Procedia Computer Science

Volume 88, 2016, Pages 1–7

7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

On the perceptual advantages of visual suppression mechanisms for dynamic robot systems

João Avelino[1], Rui Figueiredo[1], Plinio Moreno[1], and Alexandre Bernardino[1]

Instituto Superior Técnico, Lisboa, Portugal
joao.avelino@tecnico.ulisboa.pt
{ruifigueiredo, plinio, alex}@isr.ist.utl.pt

Abstract
The use of computer vision based methods to explore the surrounding environment and track individuals, is further enhanced by the ability to move the visual system, a process realized by biological organisms. However, several perceptual issues arise from the rapid movement of the image acquisition system: blurred images and visual-proprioceptive transient delays result in incorrect spatial location estimation. Inspired by the biological mechanism of saccadic suppression, our main contribution is a biologically inspired visual stability mechanism able to deal with problems arising from self-motion. Together with a state-of-the-art pedestrian detection algorithm, our proposed methodology contributes to enhancements in person position estimation, thus improving human-robot interaction behaviors.

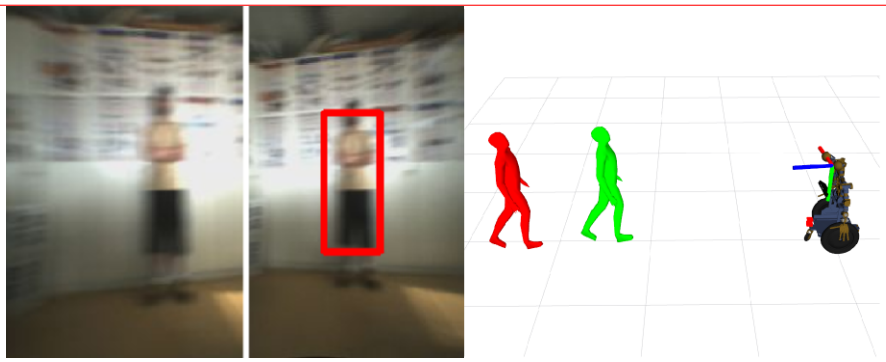
Keywords: Sensory-motor inhibition, Pedestrian Tracking, Active Vision, Visual Suppression

1 Introduction

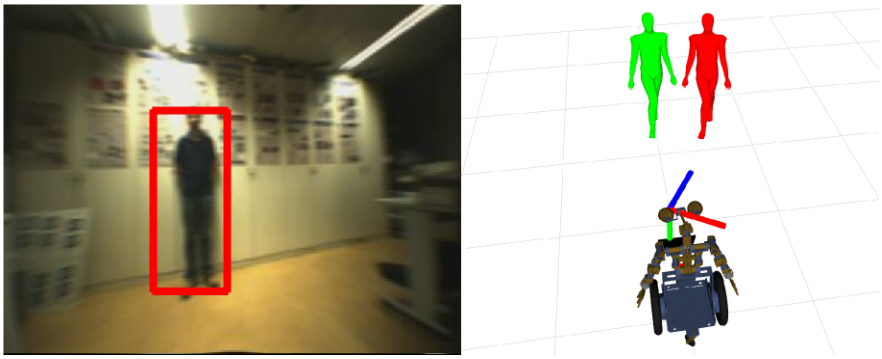
To visually explore and understand the world, humans are endowed with a set of oculomotor mechanisms to direct their eyes to specific locations in the environment. Combined head-eye movements allow shifting the gaze between different targets in the visual scene [4].

