Intel Serv Robotics (2011) 4:107–118
DOI 10.1007/s11370-010-0077-0

ORIGINAL RESEARCH PAPER

Mechatronic design of the Twente humanoid head

Rob Reilink · Ludo C. Visser · Dannis M. Brouwer · Raffaella Carloni · Stefano Stramigioli

Received: 9 June 2010 / Accepted: 16 August 2010 / Published online: 10 September 2010

Abstract This paper describes the mechatronic design of the Twente humanoid head, which has been realized in the purpose of having a research platform for human-machine interaction. The design features a fast, four degree of freedom neck, with long range of motion, and a vision system with three degrees of freedom, mimicking the eyes. To achieve fast target tracking, two degrees of freedom in the neck are combined in a differential drive, resulting in a low moving mass and the possibility to use powerful actuators. The performance of the neck has been optimized by minimizing backlash in the mechanisms, and using gravity compensation. The vision system is based on a saliency algorithm that uses the camera images to determine where the humanoid head should look at, i.e. the focus of attention computed according to biological studies. The motion control algorithm receives, as input, the output of the vision algorithm and controls the humanoid head to focus on and follow the target point. The control architecture exploits the redundancy of the system to show human-like motions while looking at a target. The head has a translucent plastic cover, onto which an internal LED system projects the mouth and the eyebrows, realizing human-like facial expressions.

Keywords Humanoid head · Mechatronic design · Saliency-based vision system · Vision-based motion control · Human-like behavior

R. Reilink (✉) · L. C. Visser · R. Carloni · S. Stramigioli
Department of Electrical Engineering, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands
e-mail: r.reilink@utwente.nl

L. C. Visser
e-mail: l.c.visser@utwente.nl

R. Carloni
e-mail: r.carloni@utwente.nl

S. Stramigioli
e-mail: s.stramigioli@utwente.nl

D. M. Brouwer
Department of Mechanical Automation and Mechatronics, Faculty of Engineering Technology, University of Twente, Enschede, The Netherlands
e-mail: d.m.brouwer@utwente.nl

D. M. Brouwer
DEMCON Advanced Mechatronics, Oldenzaal, The Netherlands

1 Introduction

Several humanoid heads have been developed over the last years. The robotic heads, presented in the literature, can be classified according to the characteristics of the neck. In particular, it is possible to distinguish between two types: fast and slow moving necks. The former have a short range of motion, consist of two or three degrees of freedom (DOFs) and are used, in general, for object tracking. The latter have a long range of motion, consist of three or more DOFs and are optimized so to perform different expressions and behaviors. Examples of relatively fast two DOF necks are ASIMO by Honda [1], the GuRoo by University of Queensland [2], the humanoid head developed by UC San Diego [3], and Maveric [4], the fast three DOF neck created at the University of Southern California. Examples of necks with a long range of motion are WE-4RII from the University of Waseda [5], QRIO by Sony [6], Cog by MIT [7], the humanoid head by the University of Karlsruhe [8], iSHA by Waseda University [9], and iCub by the Technical University of Madrid [10].

In the Twente humanoid head, the compact mechanical design realizes a fast and long range of motion system with seven DOFs, four for the neck and three for the eyes. The

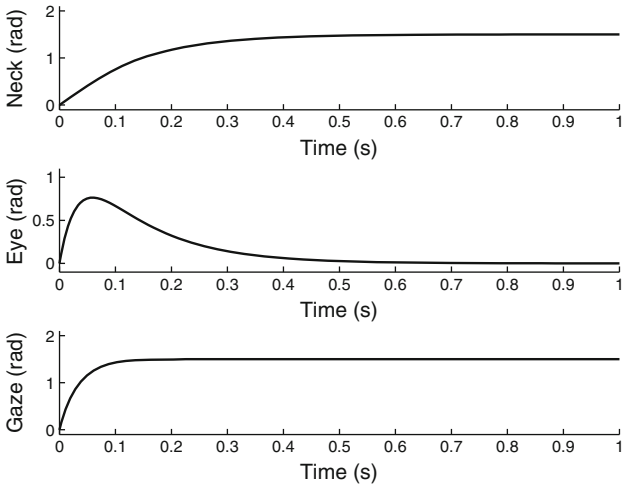

Fig. 1 A simulated saccade of a human. The gaze (*bottom*) is changed quickly due to the fast movement of the eye (*middle*), the neck follows more slowly (*top*), but in the end it is orientated towards the target

humanoid head is capable of tracking objects and of mimicking human expressions and behaviors: thanks to the redundancy of the system, the neck and the eye movements, important in non-verbal communications, are similar to those of human beings. The motion control algorithm receives the input from the vision processing algorithm that processes the camera images, and can steer the humanoid head such that it reproduces the human movements analyzed in biological studies. According to these studies, humans use the eyes to quickly change the gaze, i.e. the angle of the eyes with respect to a fixed reference, to a new target, while the heavy head moves slowly [14]. Figure 1 shows typical position (angle) trajectories for the eye and the head during a rapid change of gaze, i.e. a saccade.

The paper is organized as follows. In Sect. 2, we describe the specifications and the requirements for the mechanical design of the Twente humanoid head and, in Sect. 3, we present the details of the mechanical realization. In Sect. 4, the vision algorithm is described and, in Sect. 5, we present a motion control architecture based on a kinematic model of the system, which controls the humanoid head, realizing human-like behaviors. Section 6 describes the implementation of the expressions and, finally, conclusions are drawn in Sect. 7.

2 Specifications of the humanoid head

To realize a humanoid head that achieves a human-like behavior, the mechanical specifications are directly derived from biological data and, in particular, from the range of motion, maximum velocity and acceleration of the human neck. For the mechanical design of the neck, we adopt a combination of the most challenging specifications on the roll, tilt and pan angles, found in the literature [16–18] and summarized in Table 1.

Table 1 Characteristics of the human neck

	Tilt	Roll	Pan
Range of motion	$-71°$ to $+103°$	$\pm 63.5°$	$\pm 100°$
Max. velocity	$352°/s$	$352°/s$	$352°/s$
Max. acceleration	$3{,}300°/s^2$	$3{,}300°/s^2$	$3{,}300°/s^2$

