Real Time Facial Features Tracking using an Active Vision System

Gigel Măceşanu*, Sorin Grigorescu*, João Filipe Ferreira†, Jorge Dias† and Florin Moldoveanu*
*Department of Automation, Electronics and Computers, Transilvania University of Brasov, Romania
Email: {gigel.macesanu, s.grigorescu, moldof}@unitbv.ro
†Institute of Systems and Robotics, University of Coimbra, Portugal, Email: {jfilipe, jorge}@isr.uc.pt

***Abstract*—In this paper, an active vision system used for real time facial features tracking is presented, where a human head region is searched using a stereo camera. This region is divided such that a person's nose will be searched in a smaller spatial domain. The 2D coordinates of the nose are determined based on a Haar classifier and colour segmentation. Its 3D coordinates, computed via the constraints of the epipolar geometry, represent a reference signal used within a control system for controling the stereo camera's pan and tilt. A method for determining the system's parameters and for designing a proper controller is further proposed. The controller is design such that it integrates the maximum time delay that can be inserted into the control system by the machine vision component. Finally, the overall architecture is tested in a real facial feature tracking application.**

I. Introduction

Robotic systems that realize human behaviour imitation represent are currently heavy investigated due to their important role in applications presenting high risk of accidents to humans. The perceived scene should include, as much as possible, information about the surrounding environment. Whilst the newest video technologies used as artificial video systems are more and more robust, an adaptive system, able to modify its pose according with the pose of different objects of interest, is indicated for cluttered environments. One of the main objectives of a stereo active vision system is to track objects and reconstruct their poses in a virtual 3D space. In this sense, a stereo camera system capable to realize the movement on pan and tilt directions, while the zoom component is used to focus on specific objects, can be used.

In this work an active vision system used for robotic perception is presented. Similar approaches can be found in older papers such as [1] and [2]. Both present first uses of active camera systems as modules able to modify camera parameters, such as: *position*, *orientation*, *focus*, *zoom*, *aperture* and *vergence*. All these modifications are realized as a response to an external stimulus [2]. In literature, the active vision concept is usually corelated with visual servoing [3]. A visual servoing system uses the video information from a video camera to control the movement of a robot and can be classified, according to [3], in *position* and *image based visual servoing*. Modern active vision systems are used to create humanoid robotic heads, used to analyze a scene using 3DoF [4] or to construct a surveillance platform skilled to track multiple persons, as is presented in [5]. Fayman [6] and Yao [7] present active vision systems that involve the camera's zoom adaptation, which considers the problem of controlling the focal length, in order to keep a constant-sized image of an object moving along the camera's optical axes.

In the presented paper, the overall time-delay that is introduced by image processing algorithms is considered while integrating the feedback information for controling an active stereo camera system. To control such a system, a *Proportional* (P) controller is implemented, whose stability interval is determined based on [8] and [9].

Facial feature detection is a complex process, determined by the individuality of each person. All methods used for this task are related to head detection. To realise such a task, methods such as Haar classification or color based segmentation can be used. The first methods use the principle developed by Viola and Jones [10] and can be used for frontal and profile face detection. The second category use particular characteristics of the head, such as the color of the skin. In [11], a method that uses two stages for head detection is proposed. Namely, it firstly searches skin regions and secondly locates the faces using a region property measure. Starting from the head region, specific facial features can be determined. The nose is considered a very important feature since it is located in the head's central position and can appear even in profiles of faces. Currently, nose detection algorithms are not so common as the head detections ones. Nevertheless the Haar classifier can be succesfully applied for its detection.

The rest of the paper is organized as follows. In Section II, the head and nose detection is presented together with the 3D position estimation approach. The general mathematical model of the considered active vision architecture is given in Section III, where the stability of the tracking system is analysed. Section IV presents experimental results. Finally, conclusions are stated in Section V.

II. 3D Facial Feature Pose Estimation

Robotics systems that realize human behaviour imitation use information about the human observer in order to understand the geometrical relation between the observer and the robot. In order to deal with this chalange, we need to construct a method that is able to determine the 2D and 3D pose of an object of interest. Obtaining the 3D pose is possible if there is enough information regarding the acquisition system, such as the internal camera parameters.

