Noname manuscript No.
(will be inserted by the editor)

A Time-Delay Control Approach for a Stereo Vision Based Human-Machine Interaction System

Gigel Macesanu · Vasile Comnac · Florin Moldoveanu · Sorin M. Grigorescu

Received: date / Accepted: date

Abstract In this paper, an approach to control a 6-DoF stereo camera for the purpose of actively tracking the face of a human observer in the context of *Human-Robot Interaction* (HRI) is proposed. The main objective in the presented work is to cope with the critical *time-delay* introduced by the computer vision algorithms used to acquire the feedback variable within the control system. In the studied HRI architecture, the feedback variable is represented by the 3D position of a human subject. We proposed a predictive control method which is able to handle the high time-delay inserted by the vision elements into the control system of the stereo camera. Also, along with the predictive control approach, a novel 3D nose detection algorithm is suggested for the computation of the feedback variable. The performance of the implemented platform is given through experimental results.

Keywords Human robot interaction · Time-delay systems · Active vision · Facial features detection and tracking

1 Introduction

In recent years, the number of service robotics application scenarios centered in human environments has drastically increased [26]. Such applications span from common all-day-living assistance platforms [3] to care-giving robots deployed in hospitals and homes [19]. Although the navigation and mobile manipulation capabilities of such robots increased exponentially in the last decade, there is still a lack of proper *Human-Robot Interaction* (HRI) methods. The

G. Macesanu · V. Comnac · F. Moldoveanu · S.M. Grigorescu
Department of Automation, *Transilvania* University of Brasov, Mihai Viteazu 5, 500174, Brasov, Romania
Tel: +40-268-418-836
E-mail: {gigel.macesanu,comnac,moldof,s.grigorescu}@unitbv.ro

HRI term denotes the process through which a human person enters in contact with a robot, usually performed with the purpose of sending certain commands. With the advent of new imaging technologies, the HRI paradigm has been approached from the perspective of recognizing different human features from which robotic commands can be inferred [30]. Such features, which can be used to determine the head pose of the human with respect to the robot, are the eyes, nose and mouth. The work presented in this paper aims at controlling the orientation and zooming capabilities of the robotic vision hardware, with the goal of expanding the HRI interaction area. This challenge has been tackled through the active tracking of the human nose in a stream of stereo images acquired from an active 6-DoF stereo camera. Within the camera control structure, the nose is considered to be one of the best features to track since it can be visible from different imaging perspectives of the human head. As opposed to the nose, the center of the detected face may not always be optimally bounded by the face detection algorithms.

Humans use a variety of mechanisms through which they can send information to other persons, information regarding their state of mind or related to other human interactions. All these mechanisms used by a human to send or receive information fall under the name of *non-verbal communication* or *body language* [6]. Also, through the interaction, humans are turning their attention towards characteristics such as the human voice, the features, or the movements of a person [7]. This typical human behavior starts to develop from childhood, when only simple human features are tracked, until adulthood, when complex structures for human features understanding are gained [31].

Starting from the above described human behavior and using the advantages of current computation power, different research groups are aiming at developing robotic platforms with the capacity to mimic human behavior. Robotic systems that realize human behavior imitation use information with respect to the human observer in order to understand the geometrical relation between the observer and the robot. Such a system is the *ROVIS Human Interaction* platform presented in Fig. 1.

The objective of the proposed HRI architecture is to maximize the interaction area by maintaining the features of a person within the center of the acquired image through the control of two *Pan-Tilt-Zoom* (PTZ) cameras. The usage of two cameras is motivated by the fact that, along with the 2D location of a person in the image plane, we also compute the real distance from the camera to the person, thus allowing us to control the zoom components of the cameras. By controlling the zoom, the quality of the acquired image is improved, thus providing a proper input to the scene understanding algorithms.

In comparison to structured light sensors, such as the MS Kinect®, a stereo camera has the advantage that it can be used in outdoor environments, where, in the case of active sensors, the infrared pattern projected for the obtaining depth estimation, interferes with the natural light coming from the sun. Furthermore, a stereo camera with a wide enough baseline between the sensors (as the cameras usually mounted on robotic platforms) can deliver more precise

Fig. 1 Human-Robot Interaction example within the ROVIS system (face detection and 3D mapping).

depth data for objects further away from the camera than a structured-light device, which is dependent on the projected infrared grid. Also, laser range sensing technology, although precise in estimating distances, requires additional imaging sensors for extracting the features to track. Nevertheless, the time-delay approach presented in this paper can be directly applied for controlling the pose of other imaging systems.

A core concept in HRI is the recognition of human gestures, such as the pointing of a direction by a person, studied by Nickel in [27]. The approach proposed in [27] uses a *Hidden Markov Model* (HMM) and a stereo camera hardware setup to track the human head's pose and the arms. Also, Park proposed in [29] a *Particle Filter* (PF) based gesture recognition system for the purpose of mobile robot navigation. The *Engagement Concept*, referring to the way a person starts an interaction, maintains it and finally finishes it, has been integrated into a robotic structure that can participate in interactions with humans at the level of conversations and collaborations which involves gestures [32].

The HRI structure described in this paper falls in the area of active vision systems, set forth in the seminal work of Aloimonos [1]. In such a HRI system, one of the most crucial elements is the real-time capability of the architecture to control in real-time the 6 *Degrees-of-Freedom* (DoF) stereo camera. The active control of robot vision imaging technologies has been tackled in a number of research papers. Just to mention a few, in [36], a probabilistic framework for adapting the gaze of a single camera for human face acquisition is proposed. The control of two PTZ cameras is treated in [21] from the perspective of 3D depth computation and the calculation of the homographic transformations between the sensors. An algorithm for rectifying stereo images acquired by two PTZ sensors is presented in [38].

The robustness and stability of an active vision system is strongly dependent on the time-delay introduced by the image processing system into the control scheme [16]. Although there is a large number of stereo PTZ systems, such as the ones mentioned above, for which powerful computer vision algorithms have been developed the impact of the time-delay introduced by the image processing component into the overall control structure has been scarcely studied. There is also important to note, that, in our work, we do not try to contribute with a computation time enhancement of the image process-

ing methods, but to cope with the time-delay introduced by them. Nevertheless, the overall system can only benefit by the speed boosting of the vision algorithms through parallel computation devices (eg. GPUs, or FPGAs).

The main goal of the work presented in this paper is to create an apparatus which is able to follow a human observer's facial features (eg. nose tip) when he/she is moving on an unspecified path. In the application, the time delay corresponds to the process of face detection, image segmentation and 3D nose tip position estimation. In the same time, the delay produced by the image processing components is *stochastically variable* and depends on the effort the vision algorithms need for accomplishing their tasks. For example, if the tracked facial features are present in a large image area, the computational effort for detecting them will be low, while for small areas, that is, when the human subject is further away from the camera, the time needed for detection will be higher. A discussion and experimental results with respect to this variation will be given in the performance evaluation section. As calculated in [23], the maximum processing delay which can be introduced in the system, without destabilizing it, is approx. $0.48s$.

In time-delayed, or dead-time systems, the controller's choice and its tuning involve the usage of specific methods, such as prediction control [9], the classical Ziegler-Nichols approach [11], or the generalized form of the *Hermite-Biehler* theorem [33].

The problems of dead-time systems have been addressed in [2]. In order to fulfill the stability requirements it is needed to determine the maximum delay, also known as *delay margin*, that can be introduced in the system without affecting its stability. The compensator can thus be designed using this delay margin. Corke presented in [9] several algorithms, based on PID or Smith predictor controllers, in which an object of interest can be tracked using an active camera. Based on the classical Smith predictor, a neural structure for the control of a telerobotic system with time-delays caused by communication channels has been proposed in [20]. PID regulators for controlling dead-time systems have also been proposed in [39] and [25]. A stability interval for a P regulator used to control a dead-time system was obtained by Silva [34] using an analytical approach.

