Auton Robot
DOI 10.1007/s10514-012-9311-2

Real-time visuomotor update of an active binocular head

Michael Sapienza · Miles Hansard · Radu Horaud

Received: 13 July 2011 / Accepted: 8 September 2012

Abstract In order for a binocular head to perform optimal 3D tracking, it should be able to verge its cameras actively, while maintaining geometric calibration. In this work we introduce a calibration update procedure, which allows a robotic head to simultaneously fixate, track, and reconstruct a moving object in real-time. The update method is based on a mapping from motor-based to image-based estimates of the camera orientations, estimated in an offline stage. Following this, a fast online procedure is presented to update the calibration of an active binocular camera pair. The proposed approach is ideal for active vision applications because no image-processing is needed at runtime for the scope of calibrating the system or for maintaining the calibration parameters during camera vergence. We show that this homography-based technique allows an active binocular robot to fixate and track an object, whilst performing 3D reconstruction concurrently in real-time.

Electronic supplementary material The online version of this article (doi:10.1007/s10514-012-9311-2)

M. Sapienza (✉)
Department of Systems and Control Engineering,
University of Malta, Msida MSD 2080, Malta
e-mail: mikesapi@gmail.com

M. Hansard
School of Electronic Engineering and Computer Science,
Queen Mary, University of London, Mile End Road,
London E1 4NS, UK
e-mail: miles.hansard@eecs.qmul.ac.uk

R. Horaud
INRIA Grenoble Rhône-Alpes, 655 Avenue de l'Europe,
38330 Montbonnot, France
e-mail: Radu.Horaud@inrialpes.fr

Published online: 27 September 2012

Keywords Real-time vision · Active binocular vision · Visual tracking · 3D reconstruction

1 Introduction

The estimation of scene-structure from a binocular image pair is an important task in robot vision. Once the epipolar geometry between the two cameras is known, image-features can be matched more easily, and depth information can be recovered (Hartley and Zisserman 2004). For an active binocular robot head, which continuously fixates a moving target, a real-time method for updating the relationship between the cameras is required. Such an online method would allow independent or coupled camera vergence rotations while maintaining geometric calibration. This will enable the robot to perform 3D reconstruction while actively examining different parts of the scene, or while tracking a moving object (Grosso and Tistarelli 1995; Barreto et al. 2010; Hansard and Horaud 2008). Furthermore, this online calibration update must be computationally efficient if it is to be used alongside other complex algorithms such as object tracking and 3D reconstruction.

POPEYE is an active binocular robot (POP Consortium 2008), which reproduces the sensory configuration of the human head (Fig. 1). The orientation of the robot head (pan/tilt) is controlled by motors, as is the direction of the eyes (version/vergence). The robot can therefore direct its attention towards a visual stimulus in its surroundings using combined stereo and tracking techniques (Bellotto and Hu 2010). The advantage of an *active* vision system is that information about the environment can be gathered more efficiently, by appropriately re-orienting the cameras. Hence visual information is combined with motor control, in a feedback loop, which enables the robot to react to a dynamic environment (Bajcsy 1988).

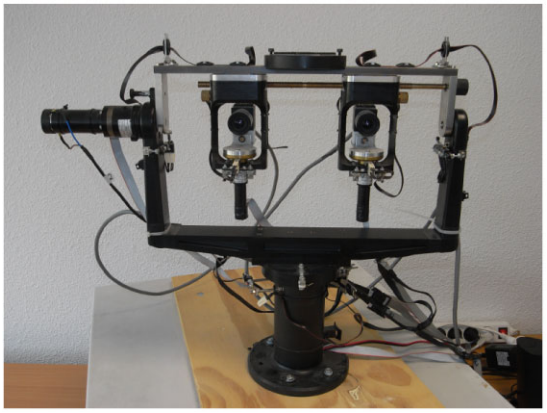


Fig. 1 The POPEYE active binocular robotic head (designed and built in the Institute for Systems & Robotics, University of Coimbra, Portugal). The binocular vergence is controlled by a pair of motors, one below each camera. The pan and tilt angles of the head are controlled by another pair of motors

1.1 Previous work

In order to preserve the calibration of the stereo system after camera vergence, one could continually recalibrate the system. This would, however, be very difficult to perform in real-time. A more efficient method would be to calibrate the system once, and then to update this calibration as the cameras move. This technique was used in Björkman and Eklundh (2002), where the epipolar geometry was updated by extracting information from feature correspondences in the current images. This required an iterative approach to estimate the stereo-head geometric parameters, which is computationally expensive. Furthermore, such an approach ignores kinematic information that can be used to improve the real-time performance of the system. The problem was also tackled in Hart et al. (2002) who developed an epipolar-kinematic model in which motor data is used to compute an updated representation of the epipolar geometry. The kinematics of the system are obtained by rotating the cameras and observing the relationship between the images of two views before and after the rotation. After this calibration procedure, the essential matrix (Hartley and Zisserman 2004) can be updated, provided that the current settings of the camera motors are known. This approach allows computation of the essential matrix in real-time. Related methods can be used to coordinate an active camera with a static camera (Horaud et al. 2006).

The POPEYE robot belongs to a class of active binocular robot heads with very accurate motors (Shih et al. 1998; Aryananda and Weber 2004; Beira et al. 2006; Miwa et al. 2002). Moreover, the design of the eye pan mechanical structures allows each camera position to be precisely adjusted to achieve close agreement between the optical and rotational centres of the cameras. This permits a direct approach to epipolar-update, as described in Sect. 2. The accuracy and negligible backlash of the DC brushed motors ensures the repeatability of the new method, as demonstrated in Sect. 3.

The present work is closely related to certain *autocalibration* procedures (Hartley 1997; Ruf and Horaud 1999; Knight and Reid 2006), in which constrained movements are used to estimate the camera parameters. The aim of the present work is somewhat different; we estimate a direct mapping from motor-settings to image-transformations, without explicitly estimating the calibration parameters.

The paper is organized as follows. The problem definition and main contributions of this work are stated in Sects. 1.2 and 1.3. Sections 2.1, 2.2 and 2.4 describe the estimation, analysis and synthesis of homographies, respectively. The new method is based on the statistical model (3), which allows for uncertainty in the homography estimates, as defined in Sect. 2.3. Section 3 describes the results of experiments using POPEYE. The conclusions of the work are stated in Sect. 4.

1.2 Problem definition

The problem to be solved is that of compensating for the effects of known camera-rotations on the images. This is important for both monocular and binocular tasks, as described below. The *monocular* geometry (which applies equally to the left and right cameras) will be described first.