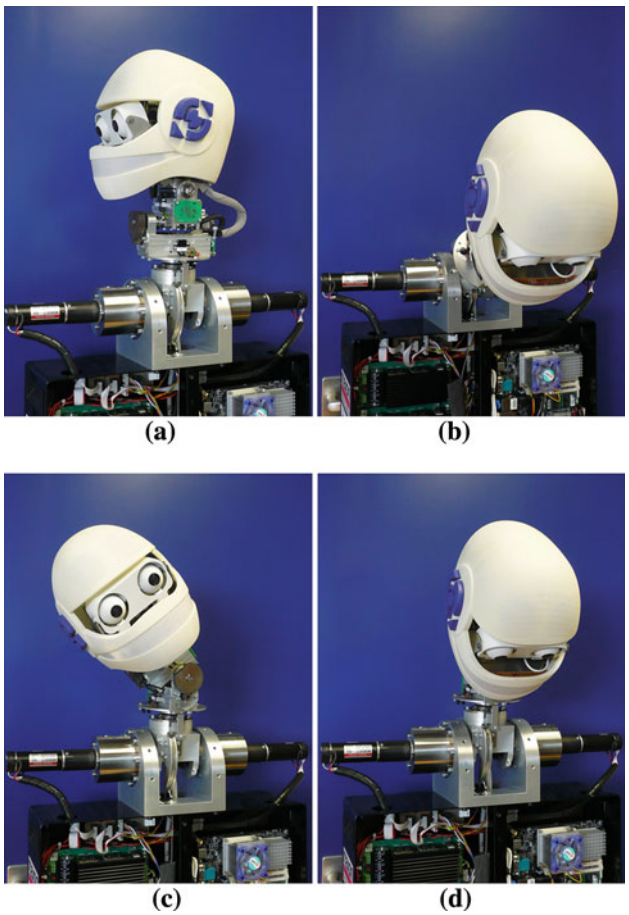

Fig. 2 The four DOFs of the Twente humanoid neck. **a** The pan motion, **b** the lower tilt, **c** the roll, **d** the upper tilt

To meet these requirements, we approximate the mobility of the human neck with four DOFs, i.e. one roll motion, two tilt motions and one pan motion. To create the tilt range specified in Table 1, the tilt of the humanoid neck is split up into two equal contributions over the lower and the upper tilt motions, which are adding redundancy to the system. The final realization is shown in Fig. 2.

Mechanically, a required stiffness of the drive system needs to be estimated. This stiffness combined with the mass of the head leads to a lowest vibration mode frequency, which limits the control bandwidth (open loop cross-over frequency). The control bandwidth, on its turn, determines how fast the head reacts to the changes in the input set points. A fast

response is necessary for tracking purposes. Therefore, based on the specifications, a control bandwidth is estimated for the lower tilt and pan direction [15]. The control bandwidths for the roll and upper tilt movements are less important since these DOFs are not required for object tracking.

According to biological studies, the assumption is that the head should be able to make a rotation h_m of 90° from stand still to stand still in a movement time t_m of 0.36 s. The motion profile is based on a third order set point trajectory, which does not excite the higher frequency dynamics of the system and, therefore, it results in a relatively smooth but fast motion. The movement time t_m is based on the maximum acceleration and velocity, as specified in Table 1. The set point error e is the position difference between the set point and the head position at the end of the specified movement at $t = t_\mathrm{m}$. The tracking of the eyes have a much higher control bandwidth and, therefore, the requirements of the head movement do not need to be too stringent. However, the eye tracking should not be compromised by the performance of the head. Based on biological data, an acceptable error is estimated at 1.8°. The definition of the phase lead factor α_lead is τ_p/τ_z, where τ_z and τ_p are, respectively, the zero and the pole of a PD-controller. A phase lead factor of 0.1 is generally used to obtain a minimized set point error [19]. For a typical PD controlled input force on a system, which acts like a mass, the required minimum control ω_c bandwidth can be estimated as follows [19]:

$$\omega_c = \frac{1}{2\pi}\sqrt[3]{\frac{32h_m}{t_\mathrm{m}^3}\frac{1/\alpha_\mathrm{lead}}{e}} \simeq 11\ \mathrm{Hz}$$

The mechanical layout of the drive system is such that a relatively large inertia (the head) is located at one end of the drive train, whereas the actuator and encoder are at the other end. The drive system in between the head and motor, the shafts and gears, leads to compliance. The encoder is located at the back end of the motor. For such a co-located control system to stay in its stable phase margin, the first mechanical resonance frequency, resulting from the head mass and drive train compliance, should be about three to four times higher than the control bandwidth, as follows from the small gain theorem [19].

The maximum mass for the head is specified at 3.7 kg excluding 0.8 kg of additional peripherals (e.g. audio, external cover), which may be added in the future.

3 Mechanical design

In this section, we describe the details of the mechanical design of the seven DOFs of the Twente humanoid head. The four DOFs of the neck consist of a differential drive for the pan and the lower tilt motion, on the top of which a series structure for the roll and the upper tilt is stacked. The eyes of the humanoid head are realized with two cameras, which can pan independently and can tilt in a combined motion so to realize three DOFs.

3.1 Kinematic structure

The four DOFs of motion of the neck could be created by means of parallel or serial kinematic mechanisms or by hybrid combinations. Based on the specifications for tracking, the head should be able to create a fast pan and lower tilt movement. This requires a relatively low moving mass with respect to the generated force in these directions. In general, parallel mechanisms result in a low moving mass, leading to high vibration mode frequencies and high control bandwidths, enabling fast movements [20]. However, parallel mechanisms have a limited range of motion in comparison to their total system dimensions. This is mainly due to the fact that singularities from input actuator to output motion should be avoided. Therefore, a serial kinematic concept is preferred. This agrees with Beira et al. [10], who concluded that actuating every single DOF separately with rotational motors is, among the tested neck configurations, the best design choice.

However, in our system, due to the stacked design of the serial structure, the weight and the inertia of the upper stages would lead to higher required torques in the motors of the lower parts. Several actuators could be mounted in the base by means of cable or capstan drive systems, thereby reducing the moving mass of the neck. However, cable drive systems have the disadvantage of increased friction, wear and hysteresis and need pre-tensioning. With a capstan drive system, it is difficult to transfer the web across several joints, which are not oriented in one plane. Therefore, in our case, cables or capstan drive systems are not preferred.

A hybrid system, a combination of a parallel and a serial system, was investigated. In such a system, the motions of the two heaviest loaded motors, the pan and the lower tilt, are generated in parallel by a differential drive. The actuators are mounted in the base and, therefore, they do not contribute to the moving mass of the head. The roll and upper tilt are created in series in the neck. The design of the neck becomes relatively compact with a low moving mass. A drawback of this kind of configuration is the increased complexity due to bevel gears in the differential drive, which require alignment with tight tolerances.

3.2 Differential design

The differential drive for the lower tilt and pan motion consists of two sun wheels, a planet wheel and a differential carrier, as shown in Fig. 3. The two sun wheels are externally driven gears. The planet wheel is gear driven by the

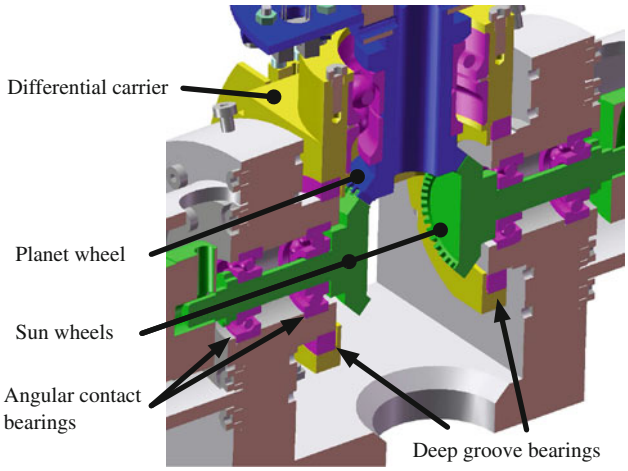

Fig. 3 The differential drive concept consists of two sun wheels, a planet wheel and a differential carrier

