An attached connected component search, that extracts regions with a minimally, heuristically size of 30×40 pixels, yields the desired head candidate. We crop the head region and rescale it to a normalized size of 64×64 pixels. A grayscale image is then being used to both compute the Sobel magnitude in order to

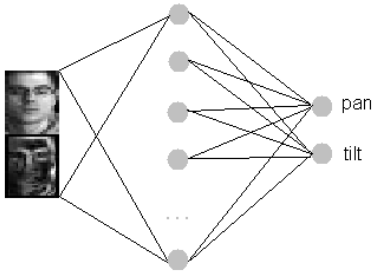

Fig. 3. Neural Network for head pose estimation: The output neurons both state a continuous hypothesis within the range of $[-90°, +90°]$. As input features, an intensity image of the extracted head region and its Sobel magnitude are used.

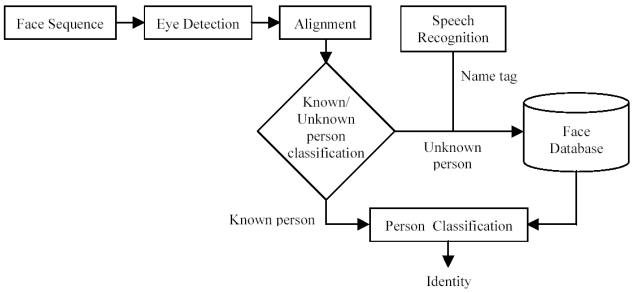

Fig. 4. Overview of the face recognition subsystem.

retrieve an edge map in both horizontal and vertical direction and to be merged into a final feature vector by concatenating the pixel values row by row. We trained a neural network to output a continuous estimate of the depicted head orientation in the range of $[-90°, +90°]$ in both pan and tilt direction. We use one output neuron to state a real value in between the given range of possible angles. The network contains 80 hidden units and is being trained with Standard Error Backpropagation. We used 100 training cycles. A cross evaluation set was used to save the network from the very best performing training iteration. Figure 3 depicts the networks topology and input features.

2.4. Face identification

Face recognition is one of the most important building blocks of a natural human-robot interaction system. A humanoid robot should recognize the people that he has met before and store associated information related to these people to conduct a natural conversation. Moreover, the robot should also detect the people that he has not met and in the case of interaction with these people, it should ask the name of them.

The intense research efforts on face recognition, especially since the beginning of 1990s, has provided significant improvements in face recognition performance on the standard databases that have been collected under controlled laboratory conditions. However, face recognition in uncontrolled environments is still a very difficult problem [17]. To provide robust face recognition and to overcome uncontrolled environmental conditions, we are utilizing multiple samples of the same face that are obtained from the video sequence. Our face recognition system takes a face sequence, which is provided by the tracking module described in section 2.1, as input. It analyzes the input face images to locate the eyes and then registers the face images according to the eye center coordinates.

Local appearance-based face recognition approach is used to extract feature vectors from each face image [18]. In this feature extraction approach, the input face image is divided into 8x8 pixel blocks, and on each block, discrete cosine transform (DCT) is performed. The most relevant DCT features are extracted using the zig-zag scan and the obtained features are fused either at the feature level or at the decision level for face recognition [18].

After extracting the feature vectors from each face image in the sequence, they are compared with the ones in the database using a nearest neighborhood classifier. Each frame's distance scores are normalized with Min-Max normalization method [19], and then these scores are fused over the sequence using the sum rule [20]. The obtained highest match score is compared with a threshold value to determine whether the person is known or unknown. If the similarity score is above the threshold, the identity of the person is assigned with that of the closest match. If the similarity match score is below the threshold, then the person is classified as unknown. A simple diagram of the face recognition system is illustrated in Figure 4.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The individual components of the proposed system have been evaluated independently. The following sections present results for gesture recognition, head-pose estimation and face identification. The audio-visual person tracker has not yet been evaluated in the integrated system. In [6], however, tracking experiments and results are presented for a different kind of sensor setup.