The standard pinhole-camera model will be used to represent the imaging process. The scene and image points will be identified by homogeneous coordinates $\boldsymbol{X} \simeq (X, Y, Z, 1)^\top$ and $\boldsymbol{x} \simeq (x, y, 1)^\top$ respectively, where '$\simeq$' indicates equality up to a non-zero scale factor. If the pose of the camera, with respect to the scene coordinate-system, is represented by the 3×3 rotation matrix $\boldsymbol{R}$ and 3×1 translation-vector $\boldsymbol{t}$, then $\boldsymbol{x} \simeq \boldsymbol{A}(\boldsymbol{R}|\boldsymbol{t})\boldsymbol{X}$, where $\boldsymbol{A}$ is the (invertible) upper-triangular matrix that contains the intrinsic parameters (Hartley and Zisserman 2004). Now consider two views, $\mathcal{V}$ and $\mathcal{V}_j$, that differ by a rotation of the *single* camera described above. Specifically, suppose that the camera is aligned with the scene-coordinate system, and that image-points $\boldsymbol{x} \in \mathcal{V}$ and $\boldsymbol{x}_j \in \mathcal{V}_j$ are observed before and after a vergence rotation (monocular pan) of angle θ_j. It follows that $\boldsymbol{x} \simeq \boldsymbol{A}(\boldsymbol{I}|\boldsymbol{0})\boldsymbol{X}$ and $\boldsymbol{x}_j \simeq \boldsymbol{A}(\boldsymbol{R}_j|\boldsymbol{0})\boldsymbol{X}$ and therefore

$$\boldsymbol{x}_j \simeq \boldsymbol{H}_j \boldsymbol{x} \quad \text{where } \boldsymbol{H}_j \simeq \boldsymbol{A}\boldsymbol{R}_j\boldsymbol{A}^{-1}. \tag{1}$$

The full-rank 3×3 matrix $\boldsymbol{H}_j$ represents the *homography* (Hartley and Zisserman 2004) that maps view $\mathcal{V}$ to view $\mathcal{V}_j$, as shown in Fig. 2. The matrix $\boldsymbol{H}_j$ will be analysed in Sect. 2.2.

Now, in the *binocular* case, let $\boldsymbol{x}_\ell$ and $\boldsymbol{x}_r$ be corresponding points in the left and right images. The epipolar geometry of the binocular system is expressed by the constraint $\boldsymbol{x}_r^\top \boldsymbol{F}\, \boldsymbol{x}_\ell = 0$ where $\boldsymbol{F}$ is the *fundamental matrix* (Hartley and

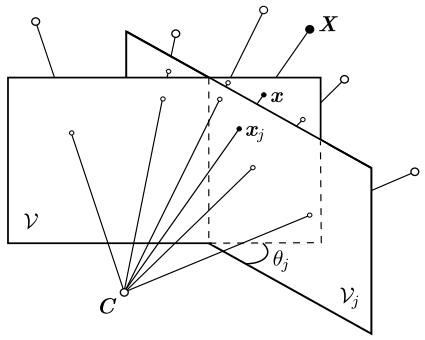


Fig. 2 A camera samples the bundle of rays that pass through the optical centre $\boldsymbol{C}$. Two different views, separated by a rotation of θ_j, are indicated by the planes $\mathcal{V}$ and $\mathcal{V}_j$. Corresponding points $\boldsymbol{x}$ and $\boldsymbol{x}_j$, which are images of $\boldsymbol{X}$, are related by a projective transformation $\boldsymbol{H}$ of the projective space P^2

Zisserman 2004). If the cameras are rotated, by angles $\theta_{\ell j}$, θ_{rj}, then the new coordinates $\boldsymbol{x}_{\ell j}$, $\boldsymbol{x}_{rj}$ will not be compatible with $\boldsymbol{F}$. However, if homographies $\boldsymbol{H}_{\ell j}$ and $\boldsymbol{H}_{rj}$ are available, then the points can be transformed *back* to the coordinate system of $\boldsymbol{F}$, so that $(\boldsymbol{H}_{rj}^{-1}\boldsymbol{x}_{rj})^\top \boldsymbol{F}\,(\boldsymbol{H}_{\ell j}^{-1}\boldsymbol{x}_{\ell j}) = 0$. Alternatively, $\boldsymbol{F}$ itself can be transformed, so that the general epipolar constraint is

$$\boldsymbol{x}_{rj}^\top \boldsymbol{F}_j\, \boldsymbol{x}_{\ell j} = 0 \quad \text{where } \boldsymbol{F}_j = \boldsymbol{H}_{rj}^{-\top} \boldsymbol{F}\, \boldsymbol{H}_{\ell j}^{-1}. \tag{2}$$

The cases $\boldsymbol{H}_{\ell 0} = \boldsymbol{I}$ and $\boldsymbol{H}_{r0} = \boldsymbol{I}$ are defined so that $\boldsymbol{F}_0 = \boldsymbol{F}$ is the original fundamental matrix. The original problem, of compensating for binocular vergence, has now been reduced to that of estimating the homographies $\boldsymbol{H}_{\ell j}$ and $\boldsymbol{H}_{rj}$.

There are two possible ways to estimate these matrices. This is an essentially monocular task, and so the left/right subscripts will now be suppressed. In the **image-based** method, $\boldsymbol{H}_j$ is estimated from $1 \le k \le N$ point-correspondences $\boldsymbol{x}_k \leftrightarrow \boldsymbol{x}_{jk}$. This method is *accurate*, but it is also *slow*, as an additional correspondence problem must be solved after every camera movement. Alternatively, in the **motor-based** method, the current motor-angle θ_j is substituted into the underlying rotation matrix $\boldsymbol{R}_j$ which appears in (1). This method is *fast*, but the results are *inaccurate*, owing to residual misalignment of the optical and rotational centres (see Sect. 2.3).

1.3 Main contributions

This paper shows that the image-based and motor-based approaches, as described above, can be combined. The resulting method has the accuracy of the former approach, as well as the speed of the latter. An outline of the new method is given below.

Firstly, in an offline calibration procedure, the image-based method is used to estimate the homographies $\boldsymbol{H}_j$ that are induced by a set of motor-angles θ_j, where $1 \le j \le M$, as described in Sect. 2.1. Each *estimated* homography determines another angle ϕ_j, which is extracted from the matrix $\boldsymbol{H}_j$, as described in Sect. 2.2. The image-based and motor-based angles are theoretically equal, $\phi_j = \theta_j$. In practice, however, there is a systematic difference between these parameters, owing to the optical and mechanical effects described in Sect. 2.3. These systematic effects can be represented by a statistical model f, and a vector of parameters $\boldsymbol{\eta}$, such that

$$\phi \approx f(\theta, \boldsymbol{\eta}). \tag{3}$$

If the parameters can be learned, then a homography-angle ϕ can be predicted from *any* current motor-angle θ, as described in Sect. 2.3. The corresponding homography $\boldsymbol{H}_\theta$ can then be quickly constructed following a real-time online procedure described in Sect. 2.4. This method is accurate, because the predicted homography $\boldsymbol{H}_\theta$ must be compatible with the original $\boldsymbol{H}_j$ at all fitted values $\theta = \theta_j$, where j indexes the M views that are used to estimate the model (3). This method is also fast, because no feature point extraction and matching is required after each camera rotation. This computational efficiency allows the system to fixate a moving target with calibrated cameras. Finally our proposed method is robust because at *runtime*, it does not depend on the scene texture, establishing feature point correspondences, or a generic scene assumption, unlike purely image-based methods (Sect. 3.5).