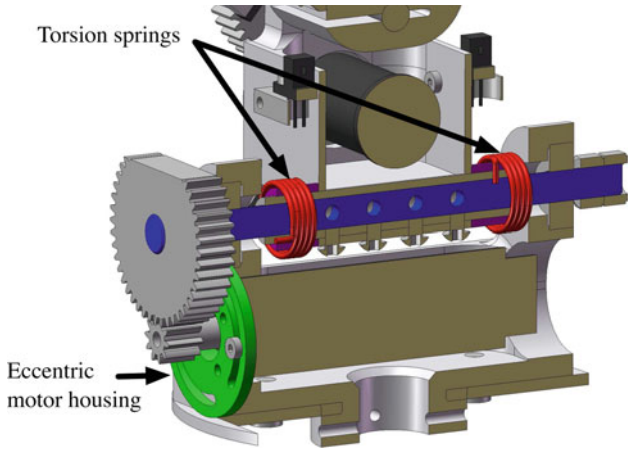

Fig. 4 The roll motion stage showing the adjustable eccentric motor housing and the gravity compensating adjustable torsion springs

two sun wheels. The sun wheels and the planet wheel rotate in pairs of medium pre-loaded angular contact bearings in O-configuration resulting in a large support stiffness. By pre-tensioning, using spring washers and by placing the bearings apart, the angular stiffness has become high, namely 8.8×10^4 Nm/rad. The large support stiffness is required because the inertia of the head on the base would otherwise result in relatively low vibration frequencies. When double row angular contact bearings had been used, the stiffness would have been 27-times lower, resulting in a vibration frequency of around 33 Hz. The paired angular contact bearing differential drive concept comes at the cost of occupying more space. Deep groove ball bearings are used to define the rotation of the differential carrier.

Strong steel (15CrNi6) is used as a material for the gears, which assures a long lifetime and occupies a small space. Each of the sun and the planet wheels are made out off one monolithic piece to assure a good alignment. The shaft of the planet wheel is made hollow to guide the cables through from the upper part down to the controller boards in the torso.

3.3 Mechanical backlash

To create smooth movements and to obtain clear camera images, backlash in the gears has been decreased by implementing an eccentric and adjustable motor housing. The motor is eccentrically fixed in the end-plate of the motor housing. By turning the end-plate, the distance between the motor and the gear can be adjusted, and thus the trade-off between play and friction can be tuned. Due to the 4.5:1 reduction ratio, the play of the head due to the play in the gear boxes is reduced at the same time.

3.4 Gravity compensation

Gravity compensation is applied in the roll and the lower tilt motion by means of pre-loaded elastic elements, so to minimize the required static motor torque and, thus, energy losses. In particular, the elastic energy in the pre-loaded springs balances the potential energy of the humanoid head. Gravity compensation is not applied in the pan and upper tilt motion. The pan motion is not influenced by gravity, and the upper tilt motion is only minimally influenced by gravity. Moreover, the gravity compensation of the upper tilt is depending on the position of the lower tilt and roll and is, therefore, more difficult to compensate.

Figure 4 depicts the realization of the gravity compensation in the roll direction by introducing two pre-loaded torsional springs, which are mounted on adjustable bushes, which enable the pre-load to be adjusted. The gravity compensation reduces the required maximum static motor torque from 0.75 to 0.18 Nm, as shown in Fig. 5. For the lower tilt, two linear springs are used and the cables, which roll over a cam, are attached to the springs. The gravity compensation in the lower tilt direction reduces the required maximum static motor torque from 2.6 to 0.45 Nm. The profile of the cam can be further optimized depending on the mass and the position of future peripherals. With a maximum continuous motor current of 3.4 A, the gravity compensation increases the minimum motor torque available for acceleration from 8.4 to 10.6 Nm. This will allow to add more peripherals to the humanoid head or to use smaller motors in future versions of the system.

3.5 Vibration mode analysis

To analyze the mechanics of the drive system with respect to the lowest vibration mode frequencies of the head, the

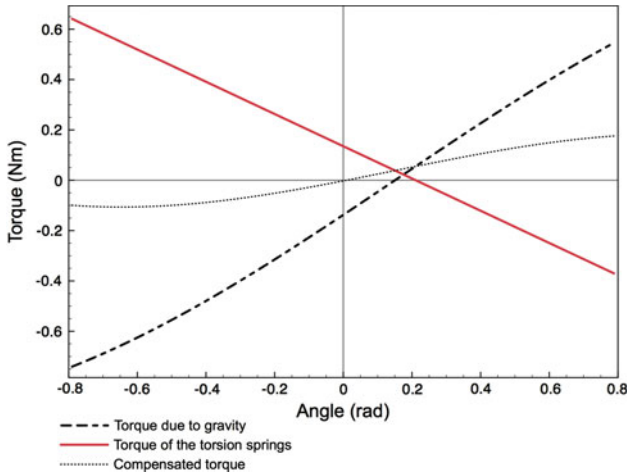

Fig. 5 Gravity compensation for the roll rotation: the required motor torque to statically position the head in the roll direction (*dashed dotted line*) is reduced by the aid of two torsion springs. The combined torque of the torsion springs (*continuous line*) reduces the required motor torque to the compensated torque (*dotted line*)

dominant stiffnesses need to be determined. The cascade of stiffnesses of each of the four DOFs will result in lower vibration mode frequencies than the stiffnesses in the support directions, because the supports use stiff parts and stiff bearings. Therefore, for each DOF, the stiffnesses of the shafts and gears are determined. With these stiffnesses and the inertia of the head, the resonance frequencies can be determined and compared with the specified control bandwidth. The lowest vibration mode turns out to be the roll of the head due to the relatively low torsion stiffness (180 Nm/rad) of the roll shaft, shown in Fig. 4. The vibration mode frequency is 31 Hz, which is about three times the specified control bandwidth derived in Sect. 2.

3.6 Motor and gearbox choice

Based on the maximum acceleration specification given in Table 1, in the worst-case scenario, i.e. with the largest possible inertia, the maximum motor torques are derived so to dimension and, therefore, choose the motors and the gearboxes. For the pan and the lower tilt, Maxon RE ∅30 60W 24V DC motors are used with a continuous nominal torque of 85 m Nm in combination with a 130:1 gearbox reduction ratio. For the roll and upper tilt a Maxon RE max ∅24 11W and an A max ∅26 11W DC motor are used, in combination with gearbox reduction ratios of 18.8:1 and 20:1, respectively.

3.7 Results

Several tests have been performed to validate the specifications of the Twente humanoid head. The mobility in the DOFs has been measured with active safety layers, namely optical switches. All the specified ranges of motion reported in Table 1 are met, except for the roll that is limited to ±39.5°, as presented in [15].

Quite a large safety margin has been taken on all components with respect to robustness. The total weight can possibly be reduced by 30–50% in the future by optimizing the material usage for the required specifications. At critical locations more advanced materials can be incorporated and, consequently, the motors can be downsized.

4 Vision

In this section, we describe the vision algorithm, which is directly derived from biological studies [11]. This is within our aim of realizing a mechanical system that can behave like human beings.

