POSTER: Robotic Visual and Inertial Gaze Control using Human Learning

José Prado
University of Coimbra, Portugal
jaugusto@isr.uc.pt

Jorge Lobo
University of Coimbra, Portugal
jlobo@isr.uc.pt

Jorge Dias
University of Coimbra, Portugal
jorge@isr.uc.pt

Figure 1: Robotic Head

ABSTRACT

Humans make use of inertial and vision cues to determine ego-motion. Bayesian models can be used to represent the human behaviour to be used in a robot. An environment may be composed by an infinite number of variables and humans deal with some of them each time a motor decision needs to be taken.

1 INTRODUCTION

One of our main challenges was finding variables that model the environment in way similar to humans. Of course it was needed to start from accepting some assumptions. In this work we are only considering visuo-inertial information. The visual information that we use is the mean of sparse optical flow based on feature tracking. We show that the mean flow is strongly related with the robotic ego-motion but alone it cannot deal with ambiguous situations that arise. Further we concluded that the fusion between inertial and visual information solves ambiguity problems.

Ambiguity is another problem that happens even on humans, if we look only to a black wall and someone shake our head, we could not know that the head was moving only by using the vision, but we are sure our head is being shake because of our vestibular system. In our robotic platform, the same ambiguity happens, when turning off the lights of the room the robot will act only with the inertial information.

Inertial sensors explore intrinsic properties of body motion. From the principle of generalised relativity of Einstein we known that only the specific force on one point and the angular instantaneous velocity.

No other quantity concerning motion and orientation with respect to the rest of the universe, can be measured from physical experiments inside an isolated closed system. Therefore from inertial measurements one can only determine an estimate for linear acceleration and angular velocity. Linear velocity and position, and angular position, can be obtained by integration.

2 RELATED WORK

Fusion of inertial and visual information was done in [1] and good results were obtained. Unit sphere projection camera model was used, providing a simple model for inertial data integration. Using the vertical reference provided by the inertial sensors, the image horizon line could be determined. Using just one vanishing point, it is possible to recover the camera's focal distance. In a typical indoor corridor scene, the vanishing point can also provide an external bearing for the system's navigation frame of reference.

The integration of inertial sensors can reduce ambiguities and improve robustness of structure from motion methods. The dual problem of motion estimation from observed structure has long been pursued. It is also said in [2] that these sensors have useful complementarities, each able to cover the limitations and deficiencies of the other. [3]

3 PROPOSED APPROACH

Our robotic implementation of gaze control basically perform visuo-inertial servoing. But the sensory inputs are being processed using Bayesian inference using the current instantaneous data and a previous probabilistic learned table.

The image optical flow provides motion data from the visual sensor. We analysed and implemented a variation of Lucas Kanade sparse optical flow building upon the ideas in [4]. Thus, we also decided to rely in the following assumption: the information from the image motion which is strongly related with the human decision when reacting to some stimulus is the mean flow.

It is intuitively preferable to have a real time application, since the idea was to show the flow arrows in the screen and also the inertial data on the fly to influence the human during the learning phase. Thus, performance challenge was instantaneously added to our project idea together with the purpose of using a robust

 ISBN 978-80-86943-95-4

flow algorithm when dealing to changes of lighting, size or image motion.

3.1 Sparse optical flow

We used an algorithm based on a pyramidal implementation of the classical Lucas-Kanade algorithm. An iterative implementation of the Lucas-Kanade optical flow computation provides sufficient local tracking accuracy. RANSAC was implemented to remove the outlier vectors and then the meanflow was calculated with the remaining flow vectors (fig. 2).

3.2 Bayesian Model of Visuo-vestibular-Based Gaze Control

The contribution in the mean flow calculation allows a better characterisation of the human decision.

The purpose is about reaching to a feasible Bayesian model for a robotic gaze control following the ideas presented on[5] and [6]. From the camera image we extract Fd^t and Fa^t which are the direction and the amplitude of the mean flow. From the IMU(Inertial Measurement Unit) we get the angles (Roll, Pitch and Yaw) that are further shown in this paper as R^t (for Roll), P^t (for Pitch), and Y^t (for Yaw).The actuator control acts based on current angular position and on instantaneous flow information of the system. The pan-tilt unit are controlled with combined commands for target position and velocity. The motors move to the desired target position with the selected velocity and stop. The motor model takes this into account by having the current motor command depending on the current state and also on the probabilistic table filled out with the human reaction information.