In our previous work, we have used the generalized *Hermite-Biehler* theorem to develop a *Proportional* (P) controller for compensating the delay present in the visual control system of a 6-DoF active stereo camera [23][24]. In the current work, we try to overcome the limitations of the P controller through the development of a *Proportional-Integrative-Derivative* regulator using prediction control. As in classical predictive control, the main characteristic of the approach is the extraction of the dead-time component outside the feedback loop. As it will be shown in the experimental results section, the control precision, as well as the computation time, have been improved using the presented approach.

In this paper we propose a control approach for a stereo active vision system used in HRI, which inherently incorporates dead-time introduced by the image processing elements. The rest of the paper is organized as follows.

In Section 2, the face detection and nose tip 3D position estimation algorithms are detailed. The descriptions of the mathematical model and of the control approach for the 6-DoF active stereo camera are given in Sections 3 and 4, respectively. Finally, before conclusions and outlook, performance evaluation results are presented in Section 5.

2 Human Head Pose Estimation

The first step in estimating the pose of a head is to detect it at the 2D image level. For this purpose, two *boosting* classifiers trained for recognizing human faces [37] have been used. The 3D orientation of the head is given through the detection of the nose tip, as it will be further explained.

2.1 Face Detection in the 2D Image Domain

The boosting approach is a general framework used to improve the accuracy of a certain machine learning algorithm [15]. This is performed by combining a weighted voting scheme using N hypotheses which have been generated by a repeated training built around different subsets of training data. A boosting classifier is composed of a so-called *weak* and *strong* learner, or *classifier*:

- **weak learner**: has to perform only slightly better than random guessing, that is, its overall classification error has to be smaller than 50%. The hypothesis h_{weak} is obtained from a learning algorithm;
- **strong learner**: from a set of N weak learners, a strong learner, or classifier, is obtained as a linear combination of weak learners.

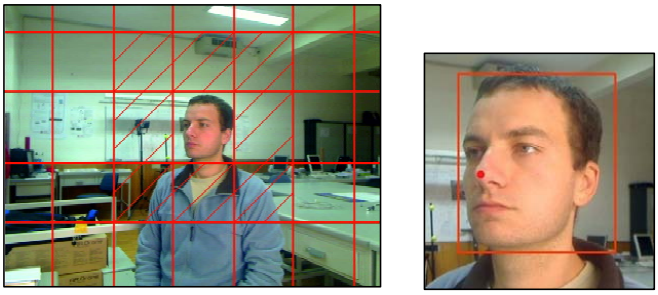
The two classifiers used for detecting human faces have been trained offline with frontal and profile faces, respectively. A number of face samples used for training may be seen in Fig. 2. The training data is composed of 4000 positive and negative manually selected image regions containing human faces. As described in [5], for each region, a set of *Haar-like* features have been calculated [37]. The implied *AdaBoost* technique automatically selects those features that best describe the human faces. In recent years, along the traditional Haar-like ones, new features for object recognition, and in particular face detection, gained popularity. Among them are the *Local Binary Patterns* (LPB) [28] and the *Histogram of Oriented Gradients* (HOG) [10], originally developed for full body human recognition. In our experiments with LBP features, we have noticed a slight improvement in the detection accuracy, as well as a processing time enhancement. Since the main goal of the work presented in this paper is the delay time introduced by the image processing components, we have chosen to stick with the standardized Haar features, leaving a comparison between the several feature extraction methods for future work.

The boosting face detection algorithm is applied on each image delivered by the stereo vision system. In order to achieve a confident head detection, a series

Fig. 2 Sample faces used to train the two AdaBoost classifiers for frontal and profile face detection.

of face recognition parameters have been tuned, such as the *scale factor*, used to determine the face scale difference between each search, the *search area size*, used to bound the minimal head search region, or the *head selection confidence*, used to select the best recognized face from an image where multiple faces are present. These parameters have been obtained heuristically within the context of the HRI scenario, that is, the face of the human subject will probably cover a specific image area, given the interaction area with the robot.

Real-time head detection capabilities have been achieved through a dynamic face search window adjustment approach, as follows. At the *initialization* phase, the search ROI is considered to be the whole input image. During execution, when a face fails to be detected, the previously detected face is used as input for the *search window adjustment* algorithm, which, with every new frame, increases the search area based on the location of the last detected face. In Fig. 3(a), the construction of the new search window, centered on previous head's ROI, is presented. The obtained 2D head region of interest is used as input for the following nose detection stage.



(a) (b)

Fig. 3 Human features detection. (a) Search window computation for face detection. (b) Human nose detection.

2.2 Nose Detection

The main advantage in detecting the tip of the nose is that it can deliver a more confident pose of the head in frontal, as well as in profile images, since, for the

case of profile images, it is more visible in comparison to other features, such as the lips. As with the nose, many of these characteristics are related to their poses, color or prominence. In this paper, the fact that the nose has a specific color is used for its detection. Therefore, a color based segmentation approach, followed by a nose contour identification process has been constructed.

The nose detection method has its roots in the original lips detection algorithm proposed in [35]. The face ROI, determined previously through the face detection method, is used for constructing a search region for nose tip recognition. This new region is determined according to the head's width and height in the 2D image domain. More accurately, it represents approx. 33% of the central head region. The 33% value has been empirically chosen, taking into account the a-priori knowledge that the nose is situated in the central head region. In is important to note here that the nose segmentation approach presented in this section is strictly dependent on the face recognition method for computing the nose search region. The nose region extraction can be further improved through the calculation of additional facial features, using algorithms such as boosting classifiers trained for eyes detection, or the correlation filters proposed in the work of Bolme [4]. However, since the recognition of such extra features is also based on the existing face detection algorithm, the additional feature extraction methods would increase the processing time, adding little improvement to the nose segmentation technique.

In order to cope with variable illumination conditions, the nose segmentation has been applied on images represented in the $L * a * b$ color space, obtained by converting to Lab the acquired RGB images. The Lab color space format is composed of a *Luminance* image channel L and two channels encoding the color, a and b. The block diagram of the nose segmentation algorithm can be seen in Fig. 4. A morphological gradient filter, applied on the b channel image, is used to isolate perimeters of existing binary blobs. The new perimeters are treated as nose hypotheses [5]. For segmenting the nose, a logical **AND** operation between the a color channel and the image resulted from the morphological filter is applied.

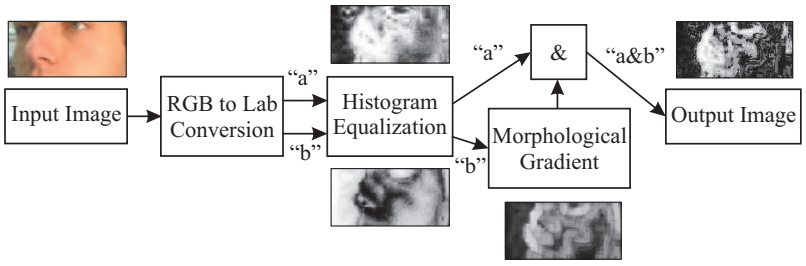
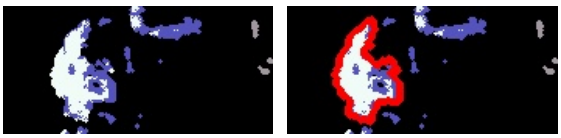


Fig. 4 Block diagram of the proposed human nose segmentation algorithm.

On the nose segmented image, all the segmented pixels are grouped into clusters based on their connectivity using a *K-means clustering* approach. Such

a clustering example can be seen in Fig. 5(a). The hypotheses clusters are further classified into the object of interest, that is the nose, and background based on their central and invariant moments. The final 2D nose tip coordinate can be seen in Fig. 3(b).



(a) (b)

Fig. 5 Nose hypotheses segmentation. (a) Pixels classification using K-means clustering. (b) Nose contour detection and recognition.