In summary, the main contribution of this paper is a novel method for real-time epipolar geometry update, which is applied to an active binocular robotic head. We demonstrate the validity of this **homography-based** method on the POPEYE robot, which can simultaneously fixate an object, track it over time, and perform 3D reconstruction in real-time.

2 Methods

The update-procedure will now be presented, in order of execution. Most importantly, Sect. 2.3 introduces a model and solution for the angle mapping (3).

2.1 Homography estimation

The initial homographies $\boldsymbol{H}_j$ are estimated by the standard *direct linear transformation* algorithm (Hartley and Zisserman 2004). Each homography has eight degrees of freedom (3×3 minus scale). This means that $k = 1, \ldots, N$ correspondences $\boldsymbol{x}_k \leftrightarrow \boldsymbol{x}_{jk}$ are required, where $N \ge 4$. It can be shown (Hartley and Zisserman 2004) that each correspondence defines a pair of homogeneous constraints

$$\begin{pmatrix} \boldsymbol{0}_3^\top & -w_k \boldsymbol{x}_{jk}^\top & y_k \boldsymbol{x}_{jk}^\top \\ w_k \boldsymbol{x}_{jk}^\top & \boldsymbol{0}_3^\top & -x_k \boldsymbol{x}_{jk}^\top \end{pmatrix} \boldsymbol{h}_j = \boldsymbol{0}_2 \tag{4}$$

where $\boldsymbol{x}_k = (x_k, y_k, w_k)^\top$, $\boldsymbol{h}_j = (H_j^{11}, H_j^{12}, \ldots, H_j^{33})^\top$, and $\boldsymbol{0}_L$ is the column-vector of L zeros. The N instances of the 2×9 matrix on the left of (4) are stacked to form a single $2N \times 9$ matrix $\boldsymbol{G_j}$. A solution for $\boldsymbol{H_j}$ can then be obtained from the right singular-vector of $\boldsymbol{G_j}$ that corresponds to the smallest singular value.

2.2 Homography analysis

Recall from (1) that $\boldsymbol{H}_j \simeq \boldsymbol{A}\boldsymbol{R}_j\boldsymbol{A}^{-1}$, where $\boldsymbol{R}_j$ is a rotation matrix. This means that the homography $\boldsymbol{H}_j$ is a *conjugate rotation*, having the same eigenvalues as $\boldsymbol{R}_j$ (Hartley 1997). Furthermore, $\boldsymbol{H}_j$ can be decomposed as

$$\boldsymbol{H}_j \simeq \boldsymbol{U}\,\boldsymbol{D}_j\boldsymbol{U}^{-1} \tag{5}$$

where $\boldsymbol{D}_j$ is a diagonal matrix of eigenvalues, and $\boldsymbol{U}_j$ is a matrix of eigenvectors. The decomposition has the following structure (Hartley 1997; Ruf and Horaud 1999), which involves a complex-conjugate pair of eigenvalues:

$$\boldsymbol{D}_j \simeq \operatorname{diag}\big(\lambda_j, \lambda_j^*, 1\big) \quad \text{where } \lambda_j = \exp(\mathrm{i}\phi_j) \tag{6}$$

and

$$\boldsymbol{U} = (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}). \tag{7}$$

The angle ϕ_j and axis $\boldsymbol{w}$ correspond to the rotation $\boldsymbol{R}_j$. The complex vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ represent the *circular points* in the plane orthogonal to the rotation axis (Hartley and Zisserman 2004). Under pure rotations, the axis of rotation is fixed, and therefore the matrix $\boldsymbol{U}$ remains unchanged. In practice, there are M eigenvector matrices $\boldsymbol{U}_j$, which are only approximately equal, owing to misalignment of the optical and rotational centres. Thus, a synthesis matrix $\bar{\boldsymbol{U}}$ may be chosen as that associated with the largest motor-angle θ_j, or estimated with a suitable averaging procedure in an offline learning stage.[1] Following this, the pre-computed eigenvector matrix $\bar{\boldsymbol{U}}$ will be used in Sect. 2.4 for homography synthesis in a real-time online procedure, since only λ_j needs to be computed at runtime. The angle ϕ can be computed from the motor-rotation θ as described in the following section.

2.3 Motor-based parametrization

This section addresses the chief issue, which is the relationship (3) between the motor parameter θ, and the homography parameter ϕ. The model f in (3) must account for two systematic effects. Firstly, the existence of a 'pinhole' optical centre is only an approximation for real cameras. Secondly, the approximate optical centre may not coincide with the rotational centre of the system. This means that camera rotations will be accompanied by small translations (Hayman and Murray 2003). It is important to note that these effects need only be modelled over a limited range of angles, as determined by the mechanics of the system. Furthermore, this range has a natural origin, which corresponds to the straight-ahead camera position. All motor-counts θ are specified in relation to this origin.

The preceding considerations suggest the zero-offset linear model $\phi \approx \eta\,\theta$, which will be experimentally validated in Sect. 3. This model states that the M available angle-pairs $\{\theta_j, \phi_j\}$ are related by

$$\phi_j = \eta\,\theta_j + \epsilon_j \quad \text{where } 1 \le j \le M. \tag{8}$$

The random errors ϵ_j are due to the optical and mechanical effects described above, as well as to lens distortion and feature mis-localisation. The unknown parameter η can be estimated, given one or more angle-pairs.

The estimation procedure requires the definition of a *distance metric* in which the observed discrepancies $\delta = \phi - \eta\,\theta$ can be minimized. Two possible metrics are $|\delta|_R \in [0, \infty]$, the standard Euclidean metric on R^1, and $|\delta|_S \in [0, 1]$, an angular metric on the circle S^1. These are respectively defined as

$$|\delta|_R = \sqrt{\delta^2} \quad \text{and} \quad |\delta|_S = \tfrac{1}{2}(1 - \cos\delta). \tag{9}$$

The Euclidean metric is not suitable for general angular problems because, for example, $|\delta|_R \neq |\delta + 2\pi|_R$. However, using the Taylor approximation $\cos(\delta) = 1 - \frac{1}{2}\delta^2 + \mathcal{O}(\delta^4)$, it is clear that

$$|\delta|_S = \frac{1}{4}|\delta|_R^2 + \mathcal{O}\big(\delta^4\big). \tag{10}$$

Hence, for small values of δ, it is possible to use $|\delta|_R^2$ rather than $|\delta|_S$. This is important, because the Euclidean metric is much easier to work with. The approximation can be quantified by the relative error

$$E_{RS}(\delta) = \frac{\operatorname{abs}(|\delta|_S - \frac{1}{4}|\delta|_R^2)}{|\delta|_S}. \tag{11}$$

The *maximum* relative error $E_{RS}(\phi_j - \theta_j)$ is 0.12% for the data-set used here, and so the Euclidean metric will be used. It will be demonstrated in Sect. 3.3 that the results of using $|\delta|_S$ are, for practical purposes, identical.