The following variables are used:

- S_t: is a tuple with the following four variables transformed in possible motor reactions
  - R^t: (roll) angle of the human-reaction for a given state
  - Y^t: (yaw) angle of the human-reaction for a given state
  - Fd^t: direction of the vector of the mean flow (comes from vectored product between u and v) (Radians)
  - Fa^t: amplitude of the vector of the mean flow
- M_t: is a movement variable with the following scope (UP, DOWN, LEFT, RIGHT, STOP)
  - The five states of M_t are concluded by doing atomisation of the raw values in the following variables
    - pan motor velocity: $\mathscr{P}_\omega$ — pan motor target position: $\mathscr{P}_\theta$
    - tilt motor velocity: $\mathscr{T}_\omega$— tilt motor target position: $\mathscr{T}_\theta$
- H_t: is the human reaction to be learned (UP, DOWN, LEFT, RIGHT, STOP)

To simplify notation, state variables are grouped in a vector $S = (R^{0:t}, P^{0:t}, Y^{0:t}, Fd^{0:t}, Fa^{0:t})$ and motor variables are considering to be in the range U,D,L,R,S after atomisation from $M = (\mathscr{P}_\omega, \mathscr{P}_\theta, \mathscr{T}_\omega, \mathscr{T}_\theta)$. The Bayesian program that show the relation between these variables is shown in (fig. 3).

4 EXPERIMENTAL PARADIGM AND PROTOCOLS

4.1 Apparatus and Stimuli

We want to use a HMD (Head Mounted Device) to pass the visual stimulus from the robot to the human subject. However currently we used the input from the keyboard reflexes when the subject looks to the screen. Our robotic head (fig.1) is a common platform consisted by many sensors, but basically those that we are using are a stereo camera, a pan-tilt unit and a inertial sensor. Those sensors are attached in the same structure. Thus, every motor command sent to the pan-tilt unit will reflect on IMU (Inertial Measurement Unit) and also in the camera images. Consequently the sent motor commands will also have a direct influence on the calculated optical flow and inertial data.

.

Attitude Heading Reference System (AHRS) are 3-axis sensors that provide heading, attitude and yaw information for aircraft. AHRS are designed to replace traditional mechanical gyroscopic flight instruments and provide superior reliability and accuracy. We used a low grade AHRS that is a Xsens MTi , which uses a combination of 3-axes accelerometers, gyroscopes and magnetic sensors to output estimates of its own orientation in geo-referenced coordinates. The AHRS orientation can be given in the form of a rotation matrix ${}^{\{\mathscr{W}\}}\mathbf{R}_{AHRS}|_i$ which register the AHRS sensor axes with the north-east-up axes. The camera is a videre STH-MDCS3-9cm stereo camera. Videre Design makes stereo imaging hardware for real-time 3D imaging using the triangulation principle. Stereo images are transferred to a PC using the IEEE 1394 (FireWire) bus.

4.2 Subjects

Five human subjects with normal working visual and vestibular systems. In those subjects with visual distortion this should be compensated by using glasses or lens, thus the distortion perform no impact to the experiment. We will mix male and female according to availability of the persons. The subjects should be at least three naive and two authors.

4.3 Protocol

By using the HMD to give the robot images to the human eyes the visual connection becomes direct between the robot and the human. We can not inject artificial inertial sensor data into human brain. Thus, what is possible to be done is a indirect correlation where during the tests, the human will use it's own vestibular system while the robot will use the artificial inertial system. Gray scale images are the input, with several visual detectable features on the environment. Visual Features are necessary by the human brain to have notion of motion. If a human is moving sideways in front of a perfectly white wall, once the acceleration stabilises subject will have no sensation of motion. However if this same wall is full of visual detectable features, human will naturally detect the motion only by the visual influence. We have the same response on artificial optical flow algorithms, and that's why we are interested on considering the optical mean flow as a artificial visual ego-motion notion measurement variable.