2.3 3D Pose Estimation

The tip of the nose, calculated in the 3D real world space, is considered to be the feedback variable of the active vision control system described in this paper. The 3D pose of the nose is obtained from its recognition in the input stereo image, that is, the left and right image pair. The 3D reconstruction procedure pair takes as input the stereo rectified images, the 2D image nose coordinates and the internal parameters of the stereo camera (e.g. focal length and optical center). The computation of the 3D nose pose is given by the following relations [18]:

$$X = x_l \frac{b}{d}, \tag{1}$$

$$Y = y_l \frac{b}{d}, \tag{2}$$

$$Z = f \frac{b}{d}. \tag{3}$$

where, $P = (X, Y, Z, 1)$ are the 3D homogeneous coordinates of the nose, projected on left and right image planes as p_l and p_r:

$$\begin{cases} p_l = (x_l, y_l, 1), \\ p_r = (x_r, y_r, 1), \end{cases} \tag{4}$$

where b represents the distances between the two sensors of the stereo cameras, f is the focal length and $d = x_l - x_r$ is the disparity between the two projections of the nose tip in the image planes. The computed 3D pose of the nose is used as feedback variable for the camera control system.

3 Modeling of a 6-DoF Active Vision System

The main goal of the proposed active vision platform is to expand the human-robot interaction area by controlling the orientation of the camera system. The basic block diagram of the proposed architecture can be seen in Fig. 6(a). The feedback variable is represented by the head detection system described in the previous section. The position error for the control system is given by the 3D pose of the tip of the nose and a 3D reference coordinate point W located at the optical center of the left camera, as illustrated in Fig. 6(b). The goal of the control system is to automatically drive the two PTZ cameras which make up the stereo vision platform, thus ending up with a 6-DoF system (e.g. 2x pan, 2x tilt and 2x zoom). The mathematical model of the camera's servo-drive is the same for the pan, tilt and zoom. In the following, the modeling and control of the left sensor's pan module will be described, the design of the other five units being analogous.

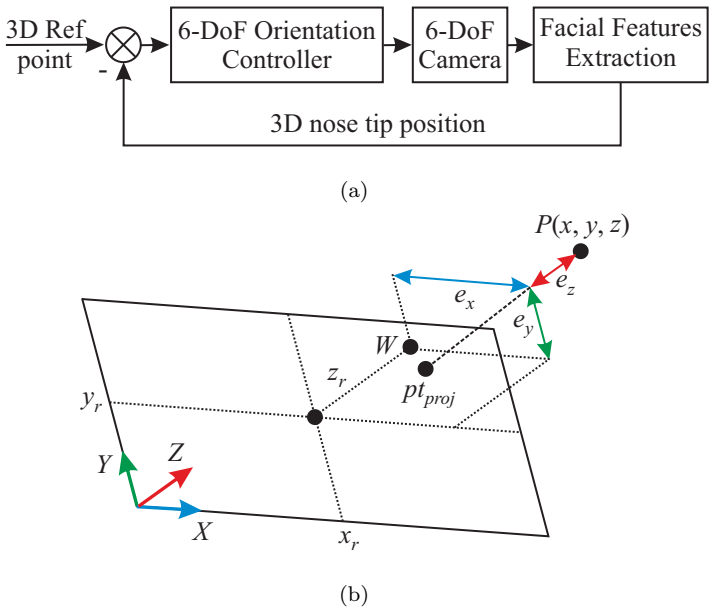


Fig. 6 (a) Basic block diagram of the proposed stereo active vision system. (b) Error definition in the 3D Cartesian space.

The detailed block diagram of the control system for a single module, that is the pan one, is presented in Fig. 7. This mathematical model of the plant represents the servo-drive which adapts the pan orientation of the left sensors of the stereo camera. In our experiments we have used two Sony Evi-D70P® PTZ digital video cameras. The inner-loop within the block diagram from Fig. 7 correspond to a standard servo-drive model for a *Direct Current* (DC)

motor [12]. All six drives of the stereo PTZ system have the same dynamic model. The blocks composing the inner-loop are:

– the plant model $P(s)$, that is the DC motor, modeled as a first order lag element, with the transfer function:

$$P(s) = k_p/(1 + T_p s); \tag{5}$$

– the controller of the inner position loop $R(s)$, described as a *Proportional* (P) controller:

$$R(s) = k_r; \tag{6}$$

– the integrative element $I(s)$ used to integrate the pan's velocity in order to extract its position:

$$I(s) = k_i/s. \tag{7}$$

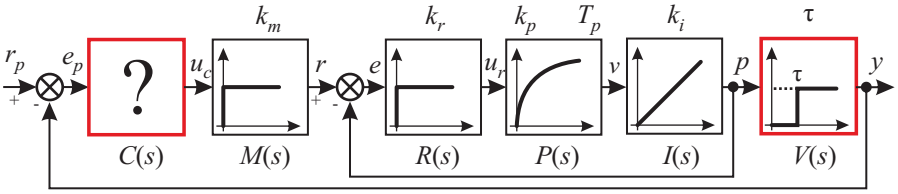


Fig. 7 Detailed block diagram of the pan control system within the proposed active vision architecture.

The signals propagating through the inner-loop are the command signal u_r which drives the DC motor, the output pan velocity v and the integrated pan's position p. The outer-loop elements from Fig. 7 are the following:

– the unknown system controller $C(s)$, designed in the next section;
– the conversion module $M(s) = k_m$, modeled as a P element, used by the architecture to transform measured pixel metric into real-world degrees.
– the visual data processing block $V(s)$ modeled as a time-delay transfer function [9]:

$$V(s) = e^{-s\tau}. \tag{8}$$

where, the delay τ represents the time needed to process a pair of images in order to extract the 3D pose of the human nose.

The r_p signal is the reference position given by the reference coordinate 3D point W. The difference between W and the feedback position variable y, representing the pose of the human head, determines the error signal e_p. e_p represents the error between the current orientation of the camera and the position along the X image axis of the human head, or nose tip:

$$e_p = r_p - y. \tag{9}$$

e_p is further used as input for the overall system controller $C(s)$. For the considered pan case, the final objective of $C(s)$ is to maintain the error along the X Cartesian axis at the lowest possible level.

Having in mind the above explanations, the transfer function of the inner-loop from Fig. 7 can be express as:

$$G_{il}(s) = \frac{R(s)P(s)I(s)}{1 + R(s)P(s)I(s)} = \frac{k_r k_p k_i}{s(sT_p + 1) + k_r k_p k_i} \tag{10}$$

where the values of the parameters have been determined using a standard system identification toolbox.

Starting from Eq. 10, the open-loop transfer function of the entire system can be written as:

$$G_{ol}(s) = C(s)M(s)G_{il}(s)V(s), \tag{11}$$

Replacing the expression of $M(s)$ and $G_{il}(s)$ in the above expression we end up with:

$$G_{ol} = C(s)\frac{k_m k_r k_p k_i}{s(sT_p + 1) + k_r k_p k_i} e^{s\tau}. \tag{12}$$

Although the inner-loop model $G_{il}(s)$ and the $M(s)$ element are both linear, the time-delay introduced by the visual processing algorithms, which calculate the feedback variable y, makes the overall feedback system to be a highly nonlinear one. The process modeled by the $G_{ol}(s)$ transfer function is time-delay dependent, since it is always influenced by the processing time required by the vision component.