4.1 Human-like behavior: tracking and saccade

Human eye movements are steered by two different kinds of behaviors: top-down and bottom-up attention [11]. Top-down attention is a deliberate eye movement, which is task-driven (e.g. tracking a ball) and requires understanding of the scene. On the other hand, bottom-up attention is the unconscious eye movement initiated by visual cues (e.g. saccades due to movements or to bright colors in the image) and it requires no understanding of the scene.

In the Twente humanoid head, we use the model presented by Itti [11–13]. The model estimates the areas of the image considered interesting by humans, thus realizing bottom-up attention behavior. In particular,

- during tracking, the eyes follow the focus of attention (FOA), which is defined in the image plane and is moving slowly;
- during a saccade, the eyes move from one FOA to the next one at their maximum speed. This happens when the distance between the new and the previous FOA is larger than a certain threshold. During a saccade, the camera input is inhibited since it is severely distorted by motion blur [12].

4.2 Saliency algorithm

The input to the model is the camera color image $I(\mathbf{p}^I)$, where $\mathbf{p}^I$ denotes a point in image coordinates. The model transforms $I(\mathbf{p}^I)$ to a saliency map $S(\mathbf{p}^I)$ that gives a measure of interestingness to each pixel in the input image. On the resulting saliency map, the most interesting point, i.e. the FOA, is determined using a winner-take-all (WTA) network, which selects as the FOA the pixel with the highest saliency

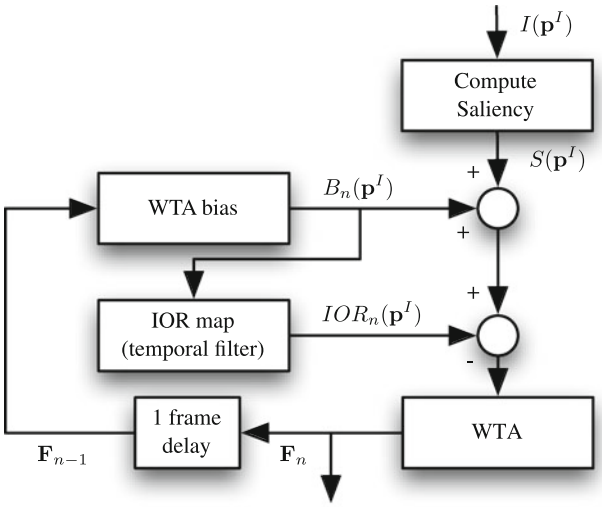

Fig. 6 Computation of the focus of attention: the image I at the point $\mathbf{p}^I$ in image coordinates is transformed into a saliency map S, which is used to compute the focus of attention $\mathbf{F}_n$. The computation of $\mathbf{F}_n$ is influenced by the winner-take-all (WTA) network, the WTA bias B_n and the inhibition of return of frame n, IOR_n

value. Two additional mechanisms influence the selection of the the FOA: the inhibition of return (IOR) map, which prevents looking at the same point all the time, and the WTA bias, which adds hysteresis to prevent undesired jumps in the FOA.

In the remainder of this section, we will describe in more detail how the FOA, denoted by $\mathbf{F}$, is determined. The data flow of the FOA computation is shown in Fig. 6.

4.2.1 IOR map

The IOR map is used to prevent the FOA from staying constant over time, by giving a negative bias to those regions of the saliency map that were attended recently. In particular, the IOR map is a first-order temporal low-pass filter, whose input is the WTA bias $B_n(\mathbf{p}^I)$. The IOR map of frame n, $\text{IOR}_n(\mathbf{p}^I)$, is computed as

$$\text{IOR}_n(\mathbf{p}^I) = \beta\ \text{IOR}_{n-1}(\mathbf{p}^I) + \gamma\ B_n(\mathbf{p}^I) \tag{1}$$

where β and γ are constants and $\text{IOR}_{n-1}(\mathbf{p}^I)$ denotes the IOR map of the previous frame. This first-order difference equation causes the IOR map values to increase around the previous FOA, while it decays everywhere else as $0 < \beta < 1$. As a result, the IOR map has a higher value where the FOA was recently.

4.2.2 WTA bias

The WTA bias is given to a region surrounding the previous FOA in order to create a hysteresis. We will denote the WTA bias as $B_n(\mathbf{p}^I)$, where subscript n denotes the frame number. $B_n(\mathbf{p}^I)$ is computed as

$$B_n(\mathbf{p}^I) = \alpha G_\sigma(\mathbf{p}^I - \mathbf{F}_{n-1})$$

where α is a constant, $\mathbf{F}_{n-1}$ denotes the previous FOA and $G_\sigma(\mathbf{x})$ is a 2D Gaussian function:

$$G_\sigma(\mathbf{x}) = e^{-\frac{||\mathbf{x}||_2^2}{2\sigma^2}} \tag{2}$$

where the constant σ is the size of the Gaussian.

Giving a positive bias to the region surrounding the previous FOA prevents undesired jumping between multiple targets with an almost equal saliency. Since not only the previous FOA is biased but also a region around it, a target can also be tracked if it has moved with respect to the previous frame. The maximum speed at which a target can be tracked is limited by the frame rate and the size of the bias σ.

4.2.3 Calculation of FOA

The saliency map, the IOR map and the WTA bias are summed and fed into the WTA network, given by

$$\mathbf{F}_n = \arg\max_{\mathbf{p}^I} \left(S_n(\mathbf{p}^I) - \text{IOR}_n(\mathbf{p}^I) + B_n(\mathbf{p}^I) \right)$$

It follows that the next FOA is the most salient location, biased negatively for regions that were recently attended and biased positively in order to stay at the current location. The constants α, β and γ can be adjusted to influence the dynamic behavior of the FOA.

4.3 Modeling the camera transformations

To determine where the system should look at, we extend the algorithm developed by Itti [11] so to use a moving camera as the input source. This extension requires a model of how the moving camera perceives the environment. The image that is captured by the camera is a projection of the environment. The properties of this projection are determined by the position and orientation of the camera, the lens, and the camera itself. In the Twente humanoid head, the cameras mainly rotate around their optical center, since the translation is limited. Limiting the description of the camera movement to rotations allows for a significant reduction of the complexity of the problem, at the cost that translations of the camera in the actual setup will cause a deviation from the model. In order to correct for the effects of the changing camera orientation, we model the transformation from points in the environment to pixels on the CCD camera.

