On the Kinematics of Robot Heads

Paul M. Sharkey, David W. Murray, and Jason J. Heuring

Abstract—The following presents a standardized approach to the kinematics of a generalized stereo robot head, providing both forward and inverse kinematic solutions as well as a discussion on the head Jacobian. The paper is intended as a comprehensive tutorial and as a standard notation reference for researchers in the field of active vision.

Index Terms—Inverse kinematics, Jacobians, kinematics, robot heads.

I. INTRODUCTION

Over the past seven years, interest in the use of robot heads, so called from their resemblance to the primate vision systems, is evidenced by the development of a wide range of devices with different designs, capabilities, and performance. Details of some examples are given in [1]–[7]. What has been lacking, however, is a standardized approach to the notation used in describing the motions of such devices. This paper presents a straightforward logical approach for describing the motion of a generalized robot head from the user’s (i.e., vision researcher’s) point of view rather than the standard robotics notation (Denavit-Hartenburg, see [8]).

The generalized robot head, described schematically in Fig. 1, has nine mechanical degrees of freedom (DOF). The neck comprises intersecting roll and pan axes (or roll and yaw in nautical/aviation terms). Both cameras are separated by means of a motorized baseline (a prismatic joint), after which each camera is driven in elevation (also called pitch or tilt), vergence, and cyclotorsion. It is assumed that the platform is symmetrical, i.e., that the two neck axes intersect and that the “eyes” are symmetrical with respect to the neck.

The order in which each rotation/translation is achieved will have a considerable effect on the kinematic transformations of the head. The first axis of the generalized robot head is assumed to be the roll axis, followed by the pan axis. The camera axes are configured in the more common Helmholtz configuration (elevate then verge [9]) with cyclotorsion as the final rotation. Examples of the Helmholtz configuration (usually where the elevation axes are constrained to a common elevation) are wide and varied, one such design is Yorick [5]. An extension to the Fick configuration (verge then elevate [9]) is also provided; see for example the KTH head [4]. Kinematics for simple pan/tilt devices and three camera heads (such as Triclops [7]) are easily obtained.

The mechanical DOF are coupled to the optical DOF by six offsets (three in translation and three in rotation) representing the offset between the camera mount and the CCD array of the camera. This final transform is fixed and, though not necessarily directly measurable, may be minimized to within certain tolerances in the design of the head.

The three optical DOF of each camera are provided by motorizing the iris, focus, and zoom for each camera. The first DOF controls the light level by opening and closing the aperture of the camera and has no effect on the kinematics of the head. It does, however, have direct relevance to the image processing and lens control (if not automatic). Both focus and zoom affect the kinematics by shifting the optical origin of the camera (on the CCD array) as the focus and zoom levels are changed. Furthermore the zoom DOF is coupled directly to the kinematics of the platform as the extent of an object of interest in the image will depend on the level of zoom as well as how close (or far) the platform can place the camera with respect to the object.

Mechanical offsets in terms of unknown twist angles and link lengths will remain within the tolerances of the design but may still be incorporated into the kinematic development. However, with good head design these offsets should be minimized, allowing the optical calibration to identify them if required. These offsets are, therefore, not included in this paper.

II. COORDINATE FRAMES

The definition of coordinate frames will facilitate the development in subsequent sections of the kinematics of the head via homogenous transformation matrices between each coordinate frame. The standard method for defining frames in robotics is the Denavit-Hartenburg (D-H) notation [8] whereby frames are selected in pre-defined ways using four parameters for each degree of freedom of the manipulator: link twist $\alpha_i$, link length $a_i$, joint distance $d_i$, and joint rotation $\theta_i$. Using the four parameters a standard homogenous transformation matrix from link $i-1$ to link $i$ is given as

$$T_i = \text{Rot}(\alpha_{i-1}, \alpha_{i-1}) \cdot \text{Trans}(a_{i-1}, d_{i-1}) \cdot \text{Rot}(\alpha_i, \theta_i). \tag{1}$$

For active vision researchers the standard D-H definitions, when applied to a robot head, present coordinate frames that do not coincide with standard image frame definitions. Additionally, D-H requires careful consideration in placing each frames to ensure that...
the standard D–H notation may indeed be applied. Such placements (for example requiring, for translation for prismatic joints), vision researchers assign the z-axis to point in the direction in which the camera is looking—the value of the z coordinate represents the distance to the object in the center of the frame. Assigning positive values along the y-axis to features appearing in the top half of the image, and using the standard right hand rule, positive values in the x-axis denote features to the left of center in the image. Positive rotations about the z-axis result in objects rotating in a clockwise direction about the center of the image. This orientation of the optical frame (x, y, z -> left, up, distance) is notionally extended to preceding coordinate frames of the robot head such that, from a vision point of view, the z-axes of all other frames represents the direction in which that coordinate frame is “looking”. A major advantage with this approach is the reduction of twist angles in the homogeneous transforms, and it is the approach we adopt here.