4 Control System Design

In this section, the design of the control system's compensator $C(s)$ is detailed, taking into account the high time-delay introduced by the image processing system. In order to control a time-delay system, such as the one considered in this paper, a different control design technique is required as for the case of traditional linear approaches. This is mainly needed because a time-delay component introduces an infinity of poles in the overall transfer function of the system. The reason why this phenomenon takes place is because an exponential function, as the one used in modeling dead time components (see Eq. 8), is expanded as the following time series:

$$e^{-s\tau} \cong 1 - \frac{s\tau}{1!} + \frac{s^2\tau^2}{2!} + \cdots + (-1)^i \frac{s^i\tau^i}{i!} \tag{13}$$

The dead time introduced in the system leads, on the one hand, to its destabilization and, on the other hand, to the decrease of the system's stability

reserve. Starting from the control system's simplified block from Fig. 8, the reduced form of the overall transfer function can be written as:

$$G_{sis}(s) = \frac{C(s)G_p(s)e^{-s\tau}}{1 + C(s)G_p(s)e^{-s\tau}}. \tag{14}$$

where $C(s)$ is the system's compensator and $G_p(s)$ is the transfer function of the considered plant:

$$G_p(s) = M(s)G_{il}(s). \tag{15}$$

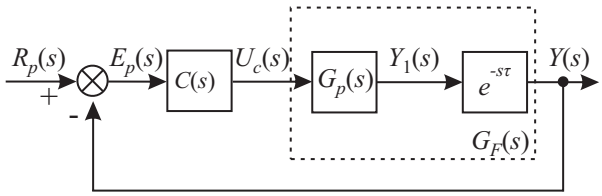


Fig. 8 Block diagram of the control structure containing dead-time.

The dead time present in the system cannot be actually separated from the process since there is no possibility to measure the signal $Y_1(s)$ from Fig. 8. In order to stabilize the plant, a *prediction control* structure can be implemented, such as the *Smith predictor* approach, illustrated in Fig. 9. The core concept of the *Smith predictor* is to move the process's dead time outside the feedback loop of the control system and to determine a controller of a time-delay free system. It is important to notice that such an approach aims at obtaining a transfer function $G^*_{sis}(s)$ which has the time-delay component outside of the feedback loop (see Fig. 10):

$$G^*_{sis}(s) = \frac{C^*_r(s)G_p(s)}{1 + C^*_r(s)G_p(s)}e^{-s\tau}. \tag{16}$$

where $C^*_r(s)$ is the compensator which controls the plant when the time delay element is outside the feedback loop.

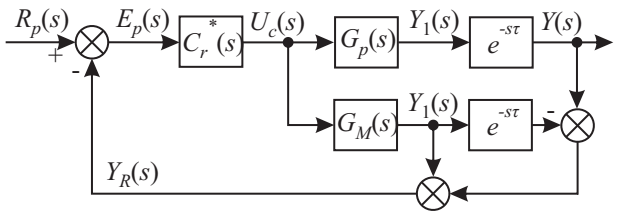


Fig. 9 Basic block diagram of a prediction based control structure.

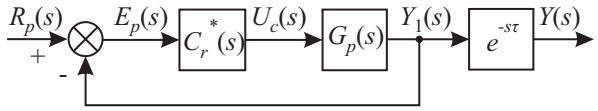


Fig. 10 The movement of the dead-time outside the control structure's feedback loop.

In order to design a controller $C_r(s)$ capable of stabilizing a system having its dead time outside the control loop, an equality between Eq. 14 and Eq. 16 has to be established. Thus, a Smith predictor based compensator is obtained, having the following control structure:

$$C(s) = \frac{C_r^*(s)}{1 + C_r^*(s)G_p(s)[1 - e^{-s\tau}]}. \tag{17}$$

Before computing $C_r(s)$, the synthesis of the $C_r^*(s)$ controller has to be done, as it will be further explained.

4.1 $C_r^*(s)$ Controller Design

Knowing the mathematical model of the open-loop system, the design of the compensator is made according to the *poles placement rule* [17], having the following Lemma in mind:

Lema 1 *Considering a control system with a unitary reaction, described by the process's transfer function $G_p(s)$ and by the process's controller $G_c(s)$, defined as:*

$$\begin{aligned} G_p(s) &= \frac{B_p(s)}{A_p(s)} = \frac{b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \ldots + b_0}{a_n s^n + a_{n-1}s^{n-1} + \ldots + a_0}, \\ G_c(s) &= \frac{Q_r(s)}{P_r(s)} = \frac{q_{n_q}s^{n_q} + q_{n_q-1}s^{n_q-1} + \ldots + q_0}{p_{n_p}s^{n_p} + p_{n_p-1}s^{n_p-1} + \ldots + p_0} \end{aligned} \tag{18}$$

It is assumed that the polynomials $B_p(s)$ and $A_p(s)$ are prime (coprime), that is, they do not have common roots. The arbitrary polynomial $P_c(s)$ of order $n_c = 2n - 1$ is considered. There exist the polynomials $Q_r(s)$ and $P_r(s)$ of order $n_p = n_q = n - 1$ which satisfy the following relation:

$$A_p(s)P_r(s) + B_p(s)Q_r(s) = P_c(s). \tag{19}$$

where $P_c(s)$ represents the characteristic polynomials and is defined as:

$$P_c(s) = p_{n_c}^c s^{n_c} + p_{n_c-1}^c s^{n_c-1} + \ldots + p_0^c. \tag{20}$$

In our work we have concentrated on developing a PID regulator, with its parameters determined according to the following lemma:

Lema 2 *Given a compensator:*

$$G_{reg}(s) = \frac{n_2 s^2 + n_1 s + n_0}{d_2 s^2 + d_1 s}, \tag{21}$$

its PID form can be obtained as:

$$G_{reg}^{PID} = k_p + \frac{k_i}{s} + \frac{k_d}{1 + sT_d}, \tag{22}$$

where,

$$k_p = \frac{n_1 d_1 - n_0 d_2}{d_1^2}, \tag{23}$$

$$k_i = \frac{n_0}{d_1}, \tag{24}$$

$$k_d = \frac{n_2 d_1^2 - n_1 d_1 d_2 + n_0 d_2^2}{d_1^3}, \tag{25}$$

$$T_d = \frac{d_2}{d_1}. \tag{26}$$

The plant $G_p(s)$, that is the stereo active vision system, is described by the following transfer function [23]:

$$G_p(s) = \frac{B_p(s)}{A_p(s)} = \frac{2.83}{s^2 + 1.19s + 2.07}. \tag{27}$$

The computation of a PID controller is conditioned by the choice of the $C_r^*(s)$ compensator, according to Eq. 21 from Lemma 2:

$$C_r^*(s) = \frac{Q_r(s)}{P_r(s)} = \frac{q_2 s^2 + q_1 s + q_0}{p_2 s^2 + p_1 s}. \tag{28}$$

The choice of the characteristic polynomial $P_c(s)$ is made such that its order is equal to the order of the left hand side expression from Eq. 19. $P_c(s)$ is obtained as a product of two second order polynomials. While, the first polynomial is intended to fulfill the imposed performances, the second one aims at constraining the values of the poles to be between three to five times higher than the natural frequency of the first polynomial [39]. Thus, the characteristic polynomial will be written as:

$$P_c(s) = (s^2 + 2\zeta\omega_n s + \omega_n^2)(s + p_c)^2, \tag{29}$$

where ζ is the *damping factor* and ω_n the natural frequency. The computation of the coefficients of the first polynomial is performed as follows [8]:

– ζ is derived from the overshooting value:

$$\zeta = -\frac{\ln(M_v)}{\sqrt{\pi^2 + ln^2 M_v}} = \frac{|\ln(M_v)|}{\sqrt{\pi^2 + ln^2 M_v}} = 2.166; \tag{30}$$

– the natural frequency is chosen according to the 2% criterion:

$$\omega_n = \frac{4}{\zeta t_s} = 4.607 \text{ rad/sec}, \tag{31}$$

where, t_s is the settling time.