Several perceptual issues arise due to the abrupt and ballistic nature of head-eye movements during saccadic gaze shifts [13]. First, fast oculocephalic motions result in blurred retinal images which degrade further processing and analysis. Second, the correct estimation of target's spatial location, requires the on-line combination of exteroceptive (visual) and proprioceptive (body posture and movement) measurements [14]. However, latencies introduced along the sensory pathways due to stimuli transduction, transmission to the brain and processing, incur in unpredictable delays that result in position estimation errors. In fact, studies from neuropsychology have shown evidence that human proprioception is ahead of vision by approximately 50 ms [3]. Still, by resorting to visual stability mechanisms [8], humans are capable of dealing robustly with the aforementioned problems [12]. The neuronal mechanisms behind visual sta-

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016



(a) Missed and wrong position estimations due to heavy motion blur



(b) Wrong position estimations due to visuo-proprioceptive delays

Figure 1: The main perceptual issues that arise in active vision systems during head-eye motion. Green and red pedestrians represent ground truth and estimated positions, respectively.

bility has been extensively addressed in the literature, but still not fully uncovered. We refer the interested reader to [15, 2] for detailed reviews.

As illustrated in Figure 1, visuomotor sensor fusion remains problematic and challenging in applications involving artificial active vision systems [1]. Likewise, robust integration, synchronization and coordination of visual and proprioceptive information are essential for accurate target position estimation and tracking. Developing artificial systems that emulate biological visual stability mechanisms is of the utmost importance for many robotic applications, namely in effective HRI and hand-eye coordination during robotic grasping.

In this work we investigate the impact of visual stability mechanisms on the pedestrian tracking problem, which deals with determining people's 3D locations relative to the observer. More specifically, we develop a biologically inspired visual inhibition algorithm that mimics the *saccadic masking* mechanism [11], which is responsible for temporary blinding the observer during fast head-eye movements. We combine our method with a state-of-the-art real-time pedestrian detection algorithm that uses the Aggregate Channel Features (ACF) [5]. The benefits of the proposed approach are twofold: better pedestrian position estimation and improved computational performance, as a result of discarding irrelevant defective sensory information, acquired during self-motion. The proposed visual inhibition algorithm can be combined with any general target estimation algorithm and was implemented in C++ to make it suitable for real-time applications involving robotic platforms with active vision heads.

The remainder of this paper is organized as follows. In section 2 we describe the various components of our biologically inspired system for pedestrian position estimation. In section 3 we evaluate our methodologies with data obtained with a real robotic platform supplied with an anthropomorphic head. Finally, in section 4 we draw the final conclusions and remarks.

2 Methodologies

In the following section we describe our cognitive architecture for robust detection and visual tracking of pedestrians. The proposed architecture depicted in Figure 2, relies on active vision and consists in the integration of several cognitive blocks: pedestrian detection, 3D position estimation and visual suppression.

2.1 System Overview

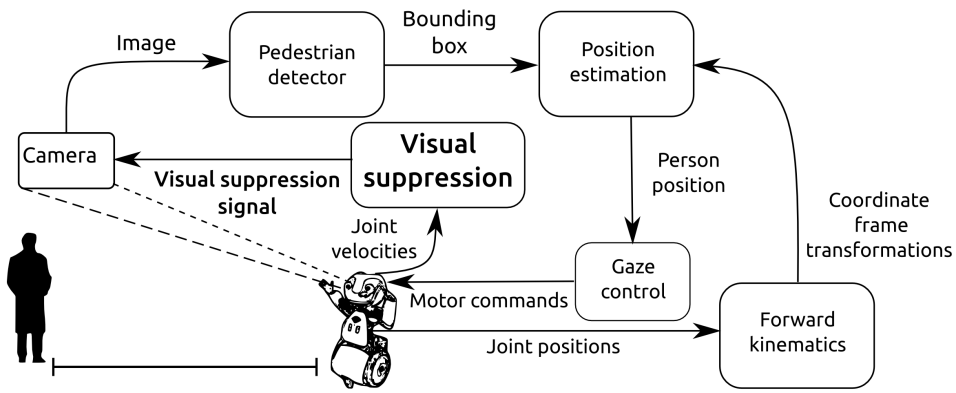


Figure 2: General overview of our architecture

We start by detecting people from visual information provided by a single camera. We use a pedestrian detector extract bounding boxes from a stream of monocular images, which are used to compute the 3D location and the height of a target person. The obtained 3D measurements are integrated over time using a Kalman Filter tracking algorithm [7]. Finally, the position and height information is used to center the target in the field of view of the robot, by controlling the gaze fixation point in 3D Cartesian space [10].