The minimum of the Euclidean error can be found by setting the derivative $\mathrm{d}/\mathrm{d}\eta$ of $\frac{1}{4}\sum_j^N |\phi_j - \eta\,\theta_j|_R^2$ to zero. This leads to the well-known estimate for the slope of a regression-line through the origin:

$$\hat{\eta} = \frac{\sum_j^M \theta_j\phi_j}{\sum_j^M \theta_j^2}. \tag{12}$$

[1] The $\bar{\boldsymbol{U}}$ matrix was obtained by computing columns $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{w}}$ as the leading eigenvectors of $\sum_j \boldsymbol{u}_j\boldsymbol{u}_j^\top$ and $\sum_j \boldsymbol{w}_j\boldsymbol{w}_j^\top$ respectively, with $\bar{\boldsymbol{v}} = \bar{\boldsymbol{u}}^*$.

Algorithm 1 Offline parameter estimation

1. Estimate homographies $\boldsymbol{H}_j$ induced by a set of motor-angles θ_j (Sect. 2.1).
2. Eigendecompose $\boldsymbol{H}_j$ to obtain (i) eigenvalues λ_j, from which the image-based rotation ϕ_j is found, and (ii) a matrix of eigenvectors $\bar{\boldsymbol{U}}$ (Sect. 2.2).
3. Estimate parameters $\hat{\boldsymbol{\eta}}$ relating image and motor-based rotation angles using (12).

Algorithm 2 Online epipolar geometry update

1. Calculate predicted homography angle ϕ using (13).
2. Create new homography $\boldsymbol{H}$ from synthesized eigenvector matrix $\bar{\boldsymbol{U}}$ and eigenvalue matrix $\boldsymbol{D}$ as in (14).
3. Update fundamental matrix $\boldsymbol{F}_j$ using the homographies $\boldsymbol{H}_{\ell j}$ and $\boldsymbol{H}_{rj}$ representing the rotation of both cameras using (2).

If the errors ϵ_j in (8) are normally distributed, $\mathrm{pr}(\epsilon_j) \propto \exp(-\frac{1}{2}\epsilon_j^2)$, then $\hat{\eta}$ in (12) is also the Maximum Likelihood estimate of the underlying parameter η.

It is now straightforward, given *any* motor angle θ, to compute a *predicted* homography angle

$$\phi = \hat{\eta}\,\theta \tag{13}$$

by analogy with the original model (8). This prediction, given the estimate $\hat{\eta}$, does not involve any other computations.

2.4 Homography synthesis

Suppose that a homography angle ϕ has been predicted from a motor-angle θ, using the fitted model (13). It is then possible, by inverting the procedure of Sect. 2.2, to create a new homography $\boldsymbol{H}$, as noted in Hartley and Zisserman (2004) and Knight and Reid (2006). First, by analogy with (6), the eigenvalue matrix is synthesized:

$$\boldsymbol{D} = \mathrm{diag}\big(\lambda, \lambda^*, 1\big) \quad \text{where } \lambda = \exp(\mathrm{i}\,\phi) \tag{14}$$

and λ, λ^* are complex conjugates. The coordinates $\boldsymbol{u}$ and $\boldsymbol{v}$ of the circular points (7) are independent of the rotation angle. It follows that, by analogy with (5), the synthesis matrix of eigenvectors $\bar{\boldsymbol{U}}$ can be combined with $\boldsymbol{D}$ as follows,

$$\boldsymbol{H} = \bar{\boldsymbol{U}}\,\boldsymbol{D}\bar{\boldsymbol{U}}^{-1} \tag{15}$$

which gives the estimated homography.

The above procedure is performed separately for the left and right cameras, so that $\boldsymbol{H}_{\ell j}$ and $\boldsymbol{H}_{rj}$ are obtained from $\theta_{\ell j}$ and θ_{rj}, respectively. It is emphasized that only (13, 14 and 15) need to be computed at run-time, with $\bar{\boldsymbol{U}}$ fixed in advance. The current fundamental matrix $\boldsymbol{F}_j$ is then obtained from $\boldsymbol{H}_{rj}^{-\top}\boldsymbol{F}\boldsymbol{H}_{\ell j}^{-1}$, as in (2). Both the offline and online parameter estimation procedures are detailed in Algorithms 1 and 2 respectively.

3 Experiments and results

In order to evaluate the new approach, a sequence of tests were carried out using the POPEYE robot, as described in Sects. 3.1 and 3.2. The experiments were set in a laboratory environment in which the POPEYE head was placed in a position to best view any surrounding activity. The resulting data is explored in Sect. 3.3, and the assumptions of Sect. 2 are validated. The performance of the new method is evaluated in Sect. 3.4 using an image-based error measure, and compared to a purely image-based method (Sect. 3.5).

3.1 Robot hardware

The robot's vision system consists of two PointGrey colour cameras, which provide images of size 1024×768. The POPEYE head has four rotational degrees of freedom, but only the left/right *vergence* motors are used for the experimental analysis. The rotations are performed by DC brushed motors, which are controlled by discrete (and repeatable) angle-steps. The mechanical system also provides a reference position (straight ahead).

3.2 Calibration procedure

A practical angular range of operation was chosen to be $\pm 20°$ around the reference position, as shown in Fig. 3. This range was split into a discrete set of nine views $\mathcal{V}_j$, each separated by $\Delta\theta = 5°$. The resulting images are representative of the viewable area of the scene. Textured cards were fixed to the facing furniture, containing stable calibration features. Each camera was rotated in turn by $\Delta\theta$ over the whole angular range of operation, at each step taking a snapshot of the viewable area. For each view, ten evenly distributed feature-points were matched and verified by standard methods (Hartley and Zisserman 2004). Three calibration runs were performed for each of the two cameras, giving a total of six data-sets.

The homography matrices mapping each view $\mathcal{V}_j$ to the fronto-parallel position $\mathcal{V}$ were then estimated, using the method of Sect. 2.1. The eigen-decomposition of 2.2 was then used to extract the homography angle ϕ_j corresponding to each motor angle θ_j. The linear model f, that characterizes the relationship between the motor-angle θ and the homography-angle ϕ, was then estimated by the method of Sect. 2.3.