A. Home Positions

As with any manipulator, a HOME position needs to be defined to extract a reference from which the kinematics may be developed. Mounting the head on a true horizontal surface \( \{ S \} \), a base frame \( \{ B \} \) can then be attached to the base of the head which maintains the same orientation as \( \{ S \} \). HOME may then be taken to be where both cameras are level with each other and are looking at a point at infinity on the horizon. HOME for each DOF can then be defined with respect to \( \{ B \} \), irrespective of \( \{ S \} \).

B. Definition of Coordinate Frames

From a vision researcher’s point of view, therefore, the coordinate frames for the generalized robot head (roll, pan, left/right elevation, left/right vergence, left/right cyclotorsion), represented, respectively, by the rotations \( \{ \theta_{HR}, \theta_{PR}, \theta_{ECL}, \theta_{ECR}, \theta_{VR}, \theta_{VRL}, \theta_{CEL}, \theta_{CER} \} \), may be defined as follows (see Figs. 2 and 3):

\[
\{ W \} \quad \text{World—a universal, fixed frame of reference.}
\]

\[
\{ B \} \quad \text{Base—the frame at the base of the head. Usually fixed, \{ B \} is related to \{ W \} by a simple translation, where the y-axis of this frame, referred to as y_{B}, points “up”. The origin of \{ B \} will be assigned in conjunction with the roll and pan axes. For heads mounted on manipulators, the relationship between \{ B \} and \{ W \} will be a dynamically changing homogenous transformation matrix \( \{ T_H \} \), being the forward kinematics of the arm itself. For heads mounted on AGV’s or other mobile platforms, \{ T_H \} may not be directly available and must be estimated.}
\]

\[
\{ R \} \quad \text{Roll—the axis of rotation is z_{R}, with z_{R} = z_{H}, i.e., \{ R \} coincides with \{ B \} when \theta_{E} = 0 \text{ (HOME). For telepresence systems this axis can be driven directly from the roll of the operator’s head movements, while for active vision systems this axis is principally used to counter any roll of the base \{ B \} (for example, AGV mounted robot heads) to provide a level view for the two cameras.}
\]

\[
\{ P \} \quad \text{Pan—as the roll and pan axes must intersect (symmetry) the axis of rotation, y_{P}, is chosen such that z_{P} represents the “pan direction” and such that y_{P} = y_{R}, \{ P \} coincides with \{ R \} when \theta_{P} = 0. Thus, \{ P \}_{\text{HOME}} = \{ P \}_{\text{HOME}} = \{ B \}. In other words, \{ P \}_{\text{HOME}} is defined where the rotation of the pan axis occurs in a plane parallel to y_{P} = 0.}
\]

\[
\{ C_{S[L/R]} \} \quad \text{Camera Separation (left/right)—This axis provides a linear mechanism for changing the baseline, defined as 2\delta, between the two cameras (cf. inter-ocular separation). An obvious choice for the frame is such that \{ C_{S[L/R]} \} and the elevation frames coincide when the \theta_{EL} = \theta_{ER} = 0. Translations occur in the x_{P} = x_{EL} = x_{ER} = x_{CSL} = x_{CSR}, direction, by \pm \delta, with \{ C_{S[L/R]} \} located at the end of that translation. \{ C_{S[L/R]} \}_{\text{HOME}} is related to \{ P \} by \pm \delta_{\text{nominal}} in x_{P}, and two other constant offsets: y_{CSL} in y_{P}, and \pm \delta_{\text{nominal}} in z_{P}.}
\]

\[
\{ E_{L/R/H/C} \} \quad \text{Independently Elevated (left/right/center)—related to \{ CS \} via the rotation of the elevation axes, by angle \pm \theta_{E} about x_{CS}, such that the z_{E} axes is coincident with the z_{V} axes of the vergence DOF\(^1\). Note that the rotation is in the negative direction to ensure that positive values of \theta_{E} indicate “up.” When in the HOME position, z_{E} is parallel to z_{P}. For robot heads}
\]

\(^1\)A centrally mounted (Cyclopean) camera may also be available, hence \{ E_{C} \}, see [7] for example.