A. Head Detection

The head detection algorithm is applied on a color image, using a Haar classifier for head recognition. In this sense, two different classifiers, for frontal and profile poses, are used. The resulted value is a head *Region of Interest* (ROI). To perform this detection we use the method from [10], taking into account the next three stages for detection:

a) construct a new image where the feature can be determined very fast. The pixel values of the new images are the sum of all above and left pixels values relating the considered pixel position:

$$g(x,y) = \sum_{x^{'} \leq x, y^{'} \leq y} f(x^{'}, y^{'}), \tag{1}$$

where, $g(x,y)$ is the new image and $f(x^{'}, y^{'})$ is the original image;

b) use AdaBoost to select some critical features and realize a classification;

c) separate the background from foreground information and provides statistical guarantees that discarded regions can contain the object of interest [10].

This algorithm is applied for each image delivered by the stereo vision system. To achieve a proper head detection, a series of detection parameters must be tuned, namely:

- *Scale factor*: is used to determine the object scale difference, between each searching;
- *Hit number*: during face searching, an object is considered to be a face if it has at least a number of hits equal with the hit number parameter;
- *Object size*: is used to set the minimal search region, where a head is searched;
- *Head selection*: when multiple faces are in the images, only the head with the highest region confidence is selected;

The obtained head region is used as input for the next nose detection stage.

To reduce the head detection time, we use a dynamic window adjustment approach. This method works in two individual stages:

a) *Initialisation*: this stage is computed in two different situations: first, when the program is starting and, secondly, when a head detection fails to detect a head. During this process the head's ROI is determined and saved in the temporary variable v_{tmp};

b) *Window adjustment*: when in previews frames a head has been determined, that is $v_{tmp} \neq 0$, a new searching window is constructed, which has as kernel v_{tmp} and is surrounded by blocks with the same size as the kernel.

In Fig. 1, the construction of the new window, centred on previews head ROI, is presented. The hash squares represent the extended window. The head position can caused two particular situations: the head is closed to the camera or the head is in a image corner position. In both cases, the new window is constructed such that cover all image, in first case, or to cover regions inside the image that could contain the head. These cases appear when the person movement is faster than the camera pose adjustment.

B. Nose Detection

Human facial features are characterised by individual proprieties. Many of these characteristics are related with their poses, color or prominence form. In the current application, the nose tip, and the fact that it has a specific color, is used for nose detection. Thereby, in the proposed nose detection algorithm, a color based segmentation followed by a nose contour identification process is used. The nose detection method has its roots in the lips detection algorithm from [12]. The stages involved in nose detection are developed as follows.

1) Face detection: The face ROI determined previously is used to construct a search region for nose tip recognition. This new region represents a stretch of the initial head region. It is determined according with the head's width and height. More accurately, it represents 33% of the central head region. The values are empirically chosen, since the nose is situated in the central of the head region.

2) Colour Based Segmentation: The image acquisition process delivers images in RGB format. In order to have a segmentation which is, as much as possible, invariant to variable lighting conditions, the RGB image is converted to the L*a*b color space representation. The Lab color space format is composed of a ”Luminance” value represented by L, and two color channels (a and b). All steps implied in color based segmentation can be see in Fig. 2, where the nose region is highlighted. In order to have a better contrast and a good intensity distribution, the image's histogram is equalized. For extracting nose information, a logical AND operation, between the a color channel (after histogram equalization) and the resulted image is applied. A morphological gradient filter, applied on the b channel image, is used to isolate perimeters of existing blobs. The new perimeters are treated as objects [13]:

$$gradient(src) = dilate(src) - erode(src), \tag{2}$$

where, the dilatation and erosion process can be interpreted as a convolution process between an image part and a kernel [13]. The kernel can be of any shape (e.g. solid square or

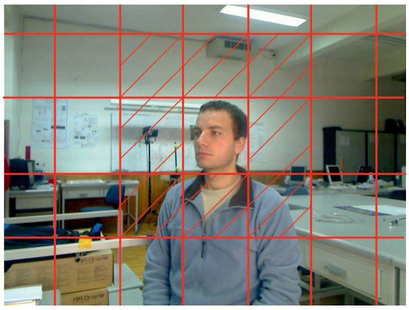


Fig. 1. A new ROI, with adjusted window.