5 RESULTS AND CONCLUSION

A first version of human-learning was implemented, using keyboard to control the robot while monitored by a human (human in the loop way of learning like in [7]). We still want to improve our way of learning by using a helmet equipped with camera and IMU and then detecting real human neck movements.

Consider that we numbered our random variables as follows:

1. is $Roll^t$, sub-variable of Imu^t variable
2. is $Pitch^t$, sub-variable of Imu^t variable
3. is Yaw^t, sub-variable of Imu^t variable
4. is Fd^t, the Flow Direction
5. is Fa^t, the Flow Amplitude

The learned table (fig. 4) is a 4D probability table with dimensions [36x5x10x5] in our test, we plotted this in five 3D graphics in order to be possible to visualise them.

It is possible to observe that for the UP, DOWN, LEFT and RIGHT movements, the main categorising random variable is Fd^t, in the other hand for the STOP movement Fd^t is very confusing, thus Imu^t will be much more useful categorising this decision.

Testing the reaction of the system

Fake stimulus were injected into the system to measure if the robot's reaction would be like expected for that stimulus. As human trained the system, we know (approximately) which stimulus to create and which reaction to expect. For example if we train a walking robot not to fall from a step, we can put a step in front of it and our expectation will be that the robot do not fall. In our case we trained the head to be centred and then we give stimulus simulating that the head would moving to one or another side "forever" during each test. We also gave stimulus for the system to believe the head was flying up like a rocket and also falling down in free fall. It was performed 100 trials with different stimulus for each expected reaction.

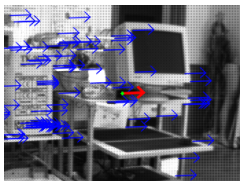


Figure 2: Sparse Flow Experimental Test

Program

- Description
  - Specification
    - Variables:
    - $S^{0:t} = (R^{0:t}, Y^{0:t}, Fd^{0:t}, Fa^{0:t})$: state variable
    - $M^{0:t}$: motor movement variables, scope {U,D,L,R,S}
    - $H^{0:t}$: human reaction variables, scope {U,D,L,R,S}
    - $Fd^{0:t}$: flow direction (Rad)
    - $Fa^{0:t}$: flow magnitude (pixels)
    - Decomposition:
    - $P(S^t \wedge M^t \wedge H^t \wedge Fd^t \wedge Fa^t =$
    - $P(R^t.P(y)^t)$
    - $.P(H|R \wedge y)$
    - $.P(Fd|H \wedge R \wedge y)$
    - $.P(Fa|Fd \wedge H \wedge R \wedge y)$
    - $.P(M|Fa \wedge Fd \wedge H \wedge R \wedge y)$
    - Parametric forms:
    - $P(M^t) = P(H^{0:t}|S^t)$
  - Identification:
  - Bayesian human-learned gaze control.
- Question:
- $P(M^t|S^t \wedge H^{0:t})$

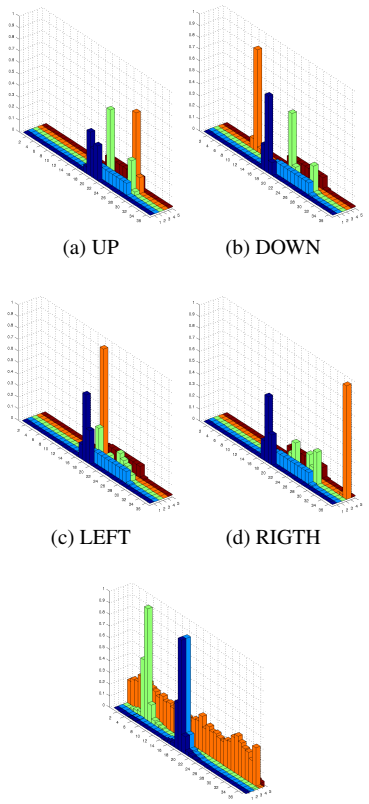
Figure 3: BayesianProgram

One stimulus for each reaction would be the trivial case to categorise, but for this preliminary results we had approximately 98% of correct decisions in 500 different stimulus to be categorised in 5 movements.