Let Ψ_C be the reference frame of either the left or the right camera, and Ψ_0 the fixed world reference frame, as depicted in Fig. 7. The complete camera transformation is due to the camera orientation, the perspective transformation and the lens distortion, i.e.:

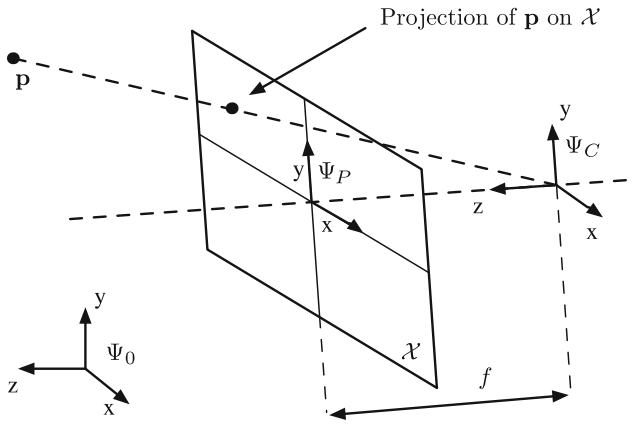

Fig. 7 The target coordinates as perceived by the cameras can be modeled by a projection on the image plane $\mathcal{X}$ using a pinhole camera model. The image plane is at focal depth f on the z-axis of the camera frame. Ψ_C denotes the camera coordinate frame

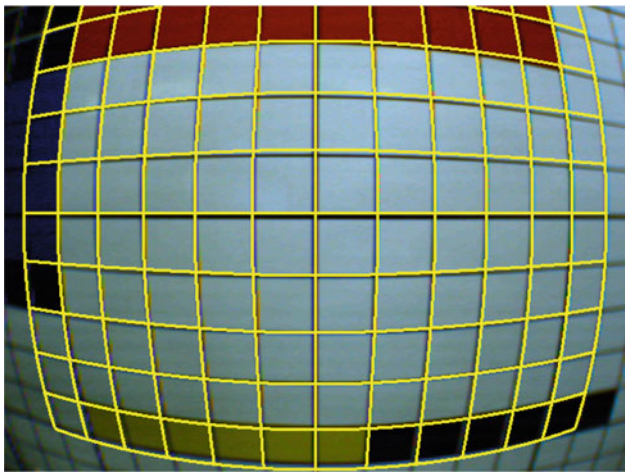

Fig. 8 The fish-eye lens of the camera causes severe distortion. This is compensated using a distortion model. This figure shows a distorted grid as computed by the distortion model, on top of a picture of a calibration pattern, taken by the camera in the head

– The orientation of the camera can be described by a transformation S_r, i.e. the rotational part of homogeneous matrix H_0^C from Ψ_0 to Ψ_C.
– The lens in the camera maps the 3D world onto the 2D image plane by the transformation $S_p : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$, which is described by the perspective transformation of the pinhole camera model [23]. Any point $[p_x\ p_y\ p_z]$ expressed in Ψ_C is projected on the image plane $\mathcal{X}$, resulting in a point $\mathbf{p}^P$ expressed in a reference frame Ψ_P according to

$$\begin{bmatrix} p_x \\ p_y \\ p_z \end{bmatrix}^C \longrightarrow \mathbf{p}^P = \begin{bmatrix} s\frac{p_x}{p_z} \\ s\frac{p_y}{p_z} \end{bmatrix}^P, \tag{3}$$

where the scale factor s depends on both the lens focal distance f and the CCD pixel pitch, and can be determined either using lens and CCD specifications or by calibration measurements. With good approximation, this transformation assumes the optical center of the lens to be equal to the center of rotation of the camera.
– The lens distortion caused by the fish-eye lens is modeled as radial distortion [26]. It is described by a transformation $S_d : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$. Any point $\mathbf{p}^P$ expressed in Ψ_P is transformed in a point $\mathbf{p}^I$ expressed in Ψ_P according to

$$\mathbf{p}^P \longmapsto \mathbf{p}^I = f(|\mathbf{p}^P|)\frac{\mathbf{p}^P}{|\mathbf{p}^P|}$$

where $|\cdot|$ is the Euclidean norm and $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a function that describes the scaling of each vector $\mathbf{p}^P$ depending on its norm. A second order polynomial function is used such that

$$f(|\mathbf{p}^P|) = a|\mathbf{p}^P|^2 + |\mathbf{p}^P|$$

where the parameter a is determined by calibration with a grid pattern. Figure 8 shows an image of a grid pattern taken by the actual camera used in the head. Note that the distorted grid, as computed by the distortion model, is also depicted. This demonstrates that the distortion model is well able to estimate the lens distortion.
The inverse of S_d is $S_d^{-1} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ and gives

$$\mathbf{p}^I \longmapsto f^{-1}(|\mathbf{p}^I|)\frac{\mathbf{p}^I}{|\mathbf{p}^I|} \tag{4}$$

Using a second-order polynomial function for f ensures it is easily invertible, which will be required later on.

The combination of these three transformations describes the mapping from a point in the world, and expressed in Ψ_0, to the corresponding point in the image captured by the camera:

$$\mathbf{p}^I = (S_d \circ S_p \circ S_r)(\mathbf{p}^0) \tag{5}$$

This image transformation model can be used to adapt the saliency algorithm to work with a moving camera, which will be described in the following subsection.

4.4 Saliency algorithm with moving cameras

To use the saliency algorithm in a system with moving cameras, it must be adapted to take the changing camera orientation into account. This means that all data, which are created in one frame and used in another, must be transformed according to this change. Moreover, when a saccade is initiated, the set point for the new camera orientation must be calculated using the described model.

4.4.1 Feed-forward saccade movement

When a saccade is initiated, the target position is known in image coordinates, i.e. $\mathbf{t}^I$. A new camera orientation has to be found such that the target position will map to the center of the image $(0, 0)^I$. Using the inverse lens distortion transformation, the corrected image coordinates $\mathbf{t}^P$ of the target are obtained. These cannot be directly mapped to the rotated world coordinates because the inverted perspective transformation

$$S_p^{-1} : \mathbb{R}^2 \longrightarrow \mathbb{R}^3; \quad \mathbf{t}^P \longmapsto d \begin{bmatrix} \frac{p_x}{s} \\ \frac{p_y}{s} \\ 1 \end{bmatrix} \tag{6}$$

is not uniquely defined, since the scaling factor d is left as unknown. However, this is not a problem since only the direction of $\mathbf{t}$ is needed to compute the required orientation change of the cameras.

During the saccade, the target cannot be detected, since the motion blur will distort the image severely. Therefore, the target is assumed to be stationary during the saccade. The position at which the stationary target would occur in the image plane is simulated using the actual joint positions, and this position is used as an input for the motion control algorithm as a feed-forward action. The accuracy of the feed-forward saccade movement is evaluated in an experiment where a target point in the image is selected. Subsequently, the required movement to get this point in the center of the image is computed using the model, and then executed. Figure 9 shows the target error as a function of the saccade distance, the target error is the Euclidean distance between the center of the image and the actual location of the selected target after the saccade and the saccade distance is measured as the sum of the absolute tilt and the absolute pan angle changes required for the saccade.

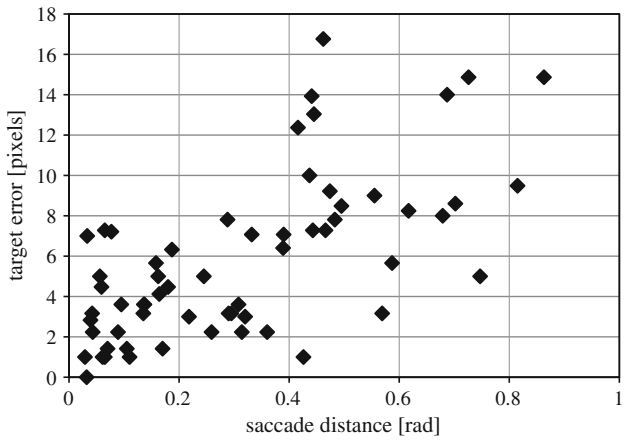

Fig. 9 Feed-forward target error as a function of the saccade movement. The image transformation model is evaluated by performing a saccade to a target point in the image, and then measuring the distance from this point to the center of the image after the saccade

Since during operation of the setup, saccades of up to 0.5 radians are frequently made, errors of over 15 pixels may be expected. The size of the WTA bias (σ in (2)) has to be chosen large enough to deal with these errors.