3.1. Pointing Gesture Recognition

In order to evaluate the performance of the gesture recognition module, we prepared an indoor test scenario with 8 different pointing targets. Ten test persons were to move around within the camera's field of view, every now and then showing the robot one of the marked objects by pointing on it. In total, we captured 206 pointing gestures within a period of 24 min. A detailed description of the results can be found in [12]. The gestures as well as the hand positions were manually labeled and then compared against the gesture detection and hand tracking hypotheses. Table 1 summarizes the most important results.

Detection rate (recall)	78%
Precision	83%
Avg. error angle	37°
Targets identified	65%

Table 1. Evaluation of pointing gesture recognition.

Pan Avg. Error	12.3°
Tilt Avg. Error	12.8°
Correct Pan Class	41.8%
Correct Tilt Class	52.1%

Table 2. Experimental results of our head pose estimator on the Face Pointing04 Database.

3.2. Head-pose Estimation

We evaluated our core head pose estimating system during CLEAR Workshop 2006 on the Face Pointing 04 Database [21]. The database consists of 15 sets of images. Each set contains 2 series of 93 images of the same person at different poses. The first set is to be used for training, the second set for evaluation. In total, there are captures of 15 different people with varying skin color and clothing. Some persons also wear glasses. Figure 5 depicts some samples from the dataset. Our system was evaluated on both horizontal and vertical head orientations. As it can be seen in Table 2, horizontal estimation performed with 12.3° mean error. Estimating the head's vertical rotation performed slightly worse with 12.77° mean error.

3.3. Face Identification

The approach is extensively tested on the publicly available face databases and compared with the other well known face recognition approaches. The experimental results showed that the proposed local appearance based approach performs significantly better than the traditional holistic face recognition approaches. Moreover, this approach is tested on face recognition grand challenge (FRGC) version 1 data set for face verification [22], and a recent version of it is tested on FRGC version 2 data set for face recognition [23], and it provided better and more stable results than the baseline face recognition system. For example, in the conducted experiments on the FRGC version 2 data set, 96.8% correct recognition rate is obtained under controlled conditions and 80.5% correct recognition rate is obtained under uncontrolled conditions. In these experiments, there are 120 individuals in the database and each individual has ten training and testing images. There is a time gap of approximately six months between the capturing time of the training and the test images. The approach is also tested under video-based face recognition evaluations and again provided better results [24, 25].

Fig. 5. Sample images from the Face Pointing 04 Database.

4. CONCLUSION AND FUTURE WORK

In this paper, we have presented an audio-visual perception system that enables a robot to gather the following information about its user: location, identity, head-pose and pointing gestures. The user is being localized by processing stereo-video and multi-channel audio in a joint particle filter framework. His focus and object-of-attention respectively are determined by head-pose estimation and HMM-based gesture recognition. Face identification is performed using a novel local-DCT-based method that has shown to be superior to conventional holistic face recognition approaches.

In future work, we plan a more tightly coupled fusion of the components, especially for the tracking and head-pose estimation modules. By mutually sharing information between those modules, we hope to both reduce processing time and to increase precision. Furthermore, an attentional system will be developed that generates robot head movements to actively explore the scene in order to search for a potential communication partner.

Acknowledgments

This work has been funded by the German Research Foundation (DFG) as part of the Sonderforschungsbereich 588 "Humanoide Roboter".

5. REFERENCES

[1] C. Fuegen, P. Gieselmann, H. Holzapfel, and F. Kraft, "Natural human-robot communication," in *To appear in: 2nd Intl. Workshop on Human-Centered Robotic Systems*, Munich, Germany, 2006.

[2] R. Stiefelhagen, C. Fuegen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, gaze and gestures," in *IEEE/RSJ Intl. Conference on Intelligent Robots and Systems*, Sendai, Japan, 2004.