By choosing the value of the p_c pole to be $3-5$ of the natural frequency and replacing it in Eq. 29, we obtain the next polynomial:

$$P_c(s) = (s^2 + 19.96s + 21.23)(s + 23)^2. \tag{32}$$

By replacing Eq. 27, 28 and 32 in Eq. 19 the following relation is obtained:

$$(p_1 s + p_2 s^2)(s^2 + 1.19s + 2.07) + 2.83(q_2 s^2 + q_1 s + q_0) = (s^2 + 19.96s + 21.23)(s + 23)^2. \tag{33}$$

The inequalities from Eq. 33 are obtained from the next matrix equation:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 1.19 & 0 & 0 & 0 \\ 1.19 & 2.07 & 0 & 0 & 2.83 \\ 2.07 & 0 & 0 & 2.83 & 0 \\ 0 & 0 & 2.83 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ q_0 \\ q_1 \\ q_2 \end{bmatrix} = \begin{bmatrix} c_{pc4} \\ c_{pc3} \\ c_{pc2} \\ c_{pc1} \\ c_{pc0} \end{bmatrix}. \tag{34}$$

Since the coefficients matrix is nonsingular, it can be resolved for p_1, p_2, q_0, q_1 and q_2. Hence, the following PID compensator is obtained:

$$C_r^*(s) = \frac{7.593s^2 + 62.42s + 61.39}{0.01541s^2 + s}., \tag{35}$$

The components of Eq. 35 can be extracted according to Eq. 23 - 26, resulting in the following values: $k_p = 61.473$, $k_i = 61.39$, $k_d = 6.645$, $T_d = 0.0154$. The $C_r^*(s)$ regulator is further inserted into Eq. 17 with the purpose of obtaining the process's controller $C_r(s)$. The system's response, for the block diagram form fig. 9 is shown in Fig. 11. As can be seen, the value of the overshooting is 23.8%, while the settling time has a value of 0.81 sec.

4.2 Overall Compensator Design

Having in mind the results presented above, the overall system's controller $C_r(s)$ can be determined according to Eq. 17, where $C_r^*(s)$ is replaced by a PID compensator, defined as in Eq. 35 and $G_p(s)$ is replaced with Eq. 27:

$$C(s) = \frac{(as^2 + bs + c)(s^2 + fs + g)}{(s + ds^2)(s^2 + fs + g) + h(as^2 + bs + c)(1 - e^{-s\tau})}. \tag{36}$$

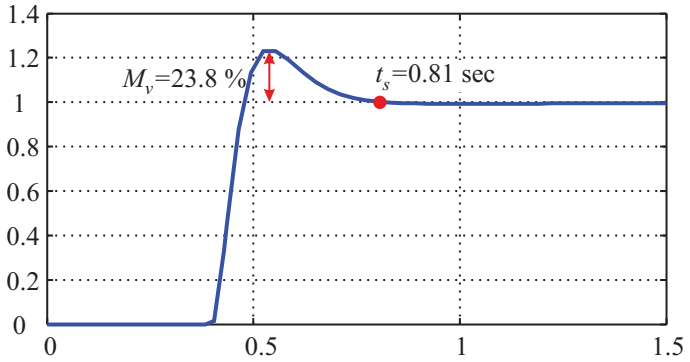


Fig. 11 The response of the proposed time-delay control system in the time domain.

The condensed block diagram for the feedback system is presented in Fig. 12. After grouping the terms in Eq. 36 the expression of $C_r(s)$ becomes:

$$C_r(s) = \frac{as^4 + (af + b)s^3 + (ag + bf + c)s^2 + (bg + cf)s + cg}{ds^4 + (df + 1)s^3 + (dg + f)s^2 + gs + h(as^2 + bs + c)(1 - e^{-s\tau})}. \tag{37}$$

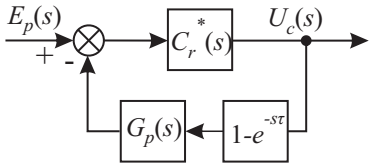


Fig. 12 Practical controller implementation.

In order to implement the compensator as a discrete controller, the continuous transfer function determined above has been transformed to the discrete time domain using the *Backward (Rectangular) Rule*:

$$s = \frac{1 - z^{-1}}{T_e} \tag{38}$$

where T_e is the *sampling time*. According to [14], the sampling frequency for a digital control system must be 4 to 20 times higher than the frequency of the closed-loop system. In our active vision application, where the system's frequency is 2.5 Hz, the value of the sampling frequency is between 10 Hz and 50 Hz. Hence, the required value of the sampling time T_e should be between 0.1 sec and 0.02 sec.

By discretizing Eq. 37 according to 38, and taking into account that $\tau = \nu T_e$, $\nu \in \mathbb{R}$, the transfer function of the numeric compensator is obtained:

$$C_r(z^{-1}) = \frac{n_4 z^{-4} + n_3 z^{-3} + n_2 z^{-2} + n_1 z^{-1} + n_0}{d_6 z^{-6} + d_5 z^{-5} + d_4 z^{-4} + d_3 z^{-3} + d_2 z^{-2} + d_1 z^{-1} + 1} = \frac{U_c(z^{-1})}{E_p(z_{-1})}, \tag{39}$$

where $\nu = \tau / T_e = 4$. The operational transfer function of the numeric compensator is:

$$\begin{aligned}(1 + d_1 q^{-1} + d_2 q^{-2} + d_3 q^{-3} + d_4 q^{-4} + d_5 q^{-5} + d_6 q^{-6}) u_c[t] = \\ (n_0 + n_1 q^{-1} + n_2 q^{-2} + n_3 q^{-3} + n_4 q^{-4}) \varepsilon_p[t].\end{aligned} \tag{40}$$

The final form of the numeric regulator is finally derived from Eq. 40:

$$\begin{aligned}u_c[t] = &- d_1 u_c[t-1] - d_2 u_c[t-2] - d_3 u_c[t-3] - d_4 u_c[t-4] - \\ &- d_5 u_c[t-5] - d_6 u_c[t-6] + n_0 \varepsilon_p[t] + n_1 \varepsilon_p[t-1] + \\ &+ n_2 \varepsilon_p[t-2] + n_3 \varepsilon_p[t-3] + n_4 \varepsilon_p[t-4],\end{aligned} \tag{41}$$

where $\varepsilon_p[t] = r_p[t] - y[t]$ (see Fig 8). Eq. 41 is used for controlling each DoF of the stereo active vision system. The numeric algorithm can be implemented either on dedicated hardware, or on typical PC computers. As it will be shown in the next section, the proposed approach performs optimally in the context of the considered active gaze following scenario.

5 Performance Evaluation

5.1 Experimental Setup

The stereo acquisition system was placed at 1.7 m above the ground, as shown in the experimental setup image from Fig. 13. The sensors become active when a human observer appears inside the camera's *Field of View* (FOV). The face detection procedure, followed by nose detection, starts the active tracking process. The experiments were performed in an indoor room, using natural and artificial illumination. The subject moved inside the room in a random way, in frontal and profile poses, covering an area of about 9 m^2. For additional details please refer to the videos accompanying this paper.

5.2 Face Detection and Tracking Results

The evaluation of the proposed facial features detection was performed using images with different poses and distances. The subject modified its position and orientation in an interval ranging from 0.3 m to 2.2 m along the pan-tilt directions and from 0.2 m to 2.8 m along the camera's zoom distance.

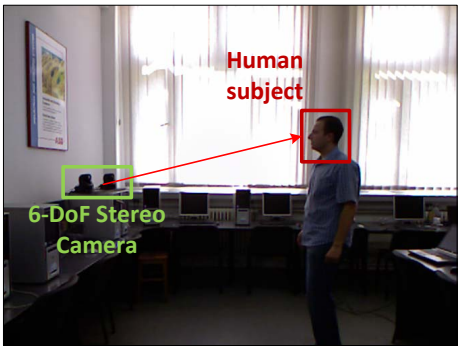


Fig. 13 Experimental setup composed of a 6-DoF stereo camera and a human subject moving freely.