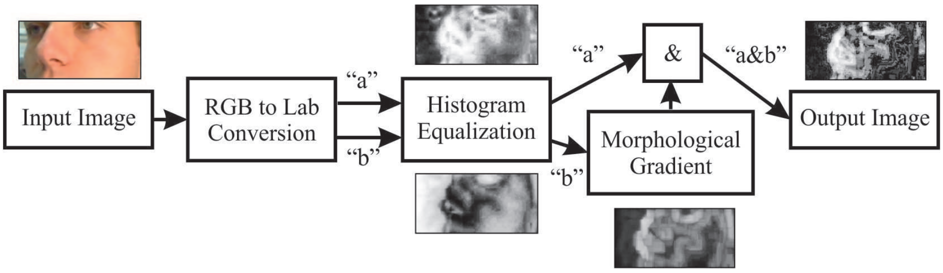


Fig. 2. Nose detection using colour based segmentation.

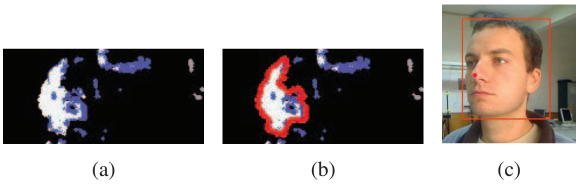

(a) (b) (c)

Fig. 3. Nose detection. a) K-means clustering results. (b) Nose contour detection. (c) Nose tip localization.

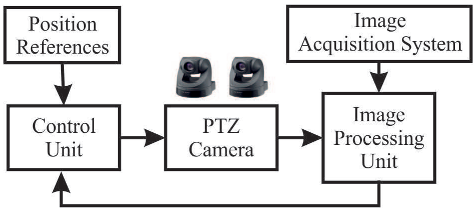


Fig. 4. Overall block diagram of the control system unit.

disks) and have a single anchor point. The effect of dilatation can be seen as a local maximum operator, while the erosion process can be seen as a minimum operator.

3) Pixels classification and nose detection: Using the results from the previews steps, all the pixels in the ROI are grouped into clusters. To do this we use *K-means clustering* algorithm. This method, groups all pixels into a desired number of clusters. The number of clusters was chosen in this work as 4. The K-means clustering algorithm is applied in five steps:

a) Initialise the method with the pre-process input image and the number of clusters, in our case $K = 4$;
b) Chose randomly a center for each cluster of pixels;
c) Associate each pixel with its nearest center;
d) Modify the cluster centres to the centroid of their pixels;
e) Return to step c), until convergence (centroid doesn't chance the position).

The resulted image, with four clusters is presented in Fig. 3(a). Using this classification, a process of contours finding starts. After this process we can classify the nose contour as being the contour with largest area, see Fig. 3(b). To determine nose tip, the center of gravity of the contour is found, using central moments. After this stage we have the 2D nose tip coordinate, see Fig. 3(c).

C. 3D Pose Estimation

The nose tip is considered to be input 3D pose of the active vision control system. The 3D position of a pair of 2D coordinates has been determined in two stages: camera calibration and pose estimation. The first process computes the internal parameters of the camera (e.g. focal length and central point deviation) and realise a stereo image rectification. Because of individual camera movement (vergence) analytical focal length in performed for each camera, starting from initial camera calibration. The second process, suppose to transforms each pair of correspondent 2D pixels into 3D points. For determining the new position, the following relations are used [14]:

$$X = x_l \cdot \frac{b}{d}, \tag{3}$$

$$Y = y_l \cdot \frac{b}{d}, \tag{4}$$

$$Z = f \cdot \frac{b}{d}. \tag{5}$$

where, (X, Y, Z) are the 3D coordinates of a point P, represented in homogeneous coordinates, projected on left and right image planes as p_l and p_r:

$$\begin{cases} p_l = (x_l, y_l, 1), \\ p_r = (x_r, y_r, 1), \end{cases} \tag{6}$$

where b represents the distances between the two sensors of the stereo cameras, f is the focal length and $d = x_l - x_r$ correspond to the disparity.