\(^2\)For clarity we use \theta_{E} to refer to either \theta_{EL} or \theta_{ER} depending on context and unless otherwise specified. We will apply a similar notation for the “eye” frames.
with common elevation configuration the elevation axes of each camera are mechanically coupled (or singly actuated) so that both cameras pitch together, hence a common elevation frame can be defined \( \{ E \} \), usually at the center of the baseline.

The above frames allow the definition of two supplementary frames.

\( \{ G \} \) - Gaze—also referred to as the Cyclopean frame, the z-axis of this frame, \( z_G \), provides a gaze direction along which both cameras might verge, such that the object of interest is equidistant from each camera, with the origin being equidistant between and colinear with the two vergence frame origins. Ideally this point should be equidistant between the two optical frame origins (see \( \{ O_{L/R} \} \) below) but this is a less flexible definition as it is necessarily dependent on the camera/lens parameters, zoom and focus, as well as build quality and the precision of camera mounting. The \( \{ G \} \) frame is related to \( \{ E \} \) via a translation \((0, y_{\text{v.c}}, z_{\text{v.c}})\) in \((x_E, y_E, z_E)\)—see \( \{ V \} \) below—and will coincide with \( \{ E \} \) only if the elevation and vergence axes intersect.

\( \{ U \} \) - Unelevated Gaze—provides a direction that always points to the horizon. \( \{ U \} \) coincides with \( \{ G \} \) when in the HOME position (\( \theta_E = 0 \)). As the head is symmetrical about the neck axes of rotation, the direction of \( z_U \) is parallel to \( z_P \), and \( \{ U \} \) frame is related to \( \{ P \} \) via constant offsets \((0, y_{\text{v.c}}, z_{\text{v.c}} + y_{\text{v.c}})\) in \((x_P, y_P, z_P)\).

\( \{ V_{L/R} \} \) - Vergence (left/right)—related to \( \{ E \} \) via a translation \((0, y_{\text{v.c}}, z_{\text{v.c}})\) in \((0, y_{\text{v.c}}, z_{\text{v.c}})\), and followed by a rotation of the left/right vergence axis through \( \theta_V \). The HOME direction, \( \theta_V = 0 \), is when \( z_V \) is parallel to \( z_G \) and \( z_E \).

\( \{ C_{L/R} \} \) - Cycloptorsion (left/right)—rotates the optical frame about the \( z_C \equiv z_V \) axes, by \( \theta_C \), to align both images. Cycloptorsion is only relevant for binocular systems, and then only to aid direct matching among the images, see \([10]\) and \([11]\). Certain types of image processing for achieving stereo matching, image processing involving the raw image rather than features, benefit if the two images are aligned as well as possible before matching. Mechanical cycloptorsion can be used to achieve alignment (KTH has this facility), but rotating sparse image features (not image pixels) in software is simpler and computationally inexpensive. HOME aligns \( \{ C_{L/R} \} \) with \( \{ V_{L/R} \} \).

\( \{ O_{L/R} \} \) - Optical (left/right)—defines the optical center and orientation of the camera—all image processing results are returned with respect to these frames. \( \{ O_{L/R} \} \) are related to \( \{ C_{L/R} \} \) (or \( \{ V_{L/R} \} \) for systems without cycloptorsion) by six offsets, three in position \((u, v, w)\) and three in orientation \((\alpha, \beta, \gamma)\): all offsets are dependent on the precision of the camera and lens, and the zoom and focus settings. These offsets should be quite small (measurable to within a small tolerance) when compared to the baseline. HOME can be defined when focused at infinity and when set to a nominal field of view.

\( \{ T \} \) - Target object—describes the position of the target object of interest with reference to the world coordinate frame \( \{ W \} \). For robot heads without cycloptorsion there is little ability to change the orientation of the object through movement of the cameras so we concentrate on the position of the target as defined by a center of “gravity” of the object as seen in the image of each camera, i.e., at a distance \( d_{TL} \) or \( d_{TR} \) along the \( z \)-axes of the cameras, \( z_{GL} \) or \( z_{OR} \), respectively. For simplicity we further assume that both centers represents a single point in \( \{ W \} \).

Fig. 4 shows a general attitude of a Common Elevation robot head, assumed to be pointing at the same target position in space with no roll or cycloptorsion.