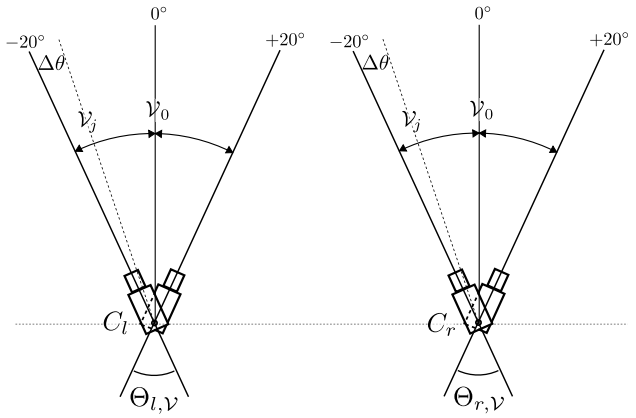


Fig. 3 An illustration of the motor-camera calibration procedure. For each camera, a set of images are extracted which are representative of the viewable area. The relationship between the motor-angle θ and the homography-angle ϕ is modelled by the method described in Sect. 2.3

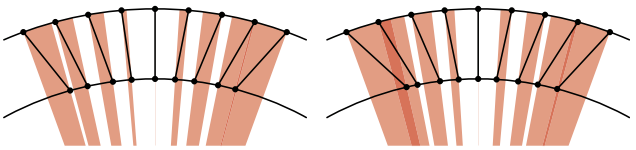


Fig. 4 Visualization of the relationship between ϕ_j (*inner points*) and θ_j (*outer points*) for the left and right cameras. The connecting lines would be purely radial if all $\{\theta_j, \phi_j\}$ pairs were equal. The area of each translucent sector is proportional to the corresponding angular discrepancy $|\phi_j - \theta_j|$

3.3 Exploratory analysis

The first task is to examine the relationship between the measured motor angles θ_j and the estimated homography angles ϕ_j. The mean absolute discrepancy $|\theta_j - \phi_j|$ is 2.92°, with a standard deviation of 1.85°. The maximum absolute discrepancy is 6.87°. The data from the first trial are plotted in Fig. 4. It can be seen that the discrepancy tends to increase for more extreme views, in agreement with the model $\phi \approx \eta\theta$ in (8).

The data can now be used to validate the quadratic approximation $|\delta|_R$ of the angular error $|\delta|_S$ in (9). Note that the latter is a data-dependent sum of $j = 1, \ldots, M$ cosinusoids, $\frac{1}{2}(1 - \cos(\phi_j - \eta\theta_j))$. The small range of the angular errors means that there is no significant difference between the minima of the two metrics, as shown in Fig. 5. The optimum η for the angular error (found by numerical minimization) differs from the regression estimate $\hat{\eta}$ in (12) by 1.9×10^{-8}.

Having validated the error-metric, the adequacy of the simple model $\phi = \eta\,\theta + \epsilon$ in (8) must also be confirmed. This can be done by plotting the *predicted* values $\phi_\theta = \hat{\eta}\,\theta$ in the same form as the *observed* values in Fig. 4. It can be seen from Fig. 6 that, with respect to the predicted values,

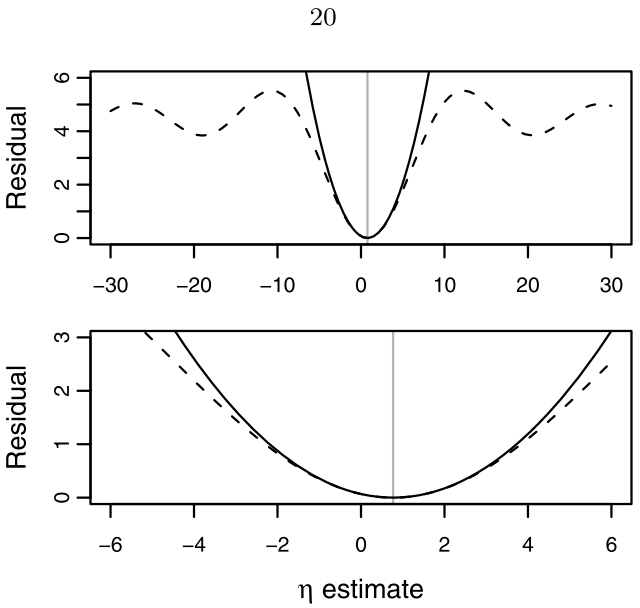


Fig. 5 Comparison of the angular (*dashed*) and Euclidean (*solid*) metrics (9), on the actual data-set. The *upper plot* shows periodic form of the angular error. The global minimum is well-approximated by a parabola, as in (10). The least-squares estimate (12) is the minimum of the parabola, indicated by the *vertical line*. For the present data, the difference between the two metrics is insignificant. The *lower plot* shows a magnified view

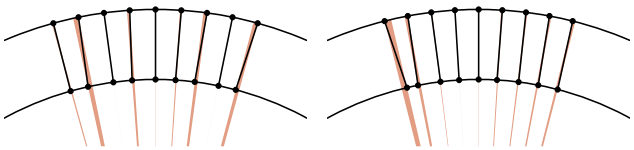


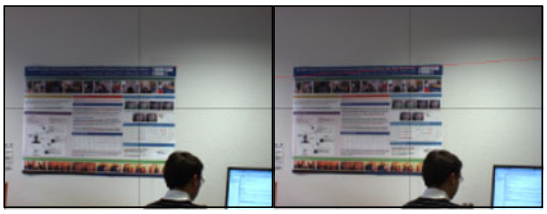
Fig. 6 Visualization of the relationship between ϕ_j (*inner points*) and $\phi = \hat{\eta}\,\theta_j$ (*outer points*) for the left and right cameras. The angular errors (represented by the translucent sectors) are greatly reduced with respect to those in Fig. 4 (note that the inner points, representing ϕ_j, are unchanged)

there is no systematic pattern in the residuals. This indicates that the linear model is adequate.

Lastly, the validity of the synthetic homographies can be checked visually, in the actual images. This is done by synthetically un-rotating the robot's current view, based on the known motor angles and fitted models, as illustrated in Fig. 7(a)–(c).

3.4 Quantitative evaluation

Consider the views $\mathcal{V}$ and $\mathcal{V}_j$, separated by an angle θ_j and related by homography $\boldsymbol{H}_j$. Recall that point $\boldsymbol{x}_k$ in $\mathcal{V}$ is mapped to point $\boldsymbol{x}_{jk}$ in $\mathcal{V}_j$ as $\boldsymbol{x}_{jk} \simeq \boldsymbol{H}_j \boldsymbol{x}_k$ (1). The error of each mapping $\boldsymbol{H}_j$ will now be evaluated over the N point-pairs in each data-set. The points are subject to errors in both images, and so the *symmetric transfer error* (Hartley and



(a) A binocular image pair taken from the fronto-parallel configuration (left camera – right camera).

(b) The left and right cameras are rotated by −10 and +12 degrees respectively, and the homographies representative of these rotations are calculated.

(c) The estimated homographies are then used to un-rotate the current images to the original fronto-parallel views (note: the field of view is truncated).