Using the previews equations we solve the 3D pose of a corresponding pair of 2D image coordinate. The new pose, represented by the calculated nose tip, is used as input for camera control system.

III. Active Vision Control Unit

The time-delay introduced by an image processing system is a crucial factor which plays an important role in the robustness and stability of an active vision system. In the current application, the time delay corresponds to the process of image segmentation, object of interest detection and 3D

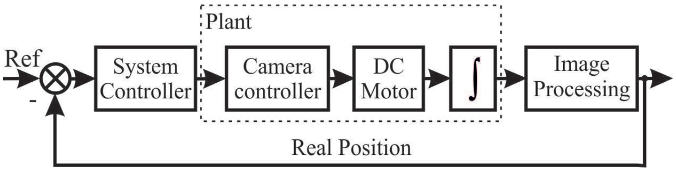


Fig. 5. Active vision model

position estimation. The main goal of the work presented in this paper is to create a system which is able to follow a human observer's facial feature (nose tip), when he is moving on an unspecified path. The whole process is constructed as a human robotic interface, that is, for robotic human behaviour imitation.

The robotic artificial vision system is composed of two Sony digital cameras, able to perform *Pan-Tilt-Zoom* (PTZ) operations. The cameras are placed on a fixed support. The block diagram of all the components involved in proposed architecture is presented in Fig. 4.

The control unit computes the signals for adjusting the current position of both cameras, represented by the PTZ camera. The position is adjusted using all six *Direct Current* (DC) motors inside the stereo camera setup. The feedback information, which contains the current position information, is compared agains a reference signal provided by the *Position References* unit. In this work the reference signal is considered to be a 3D point positioned at the center of the left camera image, along the principal rey of the sensor. These two signals, reference and current position, are used to calculate the next camera position movement. The new movement is performed such that to minimise the error between the desired and real position. The proposed architecture realises a 3D error minimisation approach:

$$e_r = pt_{int}^{ref} - pt_{int}, \tag{7}$$

where, $e_r = (e_x, e_y, e_z)$ represents the 3D error, $pt_{int}^{ref} = (x_{ref}, y_{ref}, z_{ref})$ is the reference 3D position and $pt_{int} = (x, y, z)$ represents the 3D point of interest.

A. Active Vision Model

In order to realize the camera control unit, the overall controller has to be design. The modules involved in the control configuration are shown in Fig. 5. This controller should be able to maintain the system's tracking stability when the facial feature tracking is performed. The control design strategy starts from the open-loop transfer function of the system G_{sys}^{ol}:

$$G_{sys}^{ol} = G_{plant} \cdot e^{-s\tau}, \tag{8}$$

where, G_{plant} represents the transfer function of the plant and $e^{-s\tau}$ represents the time-delay inserted into the system by the image processing component. The plant model is composed of a DC (*Direct Current*) motor with a controller, and an integral module. This integral module is necessary because the output signal of the motors represents the camera velocity and the purpose of the final application is to obtain a position control. For determining the maximum delay values that can be introduced in the overall system, a method similar to the classical approach from [15] is considered, which uses the Nyquist criterion based on the gain K_g and phase margins γ. Based on these two stability indicators, and using the methodology presented in [16], the system's maximum time-delay is obtained with a value of $\tau_{max} = 0.48\,\text{sec}$.