2.2 Pedestrian Detection

We have adopted the pedestrian detector proposed in [5], that provides a good balance between accuracy and speed, hence being suitable for real-time tracking applications. This detector employs a sliding window detection–by–classification approach: each detection window is classified as "person" or "not person". Classification is performed using boosted decision trees, trained with labeled samples of full body pedestrians, using the Adaboost algorithm [6].

The classification method relies on features that combine several image channels, including LUV, Gradient Magnitude and HOG channels, aggregated in a block-wise manner. For multi-scale detection, the method uses multi-channel pyramids. The computational burden of

constructing full pyramids is cleverly avoided by approximating in-between scales from interpolations of the coarse scales. Finally, non-maximum suppression is applied in order to avoid multiple detections (only a few pixels apart) that correspond to the same person.

2.3 3D Position Estimation

In typical HRI scenarios, people and robots stand on a common flat surface (e.g. floor). This assumption provides a sufficient constraint that allows us to extract the 3D coordinates of a person feet. Let us denote by $\mathbf{p} = [p_x\ p_y\ p_z]'$ the person's feet coordinate in the robot's base frame, x and y the coordinates in the image plane, and $\mathbf{P}$ the camera's projection matrix obtained from the robot's proprioception and camera's intrinsic parameters. The $i-th$ column of the camera matrix is denoted by $\pi_\mathbf{i}$, $\mathbf{P} = \begin{bmatrix} \pi_\mathbf{1} & \pi_\mathbf{2} & \pi_\mathbf{3} & \pi_\mathbf{4} \end{bmatrix}$. Assuming that a person is standing in the ground plane we can set $p_z = 0$. This constraint allows cancelling out the third column of $\mathbf{P}$ and obtaining a simple homography, $\mathbf{H} = \begin{bmatrix} \pi_\mathbf{1} & \pi_\mathbf{2} & \pi_\mathbf{4} \end{bmatrix}$. This allows the transformation of feet points from the image plane to the 3D robot base frame according to:

$$\begin{bmatrix} \lambda p_x \\ \lambda p_y \\ \lambda \end{bmatrix} = \mathbf{H^{-1}} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

where λ is a scale factor.

2.4 Gazing

To keep visual track of a person, we use 3D position and height information to command the robot's fixation point[1]. We are able to obtain reliable detections except if the person is too close to fully appear in the image, or too far to be detected. To avoid unreachable position commands, we set a limit to the range of possible fixation points.

2.4.1 Visual Suppression

Inspired by the mechanisms of visual suppression during rapid eye movements, we propose a biologically plausible method to prevent the degradation of the quality of pedestrian tracking due to the problems described in section 1. However, "blinding" the system should not be made simply by discarding visual information whenever the system is moving. Such an approach would render the robot blind most of the time, preventing smooth pursuit movements and eventually leading to the divergence of estimation and filtering methods due to the unavailability of recent information. For these reasons, the implementation of a suppressive approach requires careful analysis of proprioceptive feedback. More specifically, we study how joint's velocities degrade visual perception. The proposed suppression criteria analyses head-eye joints' velocities and triggers visual masking if:

$$\left|\omega^{h\_pan} + \omega^{h\_tilt} + \omega^{e\_tilt} + \omega^{e\_pan}\right| > \omega_{th} \tag{2}$$

where ω_{th} is an angular velocity, saccadic masking threshold, to be carefully selected.

[1] the point in the visual field that is fixated by the two eyes in normal vision and for each eye is the point that directly stimulates the fovea of the retina [4]

3 Experiments

The platform used in this work was the Vizzy robot [9] (see Figure 3), a humanoid upper body and head standing on top of a mobile base. Vizzy has na anthropomorphic head allowing human-like gazing behavior. Since Vizzy heavily relies on computer vision and has a wide range of possible eye-head motions, it is an ideal platform to demonstrate the benefits of the proposed visual suppression mechanism.

(a) (b)

Figure 3: The experimental scenario (a) created for evaluating the proposed visual methodologies with the Vizzy robot (b).