The facial feature detection algorithms, that is face and nose, were then applied while, in parallel, both these features were earmarked through visual inspection. The manual visual inspection corresponds to the *ground truth* used for establishing the accuracy of the method. The Euclidean distance E_d on the 2D image plane, considered as nose detection accuracy, was then computed between the automatic estimated values from the proposed algorithm and the manually determined ground truth. The variation of E_d for a number of samples can be seen in Fig. 14. The features detection algorithm was consequently considered to produce a "Hit" when the error was lower than an adapting threshold $t_h[px]$ and a "Miss" otherwise (meaning that the error was big enough to imply that the corresponding detector failed to segment the correct features), as shown in Tab. 1. t_h is defined as:

$$t_h = \frac{\alpha \cdot (W \cdot H - w_l \cdot h_l) + \beta \cdot w_h \cdot h_h}{w_h \cdot h_h - w_l \cdot h_l}, \tag{42}$$

where w_l and w_h represent the lowest and highest width of the face region, h_l and h_h are the lowest and the highest values of the face's height and W and H are the image's width and height. $\alpha = 18$ and $\beta = 2$ represent normalizing factors.

t_h is modified according to the distance between the camera and the human subject, that is, with respect to the size of the face in the image. If a face covers a wider area of the processed image, then the value of t_h is will be higher. In other words, the structure of the t_h threshold in Eq. 42 relates the size of the face to the size of the image. That is, the larger a face region is, the better the face detector should perform.

The Hit/Miss results for nose recognition are considered for frontal, as well as for left and right profiles of the face. The "NA" (Not Available) value in Tab. 1 signifies that no face candidate (neither frontal or profile) was detected. The values presented in Tab. 1 show the accuracy of the face detector over the input HMI sequence from Fig. 14, accuracy which affects the precision of nose segmentation. A thorough description and evaluation of the face recogni-

tion system used in our work can be found in the seminal work of Viola and Jones [37].

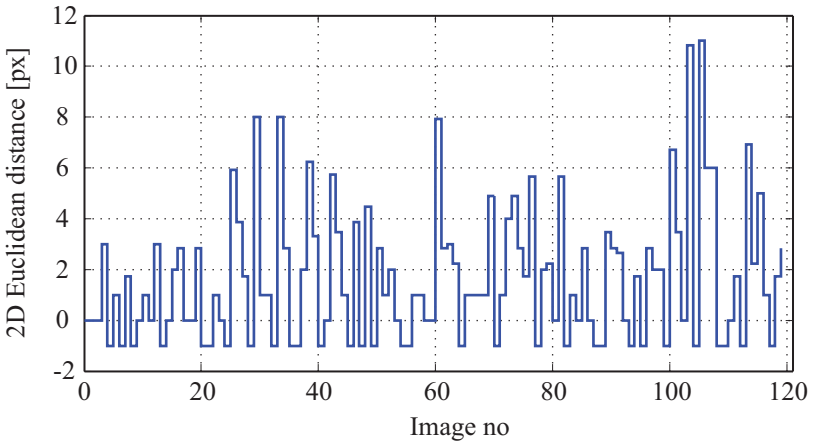


Fig. 14 2D Euclidean distance describing the accuracy of the nose detection algorithm.

Table 1 Number of face detection results in the video sequence from Fig. 14, presented as Hit/Miss/NA values.

	Profile left	Face	Profile right
Hit	32	22	41
Miss	7	4	4
NA		10	

The detection hit probability $P(hit)$ can be statistically determined from the obtained Hit/Miss/NA results (see Tab. 2). The mean error was further used to establish the standard deviations σ for each corresponding sensor model.

Table 2 Statistical data of facial feature detection algorithm.

Detector	$P(hit)$	σ
Head	0.821	4.95
Nose	0.741	1.919

The introduced delay, against which the active tracking system has to cope, varies stochastically and depends on the effort the detection methods need for processing a pair of stereo images. The variable time-delay for the described vision algorithms is illustrated in Fig. 15. As can be seen, the dead-time introduced at frame 65 is signifficantly higher. This happens due to a fast

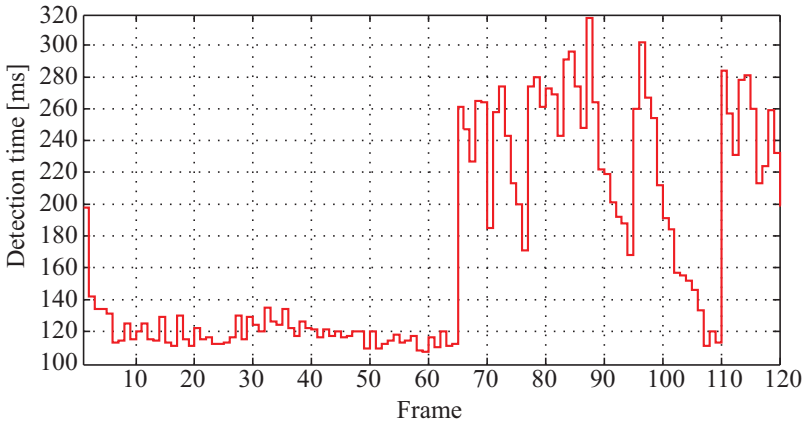


Fig. 15 The variable time-delay introduced by the vision algorithms during the active tracking procedure.

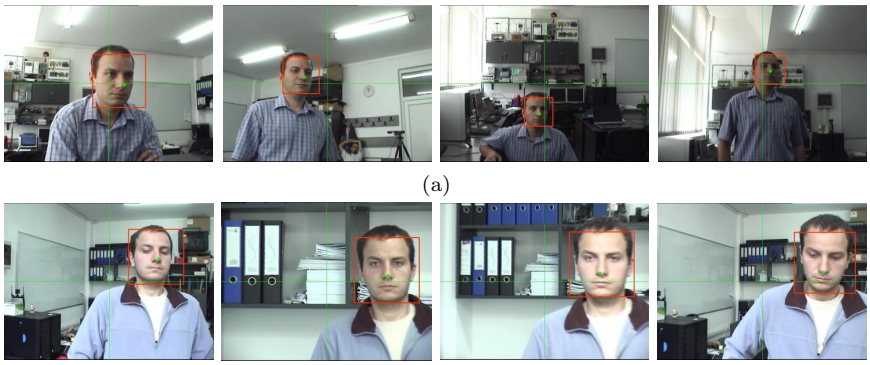
movement of the human subject located at a large distance from the camera. Since the reduced ROI for face searching could not be calculated, as described in Section 2, the time required for computing the facial features is higher. As pointed out in [23], a maximum delay of $0.48s$ can be supported by the control system, without destabilizing it.

5.3 Stereo Active Vision System's Response

The stereo camera was controlled for pan, tilt and zoom values ranging in the intervals $[-60°, 60°]$, $[-40°, 40°]$ and $[0\text{mm}, 10\text{mm}]$, respectively. For the zoom case, the controlled variable is actually the focal length of the camera which varies from 0mm to 10mm. The control commands were applied individually for the left and the right cameras, while the rotational velocity of the camera was modified using three different values: $30°/\,\text{sec}$, $50°/\,\text{sec}$ and $70°/\,\text{sec}$, for images acquired at a resolution of 640×480 pixels. Analogously, the zoom control was tested for the focal length's translational velocities of 2mm/sec, 5mm/sec and 7mm/sec. Nose samples acquired during active tracking can be seen in Fig. 16.

Active tracking results are presented in Fig. 17, where the pan-tilt real values and the position actuator variable are illustrated for different speeds. The left camera movement is presented in Fig. 17(a), 17(c) and 17(e), while the right camera movement is depicted in Fig. 17(b), 17(d) and 17(f), repsectively. Also, in Fig. 18, performance evaluation results regarding the zoom adaptation are illustrated. As can be seen from the diagrams, at a distance above 2.5 m the zoom's focal length achieves its maximum value of 10 mm.

For each experimental session, the mean error is calculated as $err[deg] = pos_{est} - pos_{real}$. The obtained error values are summarized in Tab. 3 and Tab. 4, in comparison to results delivered by our previous published method [24]. As



(a)

(b)

Fig. 16 Samples acquired during active nose tracking for the case of the pan-tilt (a) and zoom (focal length) (b) variations. The center of the green cross in the middle of the image represents the reference value for the control system, while the nose tip is the feedback variable (best viewed in color).

can be seen, based on the proposed control approach, the stereo camera was able to track nose features with good accuracy for speeds less than 70°/sec. For higher values, the camera fails to follow the features since, because of the angular speeds of the pan and tilt, an image blur effect appears in the acquired images. The different obtained mean errors, such as the case of the pan left at 70° being notable higher than the other values, is due to the nonlinearities of the PTZ drives, as well as the random movements of the human subject in the experimental area.