The control system has to adapt the two PTZ cameras which make up the stereo vision robotic platform, thus ending up with a 6-DoF system (e.g. 2x pan, 2x tilt and 2x zoom). The overall controller is design such that to maintain the system's stability when the maximum delay is inserted into the system. In order to fulfill this condition, the Hermite-Biehler theorem for quasi-polynomials is used, with the mathematical description from [8]. The proposed method for computing this type of controller conduct to a stability interval, that can be used in proposed application. Starting from Eq. 8, the overall open-loop transfer function $G_{ol}(s)$ equals:

$$G_{ol}(s) = \frac{2.38}{0.84 \cdot s^2 + s + 1.74} \cdot e^{-s\tau}, \tag{9}$$

where, τ represents the time-delay in seconds. The parameters values of the model have been determined using the Matlab® identification toolbox. The identification process was needed because of the lack of an analytical process model. For stabilising the system with $\tau = \tau_{max}$, a P (*proportional*) controller was designed. The P controller transfer function is: $C(s) = K_c$. The main goal of this method is to analytically determine the region in the K_c parameter space for which the closed-loop system remains stable. This is performed using Theorem 1 from [8], applied to a second order system.

Starting from the plant transfer function described in eq. 9, the parameters involved are: $K = 2.38$, $L = 0.48\,\text{sec}$, $a_1 = 1$ and $a_0 = 1.74$. First, we remark that $a_1^2 < 2a_0$. Thus, is used (ii) from Theorem 1 [8]. After the computation of $\alpha = 0.88$ parameter the od and ev parameters are determined. These are given by $od = 1$ and $ev = 2$. This means that the stable region can be obtained computing the roots z_l, with $l = 1, 2, 3, 4$ of equation:

$$\cot z = \frac{z^2 - 0.47}{0.57z} \tag{10}$$

The z_l roots are used to determine the set of stabilization gains given by:

$$\max_{l=2,4}\{\frac{a_1 z_l}{KLsin(z_l)}\} < K_c < \min_{l=1,3}\{\frac{a_1 z_l}{KLsin(z_l)}\} \tag{11}$$

After calculating the previous active vision system's controller, the values of K_c are within the interval:

$$-16.4754 < K_c < 1.0164. \tag{12}$$

The stability of the obtain interval can be verified using a generalized form of the Hermite-Biehler Theorem applied to quasi-polynomials [8].

IV. Experimental Results

The stereo acquisition system was placed at $1.2\,\mathrm{m}$ above the ground and reacts when a human observer appears inside the camera's field of view. The face, followed by nose detection, starts the tracking process. The experiments were performed in an indoor room, using natural illumination. The subject moved inside the room in a non-predictable way, in frontal and profile poses, covering an area of about $9\,\mathrm{m}^2$.

A camera control was realised for pan values in the interval $[-40°, 60°]$, while for the tilt in the $[-10°, 45°]$ interval. The control is applied individually for the left and the right camera. In the same time, the speed of the camera was modified using three different values: $30°/\,\mathrm{sec}$, $50°/\,\mathrm{sec}$ and $70°/\,\mathrm{sec}$. The images were acquired at a resolution of $640 \times 480\,\mathrm{pixels}$. The active tracking results are presented in Fig. 6, where the Pan-Tilt real value and the position command variable are illustrated at different speeds. The left camera movement is presented in Fig. (6(a),6(c) and 6(e)), while the right camera movement is presented in 6(b), 6(d) and 6(f). The right camera has not so good performance as the left camera, because of the additional process delay generated by the sequential image analysis (the delay is included in the total delay considered in controller design). For each experimental sessions the mean error is calculated as $er[deg] = pos_{est} - pos_{real}$. The obtained error values are summarised in Tab. I.

TABLE I
STATISTICAL POSITION ERRORS FOR PAN AND TILT MOVEMENTS

Item	Speed [°/ sec]	Mean Error [deg]
Pan Left	30	2.5365
	50	4.6072
	70	7.4130
Tilt Left	30	1.1618
	50	3.3475
	70	4.1415
Pan Right	30	14.4943
	50	13.97
	70	15.2349
Tilt Right	30	2.6462
	50	2.082
	70	3.2988

V. Conclusions

The main idea presented in the paper is the realisation of an active vision system that performs head detection and nose tracking for robotic-human interfacing. The tracking process is realised in two stages. Firstly, the overall system is analysed in order to obtain a controller that guarantee the stability of the movement. Secondly, it is designed to determine the 3D position of a facial feature, in our case, a human nose area. In proposed experimental scenarios the camera can tracked the feature with a good accuracy for speeds less than $70°/\,\mathrm{sec}$. For higher values, the camera fails to follow a specific object because of the image blur effect appear in the acquired images. The authors intend to modify the system's controller to a PI (*Proportional-Integrative*), such that to overcome the limitations of a P controller, used for real time tracking.

