Perspective Geometry Explains Perceived 3D Object Poses in Real Scenes and Pictures

Erin Koch, Famya Baig, Qasim Zaidi

Pose estimation of objects in real scenes, is critically important for biological and machine visual systems, but little is known of how humans infer 3D poses from 2D retinal images. We show that there is unexpectedly remarkable agreement across observers in estimating 3D poses from pictures, and that all observers apply the same inferential rule in all views, utilizing the geometrically derived back-transform from retinal images to 3D scenes. For a camera elevation of ϕ_c, a stick lying at the center of the ground plane with a pose angle of $\mathit{\Omega}_T$, uniquely projects to the orientation, θ_S, on the picture plane:

$$\theta_S = \mathrm{atan}(\tan(\mathit{\Omega}_T)\sin(\phi_c)) \qquad (1a)$$

Seen fronto-parallel to the picture plane (observer viewing angle, $\phi_v = 0$), the orientation on the retina θ_R = θ_S. If observers can assume that the imaged stick is lying on the ground, and if they can approximate ϕ_c , they can use the back-projection of Equation (1a) to estimate the physical 3D pose from the retinal orientation:

$$\mathit{\Omega}_T = \mathrm{atan}(\tan(\theta_R)/\sin(\phi_c)) \qquad (1b)$$

We find that observers use a variant of the proper back-projection. The perceived 3D pose, $\mathit{\Omega}_P$, is predicted by a model that adds just one free parameter, K, to (1b):

$$\Omega_P = \mathrm{atan}(\mathrm{K}\cdot\tan(\theta_R)/\sin(\phi_c)) \qquad (2)$$

The parameter K, acts to model a fronto-parallel bias in observers' judgments.
From oblique observer viewpoints, $\theta_R \neq \theta_S$, and the projection from the screen to the retina needs to be considered. The proper back-projection is given by:

$$\Omega_T = \mathrm{atan}(\tan(\theta_R)\cdot(\cos(\phi_v)/\sin(\phi_c))) \qquad (3)$$

We found that instead of updating the back-projection for oblique viewing of a picture, observers use the same back-projection rule, as for 3D viewing (or fronto-parallel picture viewing), resulting in an illusory rotation of the pictured scene as being oriented towards the observer. We used the best fitting model for 3D scene viewing (Equation (2) with K at the value that best accounted for the fronto-parallel bias). We added a fixed constant, ϕ_v, to predict that the perceived rotation of the scene will be equal to the observer viewpoint angle. We add just one free parameter, L, which multiplies ϕ_C to allow for the observation that the ground plane seems to tilt towards the observer (equivalently, the camera elevation, ϕ_c , increases) in oblique viewing conditions:

$$\Omega_P = \mathrm{atan}(K_0\cdot\tan(\theta_R)/\sin(L\,\phi_c)) + \phi_v \qquad (4)$$

The inferential rules fit the empirical data (RMSE range: (6.19-7.9), across ϕ_v). This reliance on retinal images explains distortions in perceptions of real scenes, and invariance in pictures, including the classical conundrum of why certain image features always point at the observer regardless of viewpoint. These results have implications for investigating 3D scene inferences in biological systems, and designing machine vision systems.