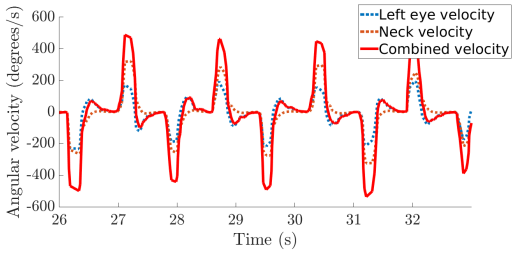
To investigate the impact of our visual stability methodology on pedestrian location estimation, we created the following experimental scenario. A person was standing still in front of Vizzy, at a known ground truth location, while Vizzy switched its gaze between two fixation points. The resulting trajectory in joint velocity space is depicted in Figure 4a. Detections along this trajectory were used to estimate the person location (x, y position in 3D robot base coordinates). The combined velocity space was then discretized into bins of $15°/s$ and the estimated positions were determined for each bin. Visual and proprioception information was acquired at 60Hz. Finally, in order to obtain statistically significant results, we let the experiment run until each bin (up to $465°/s$) had a minimum of 100 samples.

We varied the saccadic masking threshold ω_{th} (Figure 4b) and evaluated the quality of the estimated position using several statistical metrics suitable to assess the accuracy and precision of the estimates resulting in the box plot show in Figure 5.

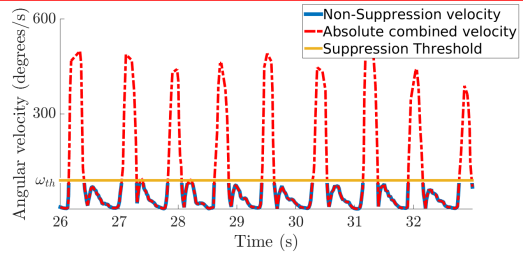
As can be seen in Figure 5, the best results were obtained for lower thresholds, which corresponds to blinding Vizzy whenever it moves. In other words, increasing ω_{th}, increases the median, the 75th percentile and the maximum, meaning less detection accuracy and precision.

In Figure 4c we show the result of applying visual suppression to the system. Choosing an $\omega_{th} = 105°/s$ leads to improved position estimates when comparing to a system without suppression ($\omega_{th} = \infty$). Furthermore, choosing an $\omega_{th} = 90°/s$ further improves the results as expected given the previous analysis of Figure 5. This value leads to improved position estimates while still accounting for reliable visual-proprioceptive information when performing smooth movements.

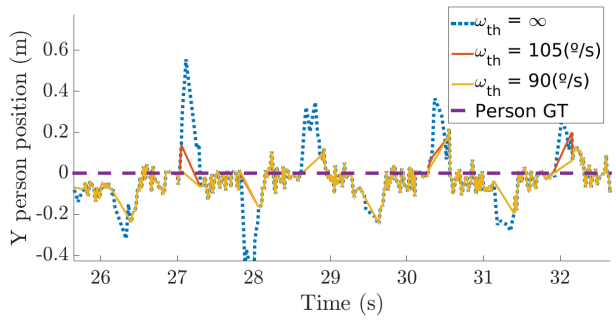
The remaining position errors (up to ± 0.1 m) are due to image quantization, intrinsic/extrinsic calibration errors, imperfect kinematics modeling and pedestrian detector oscillations.



(a) Joint angular velocity profiles



(b) Example threshold for visual suppression



(c) Estimated person trajectories using visual suppression ($\omega_{th} = 90°/s$ and $\omega_{th} = 105°/s$) and not using visual suppression ($\omega_{th} = \infty$)

Figure 4: The proposed saccadic suppression based on joint velocities proprioceptive feedback.