4.4.2 IOR map

The IOR mechanism causes a certain region to become less interesting when the camera looks at it. This causes the system to keep scanning the environment instead of staring at a single salient location. The region at which the camera has been looking is defined in the world space, while the processing of the IOR map takes place in the image space. Ideally, every point in space would correspond to a single pixel on the IOR map, independent of the camera orientation. With a stationary camera, this mapping is

$$\mathbf{p}^{\text{IOR}} = (S_d \circ S_p)(\mathbf{p}^0)$$

i.e. the same as the mapping from world space to image space when the camera is in its neutral position. To compensate for a moving camera, the transformation from image coordinates to IOR map coordinates is

$$\mathbf{p}^{\text{IOR}} = (S_d \circ S_p \circ S_r^{-1} \circ S_p^{-1} \circ S_d^{-1})(\mathbf{p}^I) \tag{7}$$

Since S_r^{-1} is a linear operator, the unknown scale factor d, introduced by S_p^{-1} in (6), enters linearly in S_p and it cancels out.

To map every pixel of the image space to the IOR map and back would require a considerable amount of processing power. Therefore, for the purpose of the IOR map this transformation is simplified to a shift with respect to the image coordinate space

$$\mathbf{p}^{\text{IOR}} = \mathbf{p}^I + \mathbf{s}$$

where $\mathbf{s}$ is chosen such that the center of the image $\mathbf{c} = (0, 0)^I$ maps according to (7), i.e. $\mathbf{c}^{\text{IOR}} = \mathbf{c}^I + \mathbf{s} = \mathbf{s}$. In particular, note that:

$$\begin{aligned} \mathbf{c}^{\text{IOR}} &= \left(S_d \circ S_p \circ S_r^{-1} \circ S_p^{-1} \circ S_d^{-1}\right) \mathbf{c}^I \\ &\quad \times (\text{no distortion in the origin}) \\ &= \left(S_d \circ S_p \circ S_r^{-1} \circ S_p^{-1}\right) \mathbf{c}^I \\ &= \left(S_d \circ S_p \circ S_r^{-1}\right) \begin{bmatrix} 0 \\ 0 \\ z \end{bmatrix}^C = \mathbf{c}^I + \mathbf{s} = \mathbf{s} \end{aligned}$$

where z cancels out in the perspective transformation S_p. This simplification results in an error in the mapping: a point $\mathbf{p}$ will not map to the same pixel in the IOR map when the camera rotates. The deviation will be larger for larger camera angles. However, the IOR map has a low spatial frequency

and therefore has a limited gradient. This means that for a given error $\mathbf{e}$, the IOR error $|\mathrm{IOR}_n(\mathbf{p}) - \mathrm{IOR}_n(\mathbf{p}+\mathbf{e})|$ is limited.

4.4.3 WTA bias

When determining the maximum salient location in the WTA stage, a bias is applied to the position of the estimated FOA target to create a hysteresis. Like the IOR map, this estimated position is defined in the world space, and a transformation to image coordinates is required. Since only a single point needs to be transformed, the actual transformation and its inverse can be used; the simplification as done with the IOR map is not necessary. However, the simplification might be acceptable since the WTA bias has also a low spatial frequency.

4.4.4 Calculation of FOA

The FOA of the previous frame is known in image coordinates, i.e. $\mathbf{F}^I_{n-1}$. This is transformed to world coordinates using the camera orientation at the time of that frame ($S^{-1}_{r_{n-1}}$), and transformed back to image coordinates of the current frame $\mathbf{F}^I_n$ using the current orientation (S_{r_n}). It follows that:

$$\mathbf{F}^I_n = (S_d \circ S_p \circ S_{r_n} \circ S^{-1}_{r_{n-1}} \circ S^{-1}_p \circ S^{-1}_d)(\mathbf{F}^I_{n-1}).$$

5 Motion control architecture

The motion control algorithm receives target coordinates from the image processing algorithm and calculates the appropriate joint velocities that make the humanoid head look at the target. To be able to design and test the control algorithm, a kinematic and dynamic model of this system has been developed [21]. In particular, the control law is based on a kinematic model of the humanoid head and the pinhole camera model.

5.1 Kinematic model

To design the control architecture of the Twente humanoid head, we need to build its kinematic model. As presented in Sect. 3, the complete system consists of a chain of rigid bodies, connected to each other by actuated joints with a total of seven DOFs: four in the neck and three in the eyes.

The model has been derived using screw theory [22], which provides the mathematical tools to describe kinematic and dynamic relations of interconnected rigid bodies. The generalized velocity, or twist, of a coordinate frame Ψ_j with respect to a coordinate frame Ψ_i, expressed in Ψ_i is given by

$$\mathbf{T}^{i,i}_j = \begin{bmatrix} \boldsymbol{\omega} \\ \mathbf{v} \end{bmatrix}$$

where $\boldsymbol{\omega}$ denotes the relative rotational velocity and $\mathbf{v}$ denotes the relative linear velocity. By fixing a coordinate frame in each rigid body such that it is aligned with the axis of rotation of the joint with which it is connected to the previous body in the chain, the twist of body j with respect to body i, expressed in the frame fixed to body i, is given by

$$\mathbf{T}^{i,i}_j = \hat{\mathbf{T}}^{i,i}_j \dot{\mathbf{q}}_j \tag{8}$$

where $\dot{\mathbf{q}}_j$ is the angular velocity of the joint that connects body j to body i and $\hat{\mathbf{T}}^{i,i}_j$ denotes a unit twist. The relative twists given by (8) can be expressed in the fixed coordinate frame Ψ_0 by applying the adjoint operator, i.e.

$$\mathbf{T}^{0,i}_j = \mathrm{Ad}_{H^0_i} \mathbf{T}^{i,i}_j$$

where H^0_i is a homogeneous matrix describing the change of coordinates from Ψ_i to Ψ_0. The homogeneous matrix depends, in general, on the configuration $\mathbf{q} \in Q$, i.e. the position of the joints defined in the joint configuration space $\mathcal{Q}$. When all relative twists of the chain of rigid bodies are expressed in Ψ_0, they may be added to obtain the relative twist of the last body in the chain with respect to the fixed world.