### C. Optical Degrees of Freedom and Calibration

In systems employing fixed focal length lenses and constant focus, the offsets between the optical axes and the mechanical DOF will be fixed and may be identified through calibration of the mount \([12]\). When active focus and zoom are employed these offsets change dynamically. The changes in these offsets, however small, tend to be highly nonlinear and can only be determined for a particular camera/lens combination by undertaking extensive calibration at different focus/zoom settings, and then employing look-up tables and interpolation techniques to incorporate their effect into the kinematics.

For systems employing zoom lenses the target description may be augmented with a measure of the extent of the object, again assumed to be independent of orientation of the object. Then a combination of zoom level and camera position will result in a desired cropping of the image around the target. The performance of motorized zoom lenses, however, tends to be poor with respect to most current head designs. Thus most systems should be able to center any object of interest along the gaze direction long before zoom can crop the image. With such separation of dynamics there will be negligible interaction between the zoom and camera positioning from a kinematic point of view.

### III. FORWARD KINEMATICS

Given the frame descriptions of the previous section homogenous transformations may be derived between two successive frames on the head. The platform kinematics developed below are for either camera parameterized by \( s \), where \( s = +1 \) for the left camera, and \( s = -1 \) the right camera:

\[
^H T_R = \text{Rot}(z_R, \theta_R);
\]

\[
^H T_P = \text{Rot}(y_P, \theta_P);
\]

\[
^H T_{US} = \text{Trans}(x_P, s\delta) \, \text{Trans}(y_P, y_{CSa}) \, \text{Trans}(z_P, z_{CSa});
\]
\[ C^ST_E = \text{Rot}(z_{CS}, -\theta_E); \]
\[ E^T_V = \text{Trans}(y_{CV}, y_{V+}) \text{Trans}(z_{E}, z_{V+}) \text{Rot}(y_{V}, \theta_V); \]
\[ v^T_C = \text{Rot}(z_C, \theta_C); \]
\[ c^T_D = \text{Rot}(\alpha, \beta, \gamma); \]
\[ o^T_T = \text{Trans}(z_0, d_T). \]

The intermediate gaze frames are given by
\[ v^T_U = \text{Trans}(y_{CV}, y_{CV+}) \text{Trans}(z_{E}, z_{CV} + z_{V+}) \]
\[ E^T_{CG} = \text{Trans}(x_{E}, -s) \text{Trans}(y_{E}, y_{V+}) \text{Trans}(z_{E}, z_{V+}). \]

Thus the homogeneous transform from the world frame \{W\} to the target frame \{T\} is given by
\[ w^T_T = w^T_B h^T_R h^T_P h^T_{CS} E^T_{EG} E^T_V E^T_C T_{O^T_T} \]
where, if we assume that \( C^T_{O^T_T} \equiv I_4 \), the 4 x 4 identity matrix, the kinematics calculation from frames \{B\} to \{T\} involves five rotations and three vector translations. The forward kinematics for the open kinematic chain from the base to the target position is found as shown in (5) at the bottom of the page.

As expected, cyclostern does not feature in this equation, reflecting the fact that the target is assumed to be a point. Now, if both cameras are directed to look at the same object (point) in \{W\}, then both left and right open kinematic chains will be closed around the target and the first constraint on the kinematics is found as
\[ H^D = h^D_{left} = h^D_{right} = w^T_{left} \times [0 \ 0 \ 0 \ 1]^T \]
\[ = w^T_{right} \times [0 \ 0 \ 0 \ 1]^T \]
(6)
i.e., substituting \( \theta_{EL}, \theta_{VL}, d_{TL}, d_T \) and \( s = 1 \) on the left-hand side, and \( \theta_{ER}, \theta_{VR}, d_{TR}, d_{TR} \) and \( s = -1 \) on the right.

A. Interchanging Roll and Pan Axes

As the pan and roll frames coincide in the HOME positions, and, with the assumption of head symmetry, they may be easily interchanged, with \( H^T_{CS} = h^T_{RP} h^T_R h^T_P h^T_{CS} \), as above, or \( H^T_{CS} = h^T_P h^T_R h^T_{RP} h^T_{CS} \), where \( h^T_P = \text{Rot}(y_{CV}, \theta_P) \); \( h^T_R = \text{Rot}(z_C, \theta_R) \); and \( H^T_{CS} = \text{Trans}(x_{R}, -s) \text{Trans}(y_{R}, y_{CS+}) \text{Trans}(z_{R}, z_{CS}). \)

\[ B^T = \text{Rot}(x_{R}, \theta_R, y_{CS+}) = \text{Trans}(x_{R}, -s) \text{Trans}(y_{R}, y_{CS+}) \text{Trans}(z_{R}, z_{CS}). \]