Table 3 Statistical position errors for pan and tilt movements, in comparison to method [24].

Item	Speed [°/sec]	Mean Error [deg] Current approach	Mean Error [deg] Method [24]
Pan Left	30	0.358	2.5365
	50	0.797	4.6072
	70	0.33	7.4130
Tilt Left	30	0.465	1.1618
	50	0.313	3.3475
	70	0.482	4.1415
Pan Right	30	0.15	14.4943
	50	0.227	13.97
	70	0.222	15.2349
Tilt Right	30	0.358	2.6462
	50	0.315	2.082
	70	0.441	3.2988

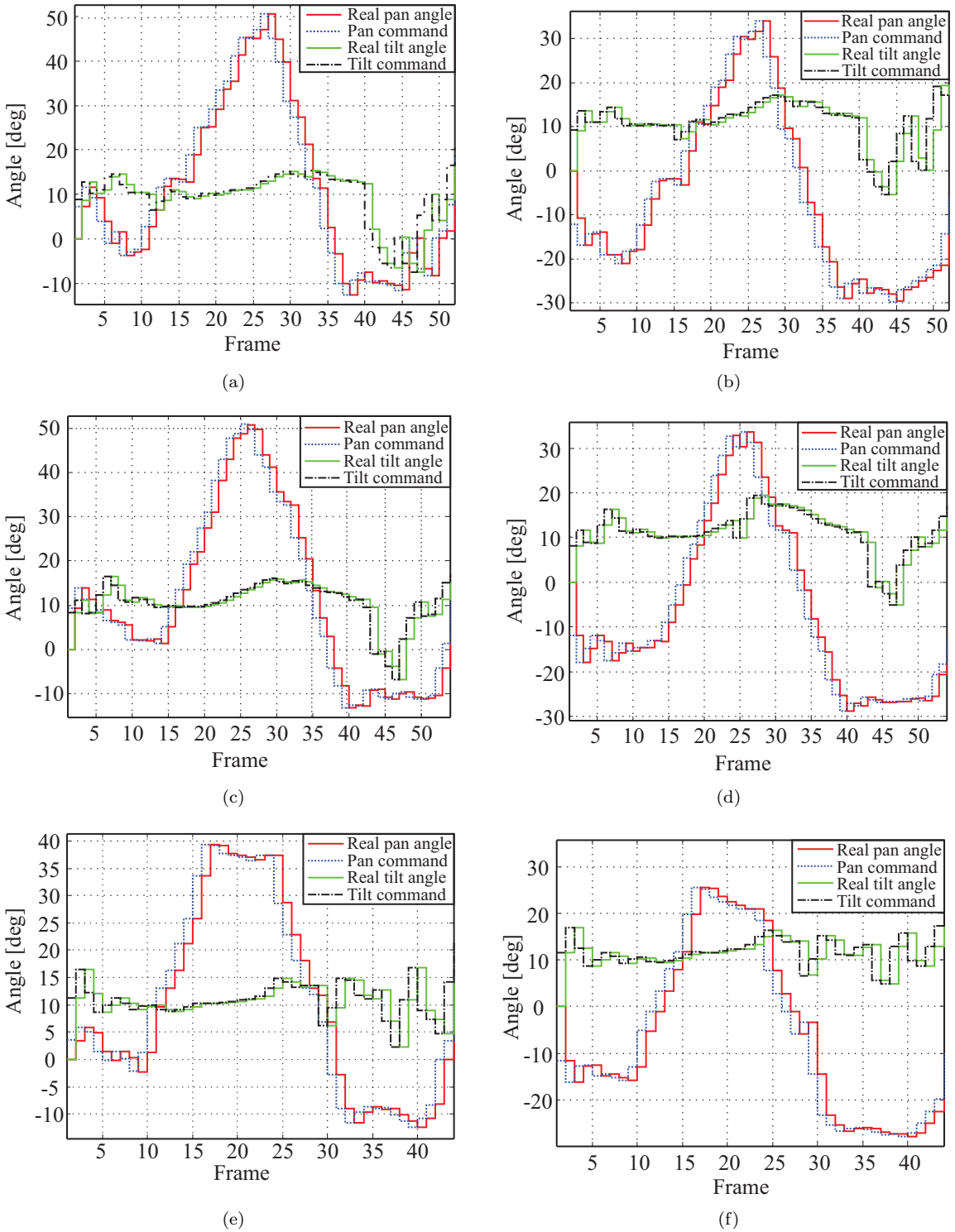
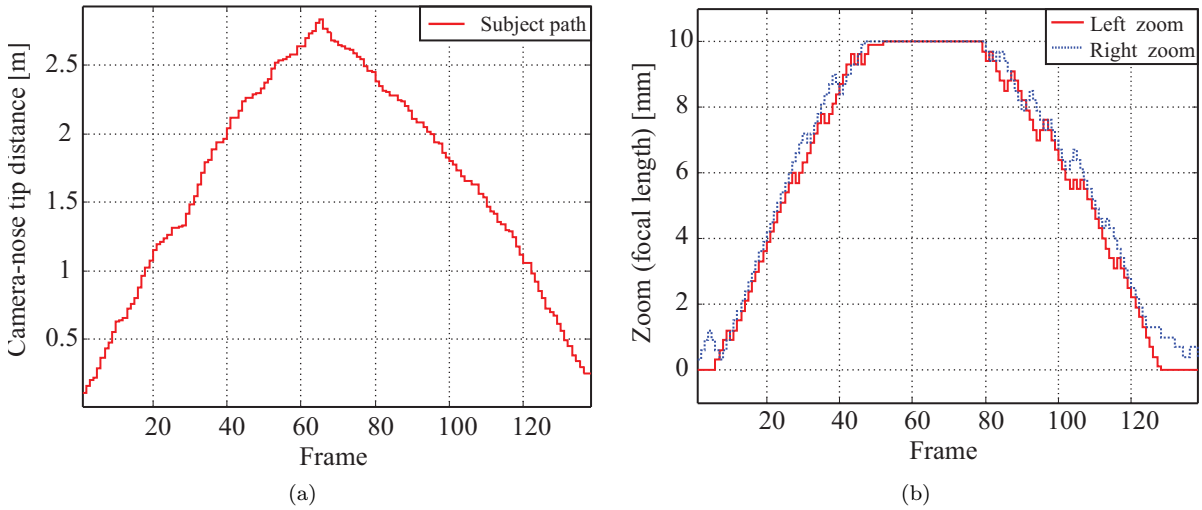


Fig. 17 Active nose tip tracking for different pan and tilt speeds. (a,b) 30°/sec. (c,d) 50°/sec. (e,f) 70°/sec.



(a) (b)

Fig. 18 Nose tip active tracking results using the zoom (focal length) controller. (a) Human subject - camera distance. (b) Adaptation of the focal length with respect to the human subject - camera distance from (a).

Table 4 Statistical position errors when controlling the focal length.

Item	Speed [mm/ sec]	Mean Error [mm]
Zoom Left	2	0.582
	5	0.434
	7	0.263
Zoom Right	2	0.895
	5	0.748
	7	0.576

6 Conclusion and future work

The work presented in the paper deals with the realization of a stereo active vision framework for HRI which can cope with the high time-delay values introduced by the image processing algorithms. As can be seen from the experimental results section, the proposed approach has been proven stable in tracking the nose feature of the human subject in different active tracking scenarios, provided that the maximal rotational velocities of the sensors are limited. The proposed method for designing the overall system controller has better results, compared with the previous work of the authors, based on the development of a proportional controller. Our new results overcame the limitations of the P controller, more exactly the new system doesn't have steady state error and the camera's oscillations are eliminated.

The methods for 3D facial feature estimation presented in this paper are being integrated into a probabilistic model for gaze tracking [22], a part of

an ongoing work on the extension of a hierarchical Bayesian framework for multisensory active perception presented in [13]. The framework can be used to drive the active vision system proposed in this paper, providing a powerful solution for applications in the field of Human-Robot Interaction.