Fig. 7 Visual validation of the synthesized homographies

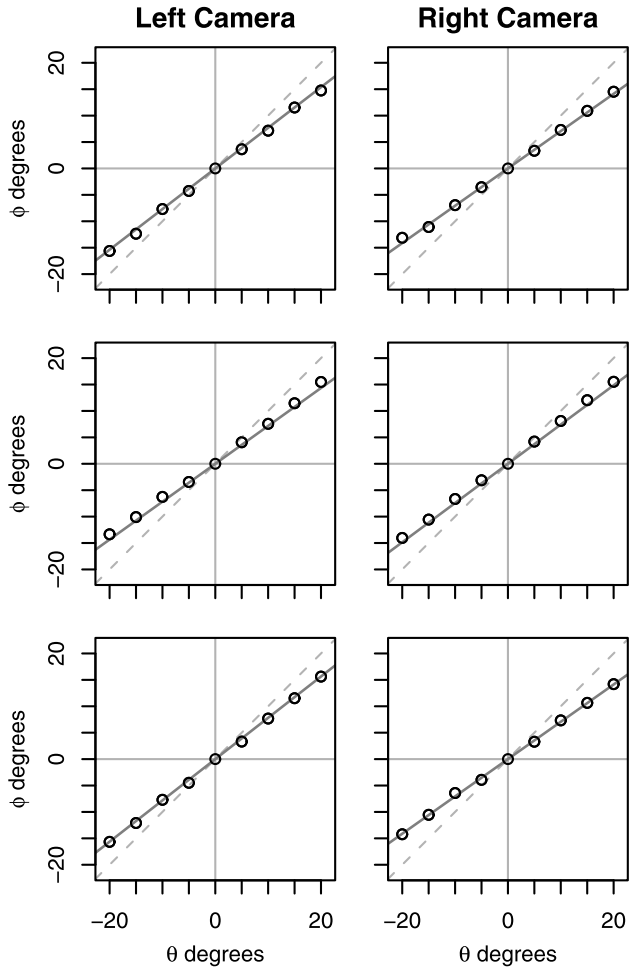


Fig. 8 Homography-angles ϕ_j plotted as a function of the motor-angles θ_j. Three trials are shown (*rows*) for each camera (*columns*). The least-squares regression model is drawn through the data points. It can be seen that the homography/motor relationship is approximately linear, with a slope $\eta < 1$. The *dashed line*, for comparison, has unit slope

Zisserman 2004) is used:

$$E^2(\boldsymbol{H}_j) = \frac{1}{2N}\sum_{k=1}^{N}\left(d^2\left(\boldsymbol{x}_k, \boldsymbol{H}_j^{-1}\boldsymbol{x}_{jk}\right) + d^2(\boldsymbol{x}_{jk}, \boldsymbol{H}_j\boldsymbol{x}_k)\right). \tag{16}$$

Here, for example, $d^2(\boldsymbol{x}_{jk}, \boldsymbol{H}_j\boldsymbol{x}_k)$ is the squared Euclidean image-distance between the measured point $\boldsymbol{x}_{jk}$ in $\mathcal{V}_j$ and the point $\boldsymbol{H}_j\boldsymbol{x}_k$ that is mapped from the reference view $\mathcal{V}$.

This criterion will be used to quantitatively evaluate the accuracy of the synthesized homographies that were obtained from the motor-camera model. Matched points were extracted from the three data-sets, with left and right views spanning ±20° of pan. For each set of images, the motor-camera calibration procedure was carried out as described in Sect. 3.2. The estimated linear relationship between the homography-angles ϕ_j and the motor angles θ_j can be seen in Fig. 8.

The model parameters were then used to synthesize the homographies needed to map the scene points from the fronto-parallel images $\mathcal{V}$ to the rotated views $\mathcal{V}_j$ at angles θ_j. For a comparison between our **motor-image** based and a purely **image-based** method, we additionally estimated the homographies with the standard direct linear transformation (DLT) algorithm (Hartley and Zisserman 2004). The actual and predicted scene points obtained with both methods were then compared in terms of symmetric transfer error, as defined in (16). A discussion of the results obtained with both methods is presented in the following section.

3.5 Discussion

The results obtained after computing the RMS transfer error for our **motor-image** method, over all six data sets,[2] was 2.09 pixels (in the 1024 × 768 images). The maximum and standard deviation of the transfer errors were 6.68 and 1.16 pixels, respectively. The purely **image-based** method

[2]The statistics were computed after excluding the 'perfect' values at $\theta = 0$.

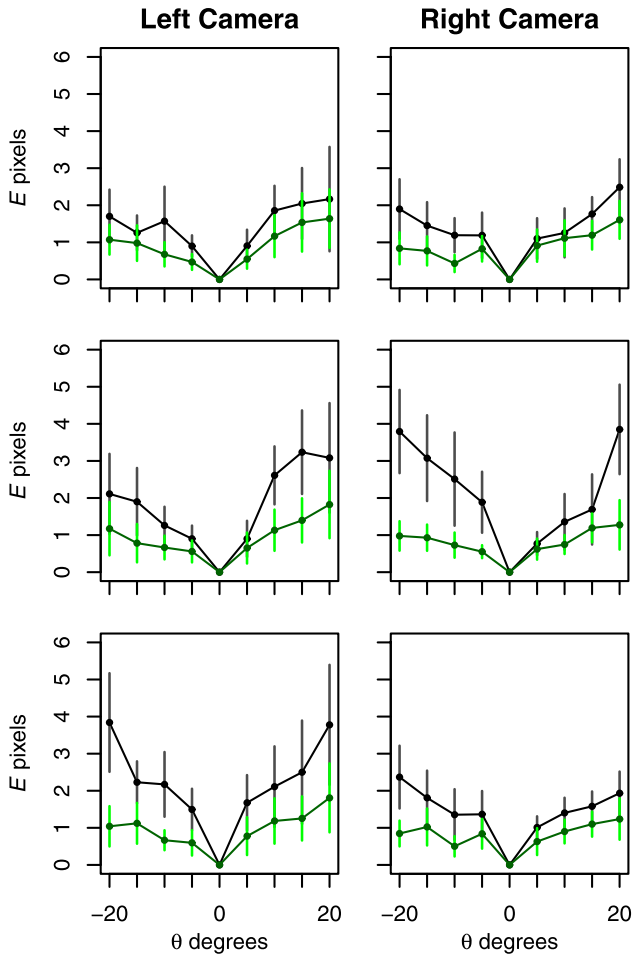


Fig. 9 Symmetric transfer error (16) as a function of the homography angle θ_j for the motor-image method (*black*), and the purely image-based method (*green*) (images of size 1024 × 768). Three trials are shown (*rows*) for each camera (*columns*), as in Fig. 8. The *vertical bars* indicate the spread $\pm\sigma$ of the errors around the corresponding mean. Note that $\boldsymbol{H}_0 = \boldsymbol{I}$, by definition, and so there is no error at $\theta = 0$. The image-based method (*green*) represents the best case scenario where there are sufficient stable points, prefect correspondences, and necessary scene structure to avoid degenerate cases. Our motor-image approach (*black*) shows performance in practice and does not require point correspondences, scene texture or a particular structure

achieved an RMS transfer error of 1.03, and a maximum and standard deviation of 3.28 and 0.58 pixels respectively. The plots in Fig. 9 show a breakdown of the errors across the two cameras, nine angles and three experimental trials.

Note that image-based results are representative of an ideal case, in which sufficient stable points and *perfect* correspondences are available to the algorithm. In practice, an automatic interest point detector and correspondence algorithm will have to cope with low textured scenes, and imperfect matches.