The methods for 3D facial feature estimation presented in this paper are being integrated into a probabilistic model for gaze tracking [17], a part of ongoing work on the extension of a hierarchical Bayesian framework for multisensory active perception presented in [18]. The framework can be used to feed the active vision system proposed in this paper, providing a powerful solution for applications such as human robot interaction.

Acknowledgment

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] R. Sharma, "Role of Active Vision in Optimizing Visual Feedback for Robot Control," in *The confluence of vision and control*, ser. Lecture Notes in Control and Information Sciences. Springer Berlin, Heidelberg, 1998, vol. 237, pp. 24–40.

[2] M. J. Swain and M. A. Stricker, "Promising Directions in Active Vision," *International Journal of Computer Vision*, vol. 11, pp. 109–126, 1993.

[3] P. Corke, *Visual Control of Robots : High-performance Visual Servoing*. England: Research Studies Press Ltd, 1996.

[4] K. Welke, T. Asfour, and R. Dillmann, "Active Multi-view Object Search on a Humanoid Head," in *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, may 2009, pp. 417 –423.

[5] Y.-Z. Gu, M. Sato, and X. lin Zhang, "An Active Stereo Vision System Based on Neural Pathways of Human Binocular Motor System," *Journal of Bionic Engineering*, vol. 4, no. 4, pp. 185 – 192, 2007.

[6] J. A. Fayman, O. Sudarsky, E. Rivlin, and M. Rudzsky, "Zoom Tracking and its Applications," *Machine Vision and Applications*, vol. 13, pp. 25–37, 2001.

[7] Y. Yao, B. Abidi, and M. Abidi, "3D Target Scale Estimation and Target Feature Separation forSize Preserving Tracking in PTZ Video," *International Journal of Computer Vision*, vol. 82, pp. 244–263, 2009.

[8] G. Silva, D. Aniruddha, and S. P. Bhattacharyya, *PID Controllers for Time-Delay Systems*, 1, Ed. Birkhuser Boston, 2004.

[9] R. Farkh, K. Laabidi, and M. Ksouri, "PI Control for Second Order Delay System with Tuning Parameter Optimization," *International Journal of Electrical and Electronics Engineering*, vol. 3, no. 1, pp. 1–7, 2009.

[10] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[11] R. Vijayanandh and G. Balakrishnan, "Human Face Detection Using Color Spaces and Region Property Measures," in *Control Automation Robotics Vision (ICARCV), 2010 11th International Conference on*, 2010, pp. 1605 –1610.

[12] E. Skodras and N. Fakotakis, "An Unconstrained Method for Lip Detection in Color Images," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 1013 –1016.

[13] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.