6 FUTURE WORK

In this work we performed a learning based on human reactions, and the Bayesian program we proposed efficiently models the ego-motion with visuo-inertial information. Adding non-representative variables to the Bayesian model may not only waste performance, rather it may also cause confusion to the algorithm when trying to categorise the correct reaction to take.

The system is running real time at 3fps, we did not use a real time operational system, so this frequency

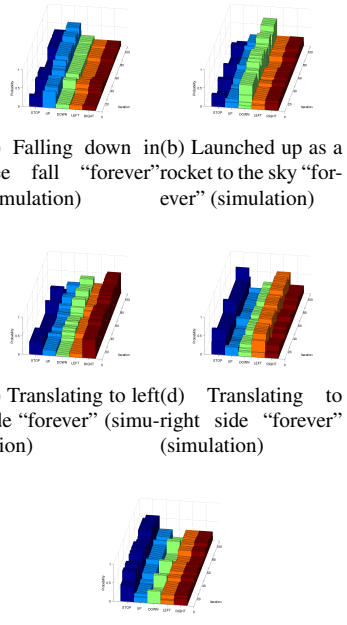


(a) UP (b) DOWN

(c) LEFT (d) RIGTH

(e) STOP

Figure 4: Probability Table - Learned data -



(a) Falling down in free fall (simulation)

(b) Launched up as a "forever" rocket to the sky "forever" (simulation)

(c) Translating to left side "forever" (simulation)

(d) Translating to right side "forever" (simulation)

(e) Stopped

Figure 5: System reaction (Imu^t vary even stopped because of Mag^t mainly)

may vary a little according to our non real time UNIX system. Learning of motor behaviours was done by our Bayesian program and it allowed fusion between inertial and visual input.

For future work, we intend to perform more experiments with different environments. Restrict the environment, for example a white point in a black background would be a way to push the human to pay attention in something near to the flow, rather then in other things that we cannot predict. Another desired set up is adding more moving objects into the scene it will be interesting to verify what happens if we try to exceed the gross errors that our RANSAC implementation is able to filter. Also we want to improve the speed of the system motor reaction, a new hardware platform is already being constructed. This new platform will have a binocular active vision system witch will allow stereo vergence with an adjustable base line, with common head tilt and neck pan, mimicking the human degrees of freedom.

ACKNOWLEDGMENTS: The authors gratefully acknowledge support from EC-contract number BACS FP6-IST-027140, the contribution of the Institute of Systems and Robotics at Coimbra University and reviewers' comments

REFERENCES

[1] Jorge Lobo and Jorge Dias. Vision and inertial sensor cooperation using gravity as a vertical reference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1597–1608, December 2003.

[2] Peter Corke, Jorge Lobo, and Jorge Dias. An introduction to inertial and visual sensing. *The International Journal of Robotics Research (IJRR) Special Issue from the 2nd Workshop on Integration of Vision and Inertial Sensors.*, 26(6):519–535, June 2007.

[3] Jorge Lobo and Jorge Dias. Relative pose calibration between visual and inertial sensors. *The International Journal of Robotics Research (IJRR) Special Issue from the 2nd Workshop on Integration of Vision and Inertial Sensors.*, pages 561–575, 2007.

[4] Andrew Lookingbill, David Lieb, David Stavens, and Sebastian Thrun. Learning activity-based ground models from a moving helicopter platform. *In Proceedings of the International Conference on Robotics and Automation*, 5:3948–3953, 2005.

[5] Pierre Bessiere, Juan-Manuel Ahuactzin, Kamel Mekhnacha, and Emmanuel Mazer. *Bayesian Programming Book*. Springer, 2006.

[6] Pierre Bessiere, Christian Laugier, and Roland Siegwart. *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*. Springer, 2008.

[7] Xavier Perrin, Ricardo Chavarriaga, Roland Siegwart, and Jose del R. Millan. Bayesian controller for a novel semi-autonomous navigation concept. *ECMR*, 2007.