The results from Fig. 9 show that in an ideal case, the purely image-based method will produce lower pixel errors. However, a collection of independently fitted homographies has many more parameters than our synthesis model, so indeed they *should* produce a better fit. Let N_c and N_r denote the number of monocular cameras and monocular camera rotations respectively. Then, the number of estimated parameters for the image-based method is $8 \times N_r \times N_c$, estimated online ($\boldsymbol{H}$ has 8 degrees of freedom). For the motor-image method, the number of parameters is $(1+8) \times N_c$, estimated offline (1 dof from ϕ, and 8 from the $\boldsymbol{U}$ matrix). $\boldsymbol{U}$ has at most eight degrees of freedom, as it is obtained from the 3×3 matrix $\boldsymbol{H}$, which is defined up to an overall scale. More importantly, it is clearly seen that the number of parameters estimated by our method does not depend on the number of camera rotations, N_r. For example in our experiments, the number of estimated parameters totalled $8 \times 8 \times 2 = 128$ for the image based method, and $(1 + 8) \times 2 = 18$ for the motor-camera method.

It is additionally important to point out that for our motor-image based technique, the scene-structure or texture is of no particular importance; each image-point could represent *any* scene-point on the corresponding ray (as indicated in Fig. 2). Hence the motor-image based results generalize immediately to any scene.

In contrast, purely image-based methods *do* depend on the scene texture and structure to compute feature correspondences and to avoid degenerate cases. In the case where $\boldsymbol{F}$ is re-estimated online, at least $n \geq 7$ point correspondences are needed after each camera rotation, subject to the points being in general 3D position. This means that the points must not lie on (or near) a plane or other 'degenerate' surface (Hartley and Zisserman 2004). The image based method to which we compare involves the re-estimation of *homographies* to update $\boldsymbol{F}$. In order to compute each $\boldsymbol{H}$, at least four (no three collinear) point correspondences per camera rotation are required. Since the motor-image method does not depend on computing $\boldsymbol{H}$ from feature point correspondences, it is not susceptible to the degeneracies described above, making it invariant to scene texture and structure. A complete theoretical error analysis of the epipolar geometry (Csurka et al. 1997; Brandt 2008) is beyond the scope of this work.

In these experiments we have shown that our method accounts for *small* misalignments in the optical and rotational camera centres. The development of a similar approach which accounts for the purposeful misalignment of camera rotation and optical centres (Hayman and Murray 2003) is left for future work.

3.6 Target tracking and 3D reconstruction

In order to validate the homography-based method in a real-world setting, the algorithm was used to extend the capabilities of the POPEYE robot by allowing sparse 3D reconstruction during camera vergence movements. Whereas pre-

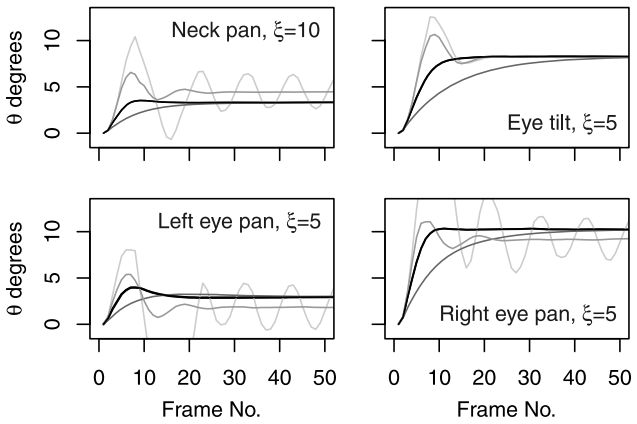


Fig. 10 Response of the active binocular head motors to a step input. A static face was detected and the cameras rotated to fixate on the face. The damping factor was tuned so as to give a fast rise-time with minimal overshoot

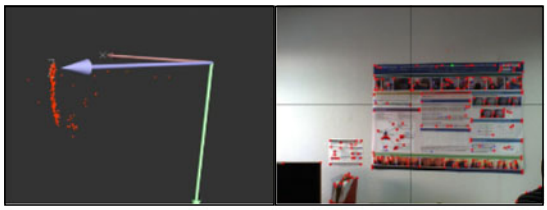
viously the cameras of the robot were static and the viewable area constrained to the initial calibration position of the cameras, it is now possible to allow the cameras to move and verge on a target of interest whilst maintaining its calibrated state. A simple gaze control feedback algorithm was implemented to keep each camera centred on a target of interest as it moved. In this scenario, the robot was placed in a room and its task was to keep track of a human face whilst performing sparse 3D reconstruction of this area of interest.

The possible movements allowed by the binocular robot are horizontal head pan rotation, a vertical tilt movement, and independent pan movements for the stereo cameras. The camera pan movements are considered to be independent and we consider the task of bringing the target position to the centre of the cameras as a monocular tracking task, unlike the tracking in Pagel et al. (1998), in which the camera movements were coupled. This leads to a simple update procedure to bring the object of interest to the centre of each camera:

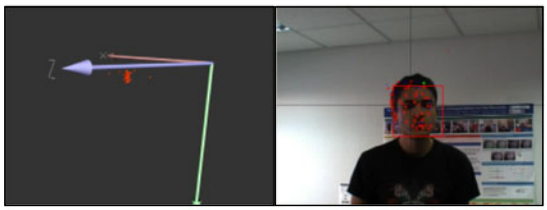
$$\delta\theta \propto \arctan\left(\frac{\delta_x}{f_x}\right) \qquad \delta\psi \propto \arctan\left(\frac{\delta_y}{f_y}\right) \tag{17}$$

where $\delta\theta$ and $\delta\psi$ are the damped pan and tilt eye motor rotations necessary to keep track of a moving target under stability. The damping factor for each motor ξ was found by giving a step-input in the form of a fixed face target to the active head. The motor response was plotted for varying ξ in Fig. 10, and the parameters close to critical damping were chosen.

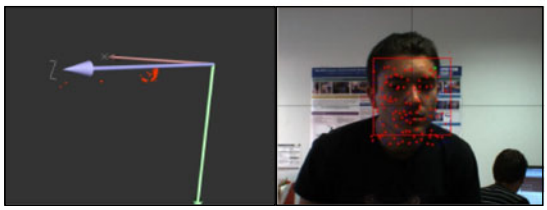
The commonly used Viola-Jones face detection algorithm (Viola and Jones 2004) was used to return the position of a single face in each of the camera images. Feature points within this face area were found using an interest point detector (Harris and Stephens 1988). A more recent method inspired by human visual attention (Frintrop and Jensfelt 2008) can easily be incorporated to determine a region of interest in the image. The detected feature points were matched after limiting the search-space, for each point, to a region around the conjugate epipolar line. The matched points were then triangulated and plotted in Euclidean space as shown in Fig. 11(a)–(c). The complete system runs in real-time (approx. 6 fps), with all image-processing implemented on standard PC hardware. A demonstration of the system in operation can be seen in the attached multimedia file.[3]



(a) A snapshot of the sparse reconstruction (left) of the plane formed by a textured poster (right) fixed to the wall of the lab.