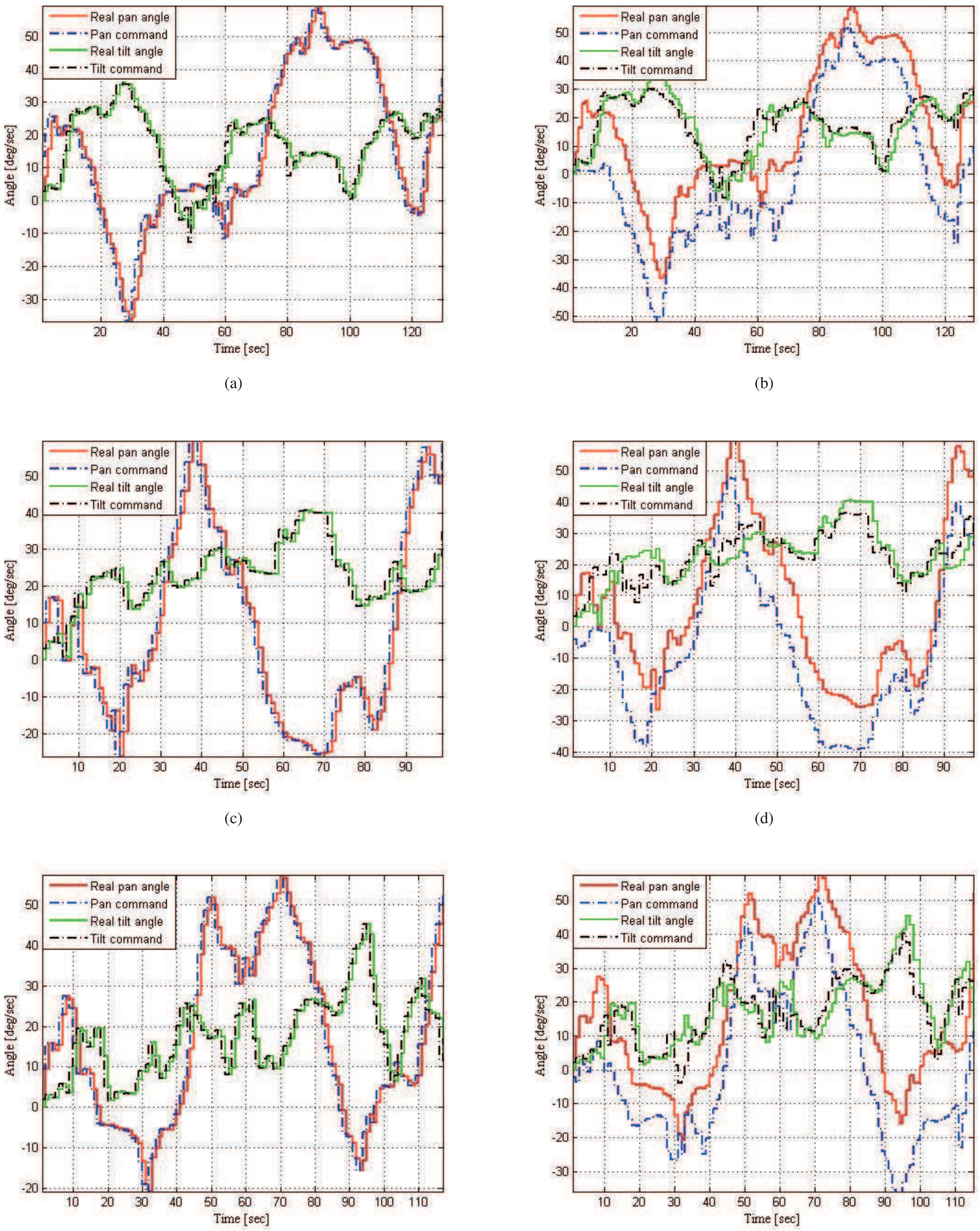
[14] S. M. Grigorescu, G. Macesanu, T. T. Cocias, D. Puiu, and F. Moldoveanu, "Robust Camera Pose and Scene Structure Analysis for Service Robotics," *Robotics and Autonomous Systems*, vol. 59, no. 11, pp. 899 – 909, 2011.

[15] R. Dorf and R. Bishop, *Modern Control Systems*. Prentice Hall PTR, 2010.

[16] G. Macesanu, S. Grigorescu, and V. Comnac, "Time-delay Analysis of a Robotic Stereo Active Vision System," in *2011 15th International Conference on System Theory, Control, and Computing*, 2011, pp. 1 –6.

[17] G. M. Macesanu, J. F. Ferreira, and J. Dias, "A Bayesian Hierarchy for Gaze Following," in *5th International Conference on Cognitive Systems*. TU Vienna, Austria, 2012.

[18] J. F. Ferreira, M. Castelo-Branco, and J. Dias, "A hierarchical Bayesian framework for multimodal active perception," *Adaptive Behavior*, 2012, published online ahead of print.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 6. Active nose tip tracking for different pan and tilt speeds. a,b) 30°/ sec; c,d) 50°/ sec; e,f) 70°/ sec