B. Fick Configuration

An extension of the kinematic model (2)-(6) to the Fick or independent gun-turret model is provided with the addition of two independent elevation frames \{E_{FICK}\} between the vergence and cyclostern axes of each camera, where \( x_{FICK} = x_{V} \), the origins of \{V\}, \{E_{FICK}\} and \{C\} coincide, and \( z_{FICK} = z_{CV} \), when \{E_{FICK}\} is in the HOME position. The new homogenous transforms are
\[ v^T_C = v^T_{EF} E^T_C \]
\[ v^T_{EF} = \text{Rot}(x_{EF}, \theta_{EF}) \]
(7)

This configuration constrains the \( z_{VL} \) and \( z_{VR} \) to a single plane, whereby simple triangulation confirms the following constraints on \( \theta_{VL}, \theta_{VR}, d_{VL}, \) and \( d_{TR} \):
\[ d_{TR} C_{VR} \equiv d_{TL} C_{VL}; \]
\[ 2\delta = d_{TR} S_{VR} - d_{TL} S_{VL} \]
\[ d_{TR} = 2\delta C_{VL}/\sin(\theta_{VR} - \theta_{VL}) = 2\delta C_{VL}/S_{VR - VL} \]
\[ d_{TL} = 2\delta C_{VR}/S_{VR - VL}. \]

(8)

Substituting into the general kinematics (5) yields (9) as shown at the bottom of the page.

Substituting the first such head further that the target is along the gaze direction, \( z_C \); i.e., the cameras verge along the gaze direction only \( \theta_{VR} = -\theta_{VL} \), whereby \( S_{VR - VL} = 2S_{VR} C_{VR}; d_{TR} = d_{VL}; \)

\[ C^P = \text{Trans}(y_{CV}, y_{CV+}) \text{Trans}(z_{E}, z_{CV} + z_{V+}) \text{Rot}(y_{V}, \theta_V); \]
\[ \text{We now substitute (6) into (10) to find } \]
\[ B^T = \text{Rot}(x_{R}, \theta_R, y_{CS+}) = \text{Trans}(x_{R}, -s) \text{Trans}(y_{R}, y_{CS+}) \text{Trans}(z_{R}, z_{CS}). \]

C. Constrained Kinematics

As the use of the first DOF, roll, is not prevalent in current head designs, for this section we consider \( \theta_R \) to be zero at all times. Let us further assume that each camera is looking at the same target so that the elevation axes work as a common elevation, i.e. \( \theta_E = \theta_{EL} = \theta_{VR}. \) This configuration constrains the \( z_{VL} \) and \( z_{VR} \) to a single plane, whereby simple triangulation confirms the following constraints on \( \theta_{VL}, \theta_{VR}, d_{TL}, \) and \( d_{TR} \):
\[ d_{TR} C_{VR} \equiv d_{TL} C_{VL}; \]
\[ 2\delta = d_{TR} S_{VR} - d_{TL} S_{VL} \]
\[ d_{TR} = 2\delta C_{VL}/\sin(\theta_{VR} - \theta_{VL}) = 2\delta C_{VL}/S_{VR - VL} \]
\[ d_{TL} = 2\delta C_{VR}/S_{VR - VL}. \]

C. Constrained Kinematics

As the use of the first DOF, roll, is not prevalent in current head designs, for this section we consider \( \theta_R \) to be zero at all times. Let us further assume that each camera is looking at the same target so that the elevation axes work as a common elevation, i.e. \( \theta_E = \theta_{EL} = \theta_{VR}. \) This configuration constrains the \( z_{VL} \) and \( z_{VR} \) to a single plane, whereby simple triangulation confirms the following constraints on \( \theta_{VL}, \theta_{VR}, d_{TL}, \) and \( d_{TR} \):
\[ d_{TR} C_{VR} \equiv d_{TL} C_{VL}; \]
\[ 2\delta = d_{TR} S_{VR} - d_{TL} S_{VL} \]
\[ d_{TR} = 2\delta C_{VL}/\sin(\theta_{VR} - \theta_{VL}) = 2\delta C_{VL}/S_{VR - VL} \]
\[ d_{TL} = 2\delta C_{VR}/S_{VR - VL}. \]

Substituting into the general kinematics (5) yields (9) as shown at the bottom of the page.