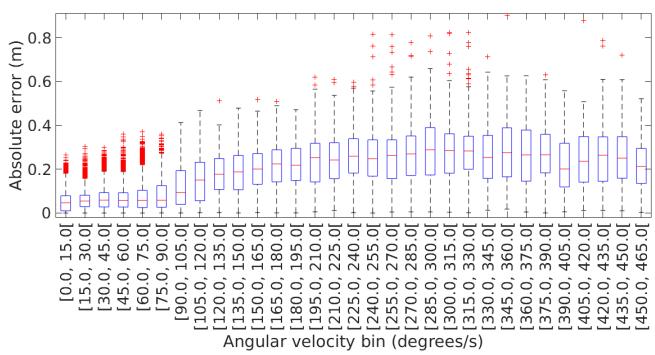


Figure 5: Box plot of the absolute error for various angular velocities

4 Conclusions

In this work we have implemented a biologically inspired visual stability mechanism that mimics *saccadic masking*, in order to deal with perceptual problems that arise during fast eye-head movements. We tested our methodology in a pedestrian position estimation scenario. The obtained results demonstrate that using proprioceptive feedback to suppress visual information during saccadic eye movements is advantageous for target position estimation tasks, since it produces smoother and more accurate target position estimates. The methodology to tune the suppression threshold depends on the particular system specifications and application require-

ments. On one hand, the lower the suppression threshold, the lower the number of samples, which might lead to the divergence of some estimation algorithms. On the other hand, increasing the threshold might overload the computational apparatus with noisy sensory information. Moreover, the threshold choice should account for the trade-off between acceptable estimation errors and task-execution speed when switching between multiple dynamic targets.

Acknowledgements

This work was supported by the FCT projects [UID/EEA/50009/2013], the project AHA [CMUP-ERI/HCI/0046/2013] and the FCT doctoral grant [SFRH/BD/105779/2014].

References

[1] Olli Alkkiomaki, Ville Kyrki, Heikki Kalviainen, Yong Liu, and Heikki Handroos. Challenges of vision for real-time sensor based control. In *Computer and Robot Vision, 2008. CRV'08. Canadian Conference on*, pages 42–49. IEEE, 2008.

[2] David Burr. Eye movements: keeping vision stable. *Current Biology*, 14(5):R195–R197, 2004.

[3] Brendan D Cameron, Cristina de la Malla, and Joan López-Moliner. The role of differential delays in integrating transient visual and proprioceptive information. *Multisensory Integration in Action Control*, page 64, 2014.

[4] R. H. S. Carpenter. *Movements of the eyes*. London Pion, 1988.

[5] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *PAMI*, 2014.

[6] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In Lorenza Saitta, editor, *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, pages 148–156. Morgan Kaufmann, 1996.

[7] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[8] David Melcher. Visual stability. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1564):468–475, 2011.

[9] Plinio Moreno, Ricardo Nunes, Rui Figueiredo, Ricardo Ferreira, Alexandre Bernardino, José Santos-Victor, Ricardo Beira, Luís Vargas, Duarte Aragão, and Miguel Aragão. *Robot 2015: Second Iberian Robotics Conference: Advances in Robotics, Volume 1*, chapter Vizzy: A Humanoid on Wheels for Assistive Robotics, pages 17–28. Springer International Publishing, Cham, 2016.

[10] Alessandro Roncone, Ugo Pattacini, Giorgio Metta, and Lorenzo Natale. A cartesian 6–dof gaze controller for humanoid robots. In *Proceedings of Robotics: Science and Systems*, AnnArbor, Michigan, June 2016.

[11] John Ross, M Concetta Morrone, Michael E Goldberg, and David C Burr. Changes in visual perception at the time of saccades. *Trends in neurosciences*, 24(2):113–121, 2001.

[12] Fabrice R Sarlegna and Pratik K Mutha. The influence of visual target information on the online control of movements. *Vision research*, 110:144–154, 2015.

[13] Benjamin W Tatler, Nicholas J Wade, Hoi Kwan, John M Findlay, and Boris M Velichkovsky. Yarbus, eye movements, and vision. *i-Perception*, 1(1):7–27, 2010.

[14] Robert J van Beers, Anne C Sittig, and Jan J Denier van Der Gon. Integration of proprioceptive and visual position-information: An experimentally supported model. *Journal of neurophysiology*, 81(3):1355–1364, 1999.

[15] Robert H Wurtz. Neuronal mechanisms of visual stability. *Vision research*, 48(20):2070–2089, 2008.