[3] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for

human-robot-interaction: A multi-modal attention system for a mobile robot," in *Proc. Int. Conf. on Multimodal Interfaces*, Vancouver, Canada, 2003, pp. 28–35.

[4] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuro Science*, vol. 3, no. 1, pp. 71–86, 1991.

[5] Hiroshi G. Okuno, Kazuhiro Nakadai, and Hiroaki Kitano, "Social interaction of humanoid robot based on audio-visual tracking," in *15th Intl. Conference on Industrial and Engineering, Applications of Artificial Intelligence and Expert Systems, IEA/AIE*, Cairns, Australia, June 2002, vol. 2358, pp. 725–735, Springer.

[6] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proceedings of the 7th International Conference on Multimodal Interfaces*, Trento, Italy, October 4-6 2005, pp. 61–68.

[7] M. Isard and A. Blake, "Condensation–conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[8] P. Viola and M. Jones, "Robust real-time object detection," in *ICCV Workshop on Statistical and Computation Theories of Vision*, July 2001.

[9] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *ICIP*, September 2002, vol. 1, pp. 900–903.

[10] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. ICASSP '94*, Adelaide, Austrailia, April 1994, pp. II–273–II–276.

[11] J. Chen, J. Benesty, and Y.A. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," in *IEEE Trans. Speech Audio Proc.*, November 2003, pp. 11(6):549–57.

[12] K. Nickel and R. Stiefelhagen, "Pointing gesture recognition based on 3d-tracking of face, hands and head orientation," in *Proceedings of the 5th International Conference on Multimodal Interfaces*, Vancouver, Canada, November, 5-7 2003.

[13] Paul P. Maglio, Teenie Matlock, Christopher S. Campbell, Shumin Zhai, and Barton A. Smith, "Gaze and speech in attentive user interfaces," in *Proceedings of the International Conference on Multimodal Interfaces*. 2000, vol. 1948 of *LNCS*, Springer.

[14] B. Brumitt, J. Krumm, B. Meyers, and S. Shafer, "Let there be light: Comparing interfaces for homes of the future," *IEEE Personal Communications*, August 2000.

[15] M. Katzenmaier, R. Stiefelhagen, T. Schultz, I. Rogina, and A. Waibel, "Identifying the addressee in human-human-robot interactions based on head pose and speech," in *Intl. Conference on Multimodal Interfaces*, State College, PA, USA, October 2004.

[16] M. Voit, K. Nickel, and R. Stiefelhagen, "Neural network-based head pose estimation and multi-view fusion," in *CLEAR Evaluation Workshop*, Southampton, UK, 2006.

[17] Zhao, Chellappa, Phillips, and Rosenfeld, "Face recognition: A literature survey," *CSURV: Computing Surveys*, vol. 35, 2003.

[18] H. K. Ekenel and R. Stiefelhagen, "Local appearance-based face recognition using discrete cosine transform," in *13th European Signal Processing Conference, Antalya, Turkey*, 2005.

[19] Robert Snelick, Umut Uludag, Alan Mink, Mike Indovina, and Anil K. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 27, no. 3, pp. 450–455, 2005.

[20] J. V. Kittler, "Combining classifiers: A theoretical framework," *Pattern Analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.

[21] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures, Cambridge, UK*, 2004.

[22] H. K. Ekenel and R. Stiefelhagen, "A generic face representation approach for local appearance based face verification," in *Workshop on Face Recognition Grand Challenge Experiments*, 2005, pp. III: 155–155.

[23] H. K. Ekenel and R. Stiefelhagen, "Analysis of local appearance-based face recognition on frgc 2.0 database," in *Face Recognition Grand Challenge Workshop (FRGC)*, Arlington, VA, USA, March 2006.

[24] H. K. Ekenel and A. Pnevmatikakis, "Video-based face recognition evaluation in the CHIL project - run 1," in *International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 85–90.

[25] H. K. Ekenel and Q. Jin, "Isl person identification system in the clear evaluations," in *CLEAR Evaluation Workshop*, Southampton, UK, 2006.