Substituting the first such head further that the target is along the gaze direction, \( z_C \); i.e., the cameras verge along the gaze direction only \( \theta_{VR} = -\theta_{VL} \), whereby \( S_{VR - VL} = 2S_{VR} C_{VR}; d_{TR} = d_{TL}; \)

\[ B^T = \text{Rot}(x_{R}, \theta_R, y_{CS+}) = \text{Trans}(x_{R}, -s) \text{Trans}(y_{R}, y_{CS+}) \text{Trans}(z_{R}, z_{CS}). \]

C. Constrained Kinematics

As the use of the first DOF, roll, is not prevalent in current head designs, for this section we consider \( \theta_R \) to be zero at all times. Let us further assume that each camera is looking at the same target so that the elevation axes work as a common elevation, i.e. \( \theta_E = \theta_{EL} = \theta_{VR}. \) This configuration constrains the \( z_{VL} \) and \( z_{VR} \) to a single plane, whereby simple triangulation confirms the following constraints on \( \theta_{VL}, \theta_{VR}, d_{TL}, \) and \( d_{TR} \):
\[ d_{TR} C_{VR} \equiv d_{TL} C_{VL}; \]
\[ 2\delta = d_{TR} S_{VR} - d_{TL} S_{VL} \]
\[ d_{TR} = 2\delta C_{VL}/\sin(\theta_{VR} - \theta_{VL}) = 2\delta C_{VL}/S_{VR - VL} \]
\[ d_{TL} = 2\delta C_{VR}/S_{VR - VL}. \]

Substituting into the general kinematics (5) yields (9) as shown at the bottom of the page.
The kinematic solutions for single eye devices can be trivially obtained by considering the transform, \( W_T = T_E T_v T_{cv} T_{cE} T_{cc} T_{CE} \), from (4), while solutions for three eye robot heads such as Triclops [7] may be found by constraining the vergence axes to the gaze direction using coordinate frames \( \{G\} \) or \( \{U\} \) (where appropriate).

### IV. INVERSE KINEMATICS

The inverse kinematics describe a method for generating the required joint angles of the head given the target point \( \theta D \). The inverse kinematics algorithm, therefore, maps the 3 DOF point into the DOF joint space of the head. For the general head, there are 7 relevant DOF (the 2 cyclostorsion axes have no effect). To develop an efficient inverse kinematics solution we must look at ways of reducing this redundancy. Assuming that the roll axis is used to level the head (for example being driven independently of the vision processing using gyro etc.), then both cameras will elevate, whether mechanically coupled or independently, to the same angle in order to view the same target, thereby reducing the DOF to 5 (pan, baseline, elevation, 2 vergence).

For robot heads with independent pan and vergence axes, two configurations are available: 1) verging the cameras in front of the head or 2) behind the head (where the vergence angles are greater than \( \pm \pi/2 \)). Most systems avoid the latter by either hardware limiting the range of the vergence axes to less than \( \pm \pi/2 \), or limiting the vergence range through the control system, and using the pan axis to bring the target with range of the vergence axes.

Given a target at \((x, y, z)\) relative to the \( \{B\} \) frame, two simple inverse kinematic solutions can be found with the following constraints:

#### A. Vergence on Gaze

In this case, \( \theta_{VR} = -\theta_{VL} \), whereby

\[
\begin{align*}
\theta_P &= \tan^{-1}(x/z) \\
(\delta \cot_{VR} + z_{va})C_E - y_{va}S_E &= \frac{x}{S_P} - z_{CS_o} = A \\
(\delta \cot_{VR} + z_{va})S_E + y_{va}C_E &= y - y_{CS_o} = B \\
-AS_E + BC_E &= y_{va} \\
\theta_E &= \cos^{-1}(y_{va}/\sqrt{(A^2 + B^2)}) - \cos^{-1}(B/\sqrt{(A^2 + B^2)}) \\
\delta \cot_{VR} &= (A + y_{va}S_E)/C_E - z_{va}.
\end{align*}
\]

The last equation indicates that the baseline \( \delta \) is directly coupled to the vergence angle \( \theta_{VR} \). A dynamic solution to this coupling can be found by slaving the dynamically slower axes off the faster one, allowing the DOF to be effectively reduced to 3. Note, when evaluating the inverse of trigonometric functions great care must be exercised to ensure that the sign of the angle is preserved. This solution becomes ill-conditioned as \( z \) tends to zero, or when the resulting vergence or elevation angles approach \( \pm \pi/2 \) (see Section VI).