Acknowledgements This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contracts number POSDRU/88/1.5/S/59321 and POSDRU/89/1.5/S/59323

References

1. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active Vision. Inter. Journal of Computer Vision **1**(4), 333–356 (1988)
2. Ayasun, S., Gelen, A.: Stability Analysis of a Generator Excitation Control System with Time Delays. Electrical Engineering **91**(1), 347–355 (2010)
3. Bohren, J., Rusu, R.B., Jones, G., Marder-Eppstein, E., Pantofaru, C., Wise, M., Moesenlechner, L., Meeussen, W., Holzer, S.: Towards Autonomous Robotic Butlers: Lessons Learned with the PR2. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China (2011)
4. Bolme, D.S., Draper, B.A., Beveridge, J.R.: Average of synthetic exact filters. In: International Conference on Computer Vision and Pattern Recognition CVPR, pp. 2105–2112 (2009)
5. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Sebastopol, Canada (2008)
6. Brooks, A., Arkin, R.: Behavioral Overlays for Non-verbal Communication Expression on a Humanoid Robot. Autonomous Robots **22**(1), 55–74 (2007)
7. Bruce, V., Young, A.: In the Eye of the Beholder: The Science of Face Perception. Oxford University Press, Oxford, United Kingdom (2000)
8. Comnac, V., Coman, S., Boldişor, C.: Sisteme liniare continue. Universităţii Transilvania, Brasov, Romania (2009)
9. Corke, P.: Visual Control of Robots : High-performance Visual Servoing. Research Studies Press Ltd, Taunton, England (1996)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: C. Schmid, S. Soatto, C. Tomasi (eds.) International Conference on Computer Vision and Pattern Recognition CVPR, vol. 2, pp. 886–893. INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334 (2005). URL `http://lear.inrialpes.fr/pubs/2005/DT05`
11. De Paor, A.M., O'Malley, M.: Controllers of Ziegler-Nichols Type for Unstable Process with Time Delay. Inter. Journal of Control **49**(4), 1273–1284 (1989)
12. Dorf, R., Bishop, R.: Modern Control Systems. Prentice-Hall, Inc., New Jersey, USA (2010)
13. Ferreira, J.F., Castelo-Branco, M., Dias, J.: A Hierarchical Bayesian Framework for Multimodal Active Perception. Adaptive Behavior (2012). Published online ahead of print
14. Franklin, G.F., Powell, D.J., Workman, M.L.: Digital Control of Dynamic Systems, 3 edn. Prentice-Hall, Inc., Boston, USA (1997)
15. Freund, Y., Schapire, R.: A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. Journal of Computer and System Sciences **55**(1), 119–139 (1997)
16. Gaudette, D., Miller, D.: When is the Achievable Discrete-time Delay Margin Nonzero? IEEE Trans. on Automatic Control **56**(4), 886–890 (2011)
17. Goodwin, G., Graebe, S., Salgado, M.: Control System Design. Prentice-Hall, Inc, New Jersey, USA (2001)

18. Grigorescu, S., Macesanu, G., Cocias, T., Puiu, D., Moldoveanu, F.: Robust Camera Pose and Scene Structure Analysis for Service Robotics. Robotics and Autonomous Systems **59**(11), 899–909 (2011)
19. Grigorescu, S.M., Lth, T., Fragkopoulos, C., Cyriacks, M., Grser, A.: A bci-controlled robotic assistant for quadriplegic people in domestic and professional life. Robotica **30**, 419–431 (2012). DOI 10.1017/S0263574711000737. URL http://dx.doi.org/10.1017/S0263574711000737
20. Huang, J., Lewis, F.L., Liu, K.: A Neural Net Predictive Control for Telerobots with Time Delay. Journal of Intelligent & Robotic Systems **29**, 1–25 (2000)
21. Kumar, S., Micheloni, C., Piciarelli, C.: Stereo localization using dual ptz cameras. In: International Conference on Computer Analysis of Images and Patterns, vol. 5702/2009, pp. 1061–1069. Munster, GE (2009)
22. Macesanu, G., Ferreira, J.F., Dias, J.: A Bayesian Hierarchy for Gaze Following. In: The 5th Inter. Conf. on Cognitive Systems. TU Vienna, Austria (2012)
23. Macesanu, G., Grigorescu, S., Comnac, V.: Time-delay Analysis of a Robotic Stereo Active Vision System. In: Proc. of the 15th Inter. Conf. on System Theory, Control, and Computing, pp. 1–6. Sinaia, Romania (2011)
24. Macesanu, G., Grigorescu, S., Ferreira, J.F., Dias, J., Moldoveanu, F.: Real Time Facial Features Tracking Using an Active Vision System. In: Proc. of the 13th Inter. Conf. on Optimization of Electrical and Electronic Equipment, pp. 1493–1498. Brasov, Romania (2012)
25. Michiels, W., Vyhlidal, T., Zitek, P.: Control Design for Time-delay Systems Based on Quasi-direct Pole Placement. Journal of Process Control **20**(3), 337–343 (2010)
26. Mozos, O.M., Marton, Z.C., Beetz, M.: Furniture Models Learned from the WWW – Using Web Catalogs to Locate and Categorize Unknown Furniture Pieces in 3D Laser Scans. Robotics & Automation Magazine **18**(2), 22–32 (2011)
27. Nickel, K., Stiefelhagen, R.: Visual Recognition of Pointing Gestures for Human-Robot Interaction. Image and Vision Computing **25**(12), 1875–1884 (2007)
28. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: International Conference on Pattern Recognition, vol. 1, pp. 582–585. Jerusalem, Israel (1994)
29. Park, C.B., Lee, S.W.: Real-time 3D Pointing Gesture Recognition for Mobile Robots with Cascade HMM and Particle Filter. Image and Vision Computing **29**(1), 51–63 (2011)
30. Pateraki, M., Baltzakis, H., Kondaxakis, P., Trahanias, P.: Tracking of facial features to support human-robot interaction. In: Proceedings of the 2009 IEEE international conference on Robotics and Automation, ICRA'09, pp. 2651–2656. IEEE Press, Piscataway, NJ, USA (2009). URL http://dl.acm.org/citation.cfm?id=1703775.1703879
31. Siciliano, B., Khatib, O.: Springer Handbook of Robotics. Springer, Berlin, Germany (2008)
32. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in Engagement for Humans and Robots. Artificial Intelligence **166**(2), 140–164 (2005)
33. Silva, G., Aniruddha, D., Bhattacharyya, S.P.: PID Controllers for Time-Delay Systems, 1 edn. Birkhauser Boston, New York, USA (2004)
34. Silva, G., Datta, A., Bhattacharyya, S.: Determination of Stabilizing Feedback Gains for Second-order Systems with Time Delay. In: Proc. of the 2001 American Control Conf., pp. 4658–4663. Arlington, Virginia (2001)
35. Skodras, E., Fakotakis, N.: An Unconstrained Method for Lip Detection in Color Images. In: Proc. of the 2011 IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 1013–1016 (2011)
36. Sommerlade, E., Benfold, B., Reid, I.: Gaze directed camera control for face image acquisition. In: IEEE International Conference on Robotics and Automation, pp. 4227–4233 (2011)
37. Viola, P., Jones, M.: Rapid Object Detection Using a Boosted Cascade of Simple Features. In: Proc. of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 511–518. Kauai, USA (2001)
38. Wan, D., Zhou, J.: Stereo vision using two ptz cameras. Computer Vision and Image Understanding **112**(2), 184 – 194 (2008). DOI 10.1016/j.cviu.2008.02.005. URL http://www.sciencedirect.com/science/article/pii/S1077314208000313

39. Wang, Q.G., Zhang, Z., Astrom, K.J., Chek, L.S.: Guaranteed Dominant Pole Placement with PID Controllers. Journal of Process Control **19**(2), 349–352 (2009)