(b) The active head has detected a face and begins to track its movement whilst reconstructing matched points in this region of interest.



(c) As the target moves closer to the camera, the vergence angle between the cameras is updated to continually maintain fixation.

Fig. 11 A 3D viewer (*left*) and the image captured from the left camera of the binocular stereo pair (*right*). In all three camera images, the matched points (*red dots*) are plotted in Euclidean space and seen with the 3D viewer, which is updated in real-time. The motion of the head, for example between (**b**) and (**c**) is well reconstructed despite the camera movement and varying vergence angle

4 Conclusion

The results described in Sect. 3.4 show that the new method of homography-generation is sufficiently accurate for practical use. For example, the method can be used to predict the

[3] Visit: http://www.youtube.com/watch?v=jcIK8AMoejo.

coordinates of image-features, after controlled rotations of the cameras. The method is also useful for binocular vision, in which the fundamental matrix is updated as described in Sect. 1.2. Furthermore, the new method is also fast, because the mapping from motor settings to image transformations is only estimated once. No additional point-correspondences are needed at run-time, and so computationally expensive feature-extraction and matching is avoided. The methods described here have enabled active tracking and image-stabilization to be performed in real-time by the POPEYE robot.

Future work will consider more complicated models for the mapping $\phi \approx f(\theta, \boldsymbol{\eta})$ which relates the motor and homography angles (3). It is expected that the method can be extended, in this way, to systems that have poor alignment between the optical and rotational centres. Finally, the idea of estimating the relationship between control-based and image-based parameters could be applied to other visual processes, such as active zoom.

References

Aryananda, L., & Weber, J. (2004). MERTZ: A quest for a robust and scalable active vision humanoid head robot. In *Proc. IEEE/RAS int. conf. on humanoid robots* (pp. 513–532).

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, *76*(8), 996–1005.

Barreto, J., Perdigoto, L., Caseiro, R., & Araujo, H. (2010). Active stereo tracking of $N \leq 3$ targets using line scan cameras. *IEEE Transactions on Robotics*, *26*(3), 442–457.

Beira, R., Lopes, M., Praga, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F., & Saltaren, R. (2006). Design of the robot-cub (icub) head. In *Proc. IEEE int. conf. on robotics and automation* (pp. 94–100).

Bellotto, N., & Hu, H. (2010). Computationally efficient solutions for tracking people with a mobile robot: an experimental evaluation of Bayesian filters. *Autonomous Robots*, *28*, 425–438.

Björkman, M., & Eklundh, J. (2002). Real-time epipolar geometry estimation of binocular stereo heads. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 425–432.

Brandt, S. S. (2008). On the probabilistic epipolar geometry. *Image and Vision Computing*, *26*(3), 405–414.

Csurka, G., Zeller, C., Zhang, Z., & Faugeras, O. D. (1997). Characterizing the uncertainty of the fundamental matrix. *Computer Vision and Image Understanding*, *68*(1), 18–36.

Frintrop, S., & Jensfelt, P. (2008). Active gaze control for attentional visual SLAM. In *Proc. IEEE int. conf. robotics and automation* (pp. 3690–3697).

Grosso, E., & Tistarelli, M. (1995). Active/dynamic stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(9), 868–879.

Hansard, M., & Horaud, R. (2008). Cyclopean geometry of binocular vision. *Journal of the Optical Society of America A, Online*, *25*, 2357–2369.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Proc. 4th alvey vision conference* (pp. 147–151).

Hart, J., Scassellati, B., & Zucker, S. (2002). Epipolar geometry for humanoid robotic heads. In *Proc. 7th int. workshop on advanced motion control* (pp. 567–572).

Hartley, R. (1997). Self-calibration of stationary cameras. *International Journal of Computer Vision*, *22*(1), 5–23.

Hartley, R., & Zisserman, A. (2004). *Multiple view geometry in computer vision* (2nd edn.). Cambridge: Cambridge University Press.

Hayman, E., & Murray, D. (2003). The effects of translational misalignment when self-calibrating rotating and zooming cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(8), 1015–1020.

Horaud, R., Knossow, D., & Michaelis, M. (2006). Camera cooperation for achieving visual attention. *Machine Vision and Applications*, *16*(6), 330–342.

Knight, J., & Reid, I. (2006). Automated alignment of robotic pan-tilt camera units using vision. *International Journal of Computer Vision*, *68*(3), 219–237.

Miwa, H., Okuchi, T., Takanobu, H., & Takanishi, A. (2002). Development of a new human-like head robot we-4. In *Proc. IEEE/RSJ int. conf. on intelligent robots and systems* (pp. 2443–2448).

Pagel, M., von Mäel, E., & von der Malsburg, C. (1998). Self calibration of the fixation movement of a stereo camera head. *Autonomous Robots*, *5*, 355–367.

POP Consortium (2008). Perception on purpose. European Project FP6-IST-2004-027268. http://perception.inrialpes.fr/POP/.

Ruf, A., & Horaud, R. (1999). Visual servoing of robot manipulators. Part I: projective kinematics. *The International Journal of Robotics Research*, *18*(11), 1101–1118.

Shih, S. W., Hung, Y. P., & Lin, W. S. (1998). Calibration of an active binocular head. *IEEE Transactions on Systems, Man and Cybernetics Part A Systems and Humans*, *28*(4), 426–442.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*, 137–154.

Michael Sapienza has received his B.Eng. in Electrical Engineering in 2009 from the University of Malta, with a thesis on monocular head pose estimation. During his stay at INRIA Grenoble Rhône-Alpes, he has developed algorithmic techniques for the active binocular robot head shown above. Back in Malta, he has carried out research on a monocular mobile robot with the Department of Systems and Control Engineering at the University of Malta, with whom he has obtained an M.Sc. degree. His research interests include computer vision, mobile robotic perception and human interaction.

Miles Hansard is a lecturer in the School of Electronic Engineering and Computer Science, Queen Mary, University of London. He is interested in geometric and statistical aspects of visual perception. His recent work is concerned with 3D reconstruction from range and colour cameras, and with image-processing in biological vision. He studied Experimental Psychology (B.Sc.) and Computer Vision (M.Res., Ph.D.) at University College London.

Radu Horaud received the B.Sc. degree in electrical engineering, the M.Sc. degree in control engineering, and the Ph.D. degree in computer science from the Institut National Polytechnique de Grenoble, Grenoble, France. He holds a position of Director of Research with the Institut National de Recherche en Informatique et Automatique (INRIA), Grenoble Rhone-Alpes, Montbonnot, France, where he is the head of the PERCEPTION team since 2006. His research interests include computer vision, machine learning, multisensory fusion, and robotics. He is an Area Editor of the Elsevier Computer Vision and Image Understanding, a member of the advisory board of the Sage International Journal of Robotics Research, and a member of the editorial board of the Kluwer International Journal of Computer Vision. He was a Program Cochair of the Eighth IEEE International Conference on Computer Vision (ICCV 2001).