#### B. Vergence off Gaze (Version)

By assuming that the baseline will only be adjusted from one fixed position to another, depending on application, we again reduce the problem to 4 DOF. A simple solution arises from noting that the pan axis will have to drive a significantly higher load inertia than either of the vergence axes. Thus the dynamic performance of the pan axis will generally be lower than that of vergence. By slaving the pan axis from \( \theta_{VR} = \theta_{VL} \), where \( \theta_{VR} = \theta_{VL} = 0 \) indicates that the target is along the gaze direction, the inverse kinematics reduces again to a 3 DOF problem. The current value of \( \theta_P \) is then used along with the position of the target to derive the required positions for \( \theta_V \). Premultiplying both sides of (9) with \( \text{Rot}(\theta_P, \theta_V) \) we obtain

\[
\begin{align*}
\frac{d_{TR} S_{VR} - \delta}{S_E(d_{TR} C_{VR} + z_{va}) + y_{va}C_E + y_{CS_o}} &= \frac{y}{x} \frac{x_{CP} - z S_P}{x S_P + z C_P} \\
\frac{d_{TR} S_{VR}}{S_E(d_{TR} C_{VR} + z_{va}) + y_{va}C_E + y_{CS_o}} &= \frac{y_{va}C_E - z S_E}{y_{va}} = \frac{y - y_{CS_o}}{y_{va}} \\
\theta_E &= \cos^{-1}(y_{va}/\sqrt{(A^2 + B^2)}) - \cos^{-1}(B/\sqrt{(A^2 + B^2)}) \\
\theta_{VR} &= \cot^{-1}(x_{va} - y_{va}C_E)/S_E - z_{va}/x_{va}.
\end{align*}
\]

and we can solve for \( d_{TR} \) (i.e., \( \theta_{VL} \)), in a similar way to before.

### V. JACOBIANS

For robot manipulators, the head Jacobian matrix describes the relationship between the velocity of the end effector and the velocities of the joints, the size of the Jacobian being \( n \times m \), where \( n \) is the number of DOF of the manipulator and \( m \) is the number of DOF of the output space. The Jacobian varies with joint positions and can be found by differentiating the forward kinematics solution with respect to time. For a point in space there are just three DOF so the Jacobian for the generalized head may be written as a 3 × 7 matrix such that

\[
\begin{bmatrix}
\dot{x} \\
\dot{y} \\
\dot{z}
\end{bmatrix} = J(\Theta, \delta)[\dot{\theta}_R^T \dot{\theta}_P \dot{\theta}_{EL} \dot{\theta}_{ER} \dot{\theta}_{VL} \dot{\theta}_{VR} \delta]^T.
\]

The importance of the Jacobian arises when, given a target position and velocity, one wishes to calculate the joint positions and velocities required to track the target. This requires the matrix \( J \) to be inverted. Problems arise when \( J \) is nonsquare or the determinant of \( J \) tends to zero—at “singularities.” When this occurs, it implies that infinite joint velocities will be required to track even small target velocities. Thus any trajectory generator that uses the inverse Jacobian should avoid known singularities of the robot. For the nonsquare Jacobian of the robot head, the singularities can be found when the rank of the Jacobian is less than \( \min(n, m) = 3 \), reflecting the 3 DOF of the output space.

The Jacobian matrix for the constrained robot head is a 3 × 5 matrix, mapping the velocities of \( \dot{\theta}_P, \dot{\theta}_E, \dot{\theta}_{VL}, \dot{\theta}_{VR}, \dot{\delta} \) to the
velocities of \([x, y, z]^T\) where