In Sect. 4, we used the reference frame Ψ_C for either the left or the right camera. Here, we use Ψ_L and Ψ_R to denote the coordinate frames fixed in the focus point of the left and right camera, respectively. It follows that the twists of these frames with respect to the fixed world, expressed in the fixed world, are then given by a simple matrix expression:

$$\begin{aligned} \mathbf{T}^{0,0}_L &= \begin{bmatrix} J_1 & J_2 & J_3 & J_4 & J_5 & J_6 & 0 \end{bmatrix} \dot{\mathbf{q}} \\ \mathbf{T}^{0,0}_R &= \begin{bmatrix} J_1 & J_2 & J_3 & J_4 & J_5 & 0 & J_7 \end{bmatrix} \dot{\mathbf{q}} \end{aligned} \tag{9}$$

in which the lower indices refer to the bodies. Each Jacobian J_i, with $i = 1, \ldots, 7$, is given by:

$$J_i = \hat{\mathbf{T}}^{0,i-1}_i = \mathrm{Ad}_{H^0_{i-1}} \hat{\mathbf{T}}^{i-1,i-1}_i$$

The column vector $\dot{\mathbf{q}}$ in (9) holds all joint velocities, i.e. the angular velocity of the lower tilt, the pan, the roll, the upper tilt, the eye tilt, the pan of the left eye and the pan of the right eye, respectively, as depicted in Fig. 10. Note that, in order to compute the actuator velocities, the differential drive needs to be taken into account.

5.2 Motion control

As described in Sect. 4, the cameras are modeled using the pinhole camera model [23], which is depicted in Fig. 7. The target coordinates $\mathbf{x} = [\mathbf{p}^L \ \ \mathbf{p}^R]^T$, as provided by the vision processing algorithm, are in the image plane $\mathcal{X}$. Obviously, the target coordinates in the image plane change when the joints are actuated. The rate of change of target coordinates

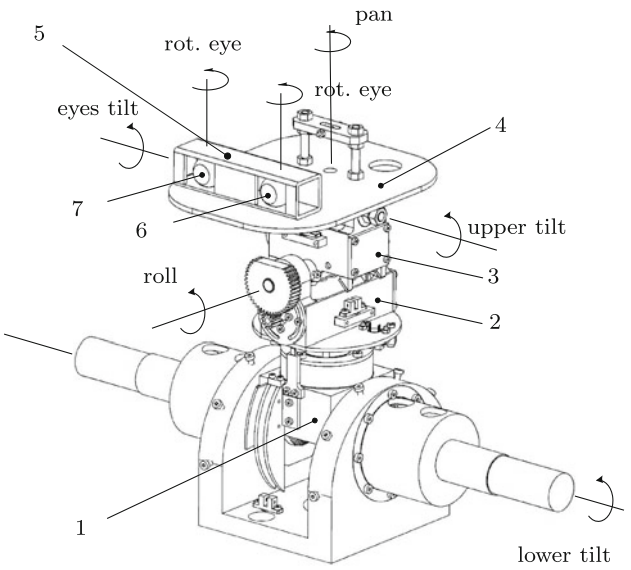

Fig. 10 Degrees of freedom of the Twente humanoid head. The system consists of seven rigid bodies, connected by rotational joints. Each degree of freedom is controlled independently

can be written in the form [22]

$$\dot{\mathbf{x}} = F\,\dot{\mathbf{q}} \tag{10}$$

where F is the map between the tangent space $T_q\mathcal{Q}$ to the joint configuration space $\mathcal{Q}$ and the tangent space $T_x\mathcal{X}$ to the target coordinates space $\mathcal{X}$, i.e. a map between joint velocities and target velocities in the image plane. Based on (10), it is possible to derive the control law that steers the head so that it looks at the target point determined by the vision processing algorithm.

We define coordinates in the image plane such that the center of the image is in the origin of the coordinate frame, i.e. $\mathbf{x}^0 = (\mathbf{0}, \mathbf{0})$. Given target coordinates $\mathbf{x}$, we may define a desired rate of change of the target coordinates, $\dot{\mathbf{x}}_\mathbf{d}$:

$$\dot{\mathbf{x}}_\mathbf{d} = K(\mathbf{x}^0 - \mathbf{x}) = -K\mathbf{x}$$

where $K > 0$ is a proportional gain. Observe that this control law is defined in image plane coordinates. Given $\dot{\mathbf{x}}_\mathbf{d}$, we can calculate the corresponding joints velocities by taking the inverse of (10). Because the system is redundant, the inverse relation is given by (see [24]):

$$\dot{\mathbf{q}}_d = F^\sharp \dot{\mathbf{x}}_d + (I - F^\sharp F)\mathbf{r}$$

where $\dot{\mathbf{q}}_d$ is the desired joint velocity that achieves $\dot{\mathbf{x}}_d$, $F^\sharp$ denotes the weighted generalized pseudo-inverse of F, and $\mathbf{r}$ is an arbitrary vector that is projected onto the null-space of F. The pseudo-inverse is defined as

$$F^\sharp := M^{-1}F^T(FM^{-1}F^T)^{-1}$$

through a metric M [25], which essentially defines the ratio between the elements of the solution and, thus, it can be used

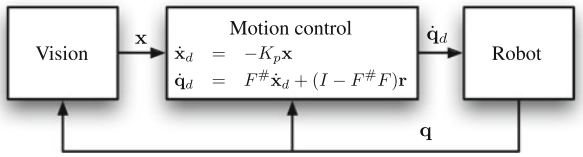

Fig. 11 Overview of the control architecture

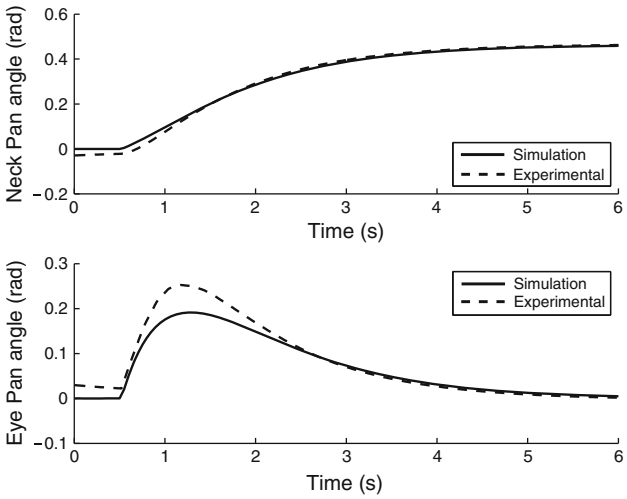

Fig. 12 A saccade performed by the simulation model and the real setup: the humanoid head pan angle (*top*) and eye pan angle (*bottom*) trajectories are very similar to those of humans

to achieve the desired human-like motions. The vector $\mathbf{r}$ is projected onto the null-space of F and, thus, it has no influence on $\dot{\mathbf{x}}$. Therefore, $\mathbf{r}$ can be used to keep the humanoid head close to natural configurations (i.e. the upright position) and to generate specific expressions while looking at a target, e.g. nodding in agreement or shaking in disagreement. Figure 11 shows an overview of the complete control structure.

To test the control algorithm, a dynamic model was built using the 20-sim simulation software [27]. The control algorithm is implemented on the real setup and the simulation experiments have been repeated. The algorithm is able to reproduce human-like motions. Figure 12 shows the behavior of the real setup during a saccade, compared to the human data depicted in Fig. 1. The small differences can be explained by differences in the model and the real setup, due to e.g. friction and mass.