\[
J_{11} = -\delta p \frac{S_{VR+VL}}{S_{VR-VL}} + C_p \left( C_E \left( 2h \frac{C_{VR}C_{VL}}{S_{VR-VL}} + z_{V_R} \right) - y_{V_R}S_E + z_{CSa} \right) = \gamma \\
J_{12} = -s_p \left( S_E \left( 2h \frac{C_{VR}C_{VL}}{S_{VR-VL}} + z_{V_R} \right) + y_{V_R}C_E \right) \\
J_{13} = \delta \frac{C_p S_{2VR} + 2s_p C_E C_{VR}^2}{S_{VR-VL}} \\
J_{14} = \delta \frac{-C_p S_{2VL} - 2s_p C_E C_{VL}^2}{S_{VR-VL}} \\
J_{15} = C_p S_{VR+VL} + 2s_p C_E C_{VR} C_{VL} \\
J_{21} = 0 \\
J_{22} = C_E \left( 2h \frac{C_{VR}C_{VL}}{S_{VR-VL}} + z_{V_R} \right) - y_{V_R}S_E \\
J_{23} = 2h \frac{S_E C_{VR}^2}{S_{VR-VL}} \\
J_{24} = 2h \frac{S_E C_{VL}^2}{S_{VR-VL}} \\
J_{25} = 2s_p C_E C_{VR} C_{VL} \\
J_{31} = -\delta C_p \frac{S_{VR+VL}}{S_{VR-VL}} - S_p \left( C_E \left( 2h \frac{C_{VR}C_{VL}}{S_{VR-VL}} + z_{V_R} \right) - y_{V_R}S_E + z_{CSa} \right) = -\gamma \\
J_{32} = -C_p \left( S_E \left( 2h \frac{C_{VR}C_{VL}}{S_{VR-VL}} + z_{V_R} \right) + y_{V_R}C_E \right) \\
J_{33} = \delta \frac{-S_p S_{2VR} - 2C_p C_E C_{VR}^2}{S_{VR-VL}} \\
J_{34} = \delta \frac{S_p S_{2VL} - 2s_p C_E C_{VL}^2}{S_{VR-VL}} \\
J_{35} = -\delta s_p S_{VR+VL} + 2C_p C_E C_{VR} C_{VL} \tag{14}
\]

Finding the singularities from the 5 \(\times\) 3 matrix above is an involved and lengthy process. However, the common factor in columns 3 and 4, \((\delta / S_{VR-VL})\), immediately yields a trivial singularity at \(\delta = 0\). By looking at the constrained options of Section IV, and assuming that the baseline is fixed, the Jacobians reduce to 3 \(\times\) 3 matrices and the singularities for each configuration can be deduced.

### A. Vergence on Gaze

The determinant is given by

\[
|J_{P,E,V}| = \frac{\delta}{2S_{VR}} \Delta \left( \delta S_E - y_{V_R}S_E + z_{CSa} \right) \\
\Delta = \frac{\delta C_{VR}}{S_{VR}} + z_{V_R}. \tag{15}
\]

The most obvious singularity occurs at \(\Delta = 0\), i.e., at \(\theta_{V_R} = \tan^{-1}(-\delta / z_{V_R})\). Thus \(\Delta = 0\) represents when the cameras are pointing directly at the origin of the gaze frame \([G]\). If the vergence and elevation axes intersect \((z_{V_R} = 0)\), this singularity occurs at \(\theta_{V_R} = \pm \pi/2\), as expected.

### B. Vergence off Gaze (Version)

In this case, we assume that the pan axis is driven, as before, from the “off-gaze” angle \(\theta_{VR} + \theta_{VL}\). The pan axis velocity will, therefore, have no impact on the trajectory generation and the Jacobian will be concerned with the mapping of velocities from the elevation and independent vergence axes to Cartesian space. The pan axis can be assumed to be stationary, say \(\theta_{VR} = 0\). The determinant in this case is

\[
|J_{E,V,L}| = \frac{\delta^2}{2S_{VR-VL}} \Delta \left( \delta C_{VR}C_{VL} \right) \\
\Delta = \frac{\delta C_{VR}C_{VL}}{S_{VR-VL}} z_{V_R}. \tag{16}
\]

With \(z_{V_R} = 0\), then the singularities occur when either \(\theta_{VR} = \pm \pi/2\) or \(\theta_{VL} = \pm \pi/2\), with \(z_{V_R} \neq 0\) accounting for the offset between vergence and elevation. Physically, when close to this singularity movement in the elevation axis will cause cyclotorsion. Another singularity occurs on the locus \(\theta_{VR} + \theta_{VL} = \pm \pi/2\). As the cameras are pointing at a single point target, the physical interpretation of this is the situation where an object is very close to one camera while being approximately \(2\delta\) from the other—an unlikely situation.

The regions in which the inverse kinematic solutions become ill-conditioned and the regions enclosing the singularities of the Jacobian are the same. However, it is noted that such regions are at the extremes of operation in terms of vision processing, in which the “ideal” operating region is along the unelevated gaze direction, \(z_{U}\), and where the target lies within a minimum and maximum range limit that is determined by the baseline between the two cameras.

### References