6 Expressions implementation

The design has been completed by adding a translucent plastic cover, which allows the implementation of expressions. A LED system is mounted in the internal part of the cover and the light is projected from inside for the realization of the mouth and the eyebrows. The movements of the mouth, together with eyelids, are coupled with the neck movements

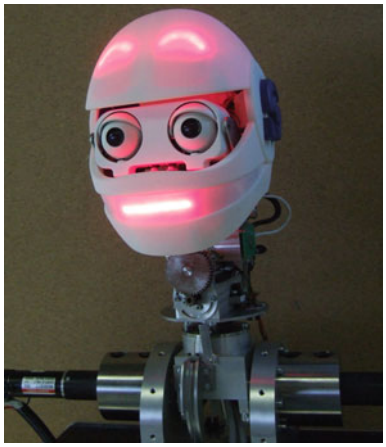

Fig. 13 Implementation of the expressions on the Twente humanoid head

according to human–machine interaction studies. The implementation of the expressions is shown in Fig. 13. A complete overview of the results can be found in [28].

7 Conclusions

In this paper, we presented the mechatronic design of the Twente humanoid head. The design features a differential drive, which helps to combine a fast tracking of the pan and tilt motion with the required long range of motion. At the same time, the differential drive concept results in a combination of relatively low moving mass and a high drive stiffness, which is required to obtain the minimum vibration mode frequency necessary for fast tracking. Other mechanical features include several methods to decrease mechanical backlash and gravity compensation in two DOFs, in order to reduce the motor size, and decrease energy loss.

The implemented vision algorithm uses a saliency map to determine the most interesting point in the image plane, i.e. the FOA. A model of the perception of the world by the cameras was used to allow the algorithm to take the motion of the cameras into account.

A motion control architecture uses the inputs from the vision processing algorithm and implements the motions according to the results of biological data. This has been achieved by actuating the redundant joints using a null-space projection method. Facial expressions are realized using LED light, enhancing human interaction. Experimental results validated the complete mechatronic design.

Future work will focus on human-machine interaction research, which will be performed on the Twente humanoid head. This will include the design of a behaviour-based supervisory control, making the movements of the head more adaptable to different situations.

Acknowledgments The Twente humanoid head is the result of a team work. The authors would like to thank Jan Bennik, Jan Leideman, Herman Soemers and Andre Hilderink for the mechanical design, Stefan Binnemars for the design the exterior shell and LED expressions, Windel Bouwman for implementing and testing the motion control software, Gerben te Riet o/g Scholten for his efforts in building the system and making it work.

References

1. Hirose M, Haikawa Y, Takenaka T, Hirai K (2001) Development of humanoid robot ASIMO. In: Proc IEEE/RSJ Int Conf Intell Robots Syst, workshop 2
2. Wyeth G, Kee D, Wagstaff M, Brewer N, Stirzaker J, Cartwright T, Bebel B (2001) Design of an autonomous humanoid robot. In: Proc Aust Conf Robot Autom, pp 44–49
3. Kim H, Cork G, Burton G, Murphy-Chutorian E, Triesch J (2004) Design of an anthropomorphic robot head for studying autonomous development and learning. In: Proc IEEE Int Conf Robot Autom, vol 4, pp 3506–3511
4. Maveric (2010) http://www.humanoids.usc.edu/HH_summary.html
5. Itoh K, Takanishi A, Miwa H, Matsumoto M, Zecca M, Takanobu H, Roccella S, Carrozza MC, Dario P (2004) Various emotional expressions with emotion expression humanoid robot WE-4RII Robotics and automation. In: Proc IEEE Tech Exhib Based Conf Robot Autom, pp 35–36
6. Geppert L (2004) Qrio, the robot that could. IEEE Spectr 41(5):34–37
7. Brooks A, Breazeal C, Marjanovic M, Scassellati B, Williamson MM (1998) The Cog project: building a humanoid robot. Lect Notes Artif Intell Comput Metaphors Analogy Agents 1562:1
8. Albers A, Brudniok S, Burger W (2003) The mechanics of a humanoid. In: Proc IEEE Humanoids. http://www.sfb588.uni-karlsruhe.de/publikationen/2003/890%20AlbersZ2_2003.pdf
9. Suzuki K, Hashimoto S (2001) Harmonized human-machine environment for humanoid robot. In: Proc IEEE Int Conf Humanoid Robots, pp 43–50
10. Beira R, Lopes M, Praca M, Santos-Victor J, Bernardino A, Metta G, Becchi F, Saltaren R (2006) Design of the robot-cub (iCub) head. Proc IEEE Int Conf Robot Autom, pp 94–100
11. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
12. Itti L, Dhavale N, Pighin F (2003) Realistic avatar eye and head animation using a neurobiological model of visual attention. In: Proc Int Symp Opt Sci Technol, pp 64–78
13. Navalpakkam V, Itti L (2006) An integrated model of top-down and bottom-up attention for optimizing detection speed. Proc IEEE Comput Vis Pattern Recognit, pp 2049–2056
14. Goossens HHLM, van Opstal AJ (1997) Human eye–head coordination in two dimensions under different sensorimotor conditions. J Exp Brain Res 114(3):542–560
15. Brouwer DM, Bennik J, Leideman J, Soemers HMJR, Stramigioli S (2009) Mechatronic design of a fast and long range 4 degrees of freedom humanoid neck. In: Proc IEEE Int Conf Robot Autom, pp 574–579
16. Wagner D, Birt JA, Snyder M, Duncanson JP (1996) Human factors design guide. Internal report of FAA William J. Hughes Technical Center

17. Zangemeister WH, Stark L (1981) Active head rotations and eye–head coordination. Ann N Y Acad Sci 374:540–559
18. Panero J, Zelnik M (1979) Human dimensions and interior space. Watson-Guptill, New York
19. van Dijk J, Jonker JB, Aarts RGKM (2010) Frequency domain approach for mechatronic design of motion systems, submitted to Mechatronics. Elsevier, New York
20. Tsai LW (1999) Robot analysis: the mechanics of serial and parallel manipulators. Wiley, New York
21. Visser LC, Carloni R, Stramigioli S (2009) Vision based motion control for a humanoid head. In: Proc IEEE/RSJ Int Conf Intell Robots Syst, pp 5469–5474
22. Murray R, Li Z, Sastry S (1994) A mathematical introduction to robotic manipulation. CRC Press, Boca Raton
23. Ma Y, Soatto S, Kosecka J, Sastry S (2006) An invitation to 3-D vision. Springer, Berlin
24. Liégeois A (1977) Automatic supervisory control of the configuration and behavior of multibody mechanisms. IEEE Trans Syst Man Cybernet 7(12):868–871
25. Doty K, Melchiorri C, Bonivento C (1993) A theory of generalized inverses applied to robotics. Int J Robot Res 12(1):1–19
26. Shah S, Aggarwal J (1994) A simple calibration procedure for fish-eye (high distortion) lens camera. In: Proc IEEE Int Conf Robot Autom, vol 4, pp 3422–3427
27. Controllab Products BV (2010) http://www.20sim.com
28. Reilink R, Visser LC, Bennik J, Carloni R, Brouwer DM, Stramigioli S (2009) The Twente humanoid head. In: Proc IEEE Int Conf Robot Autom, pp 1593–1594