OPEN ACCESS

University of BRISTOL

Leelasawassuk, T., Damen (Aldamen), D., & Mayol-Cuevas, W. (2015). Estimating Visual Attention from a Head Mounted IMU. 147-150. Paper presented at International Symposium on Wearable Computers (ISWC)., Osaka, Japan.

Peer reviewed version

Link to publication record in Explore Bristol Research
PDF-document

This is the accepted author manuscript (AAM). The final published version (version of record) is available online via ACM at http://delivery.acm.org/10.1145/2810000/2808394/p147-leelasawassuk.pdf?ip=137.222.138.48&id=2808394&acc=OPENTOC&key=BF07A2EE685417C5%2E3DCFD3605FE4B4CE%2E4D4702B0C3E38B35%2E9F04A3A78F7D3B8D&CFID=662285555&CFTOKEN=58988258&__acm__=1472654904_0ffd7ed2293980da4fc45cb3f1a140f1. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/pure/about/ebr-terms.html

Estimating Visual Attention from a Head Mounted IMU

Teesid Leelasawassuk
University of Bristol
csztl@bristol.ac.uk

Dima Damen
University of Bristol
damen@cs.bris.ac.uk

Walterio W Mayol-Cuevas
University of Bristol
wmayol@cs.bristol.ac.uk

ABSTRACT

This paper concerns with the evaluation of methods for the estimation of both temporal and spatial *visual* attention using a head-worn inertial measurement unit (IMU). Aimed at tasks where there is a wearer-object interaction, we estimate the *when* and the *where* the wearer is interested in. We evaluate various methods on a new egocentric dataset from 8 volunteers and compare our results with those achievable with a commercial gaze tracker used as ground-truth. Our approach is primarily geared for sensor-minimal EyeWear computing.

ACM Classification Keywords

H.5.2 User Interfaces: Input devices and strategies; I.5.5 Pattern Recognition: Implementation

INTRODUCTION

Wearer attention, both temporal and spatial, is linked not only to what is happening now, but importantly, helpful to anticipate what action or object interaction will be carried out next. In EyeWear computing, attention is commonly computed from tracking eyes, but this involves specialized hardware requiring precise eye-scene calibration that assumes no misalignment occurs afterwards and thus not immediately suitable or available for wearables. Current eye-gaze hardware also has limited operational conditions, does not cover the entire range of eye motions and may be affected by ambient lighting levels. An alternative proposition, explored in this paper, is to investigate how well spatial and temporal attention can be recovered from a head mounted inertial measurement unit (IMU), without requiring any inward or outward looking visual sensor. This has advantages for EyeWear design simplicity, operational robustness but also energy usage. Furthermore, IMUs are already present in many existing EyeWear devices e.g. Google Glass and Epson's Moveiro. Overall, predicting when and where the wearer is paying attention is important for deciding when to (or not to) collect activity data or provide assistance.

RELATED WORK

Few attempts in the literature aim to **predict** wearable gaze (head and or eye-gaze), but two recent works use an outward-looking head mounted camera (HMC). In [5], eye-gaze is estimated from combining a pre-computed task-specific gaze centroid and the detection of hands while users do tasks such as preparing food while seated. Their results do show how head motion, as observed from the HMC, can be a valuable cue as well as how the concentration of gaze fixations is spatially tight when compared to non wearable gaze recordings. They report an average angular error (AAE) of 8.35 degrees. Requiring hand cues however, constraints the prediction horizon, as attention (via eye-gaze) does precede action by several milliseconds [3]. The work in [7], estimates eye-gaze from multiple bottom-up sources including pixel colour, intensity, gradient and user motion. An AAE is not provided and instead performance is reported using true and false positive classification according to a saliency map generated by their method. This map often covers large portions of the image as it is mainly derived from image cues.

Figure 1: Google Glass and Commercial wearable gaze tracker (for ground-truth) bundle (L) and experimental environment setup (R).

Estimating eye-gaze without a visual sensor has been attempted as a tool for interaction using EOG [1, 2], where rates of positive detections for eye gestures or activity classification are reported instead of eye-gaze location accuracy.

On the other hand, usage of IMU signals for wearables has been widely studied with some key literature covered in e.g. [6]. We note that head motion has been somewhat less studied as a source for wearable input and, to our knowledge, the simultaneous estimation of the temporal and the spatial *visual* attention as we do here from wearable IMUs, has not been presented before. Compared to [5, 7], our spatial estimation approach is much simpler as it does not require detection of bottom up image features or detection of hands, it has been evaluated in more than a couple of activities, and we compare our results with ground-truth obtained with a commercial gaze tracker.

SENSORS, STATE AND MODELS

We use a Google Glass Explorer Edition 2.0 as it neatly features visual and IMU sensors. Furthermore, we attach it to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ISWC '15, September 7–11, 2015, Osaka, Japan.
Copyright 2015 © ACM 978-1-4503-3578-2/15/09...$15.00.
http://dx.doi.org/10.1145/2802083.2808394

an ASL Mobile Eye gaze tracker to serve as ground-truth (Fig 1L).

We recover a head motion vector H_t at time t, where $H_t = (h_t, h_{t-1}, \ldots, h_{t-N})$ is the sequence of N prior head rotational and translational motions of the t^{th} frame. Each sample h_j represents the motion of the j^{th} frame relative to its consecutive $j^{th} - 1$ frame and h is defined as $h = [\theta \quad a \quad \omega]^\top$, where $\theta = [\theta_x \; \theta_y \; \theta_z]^\top$ is the relative head orientation (Δyaw, Δpitch, and Δroll), a is the relative acceleration, and ω is the relative head angular velocity.

Temporal attention threshold

We are interested in wearer-object interaction attention periods during no translation motion, that is, when the user is stationed to interact with objects including standing or seated, e.g. assembling things. We define our temporal attention as

$$\mathrm{T}_{\text{attention}} = \begin{cases} \textit{attending}, & \text{if } a \leqslant \tau \text{ and } \omega \leqslant \nu \\ \textit{in motion}, & \text{Otherwise} \end{cases}, \quad (1)$$

where τ is the relative acceleration threshold and ν is the relative head angular velocity threshold for identifying whether the user is sufficiently stationary or not. The spatial attention estimator, explained next, only starts its prediction process when in the *attending* phase.

Spatial attention model

Our spatial attention estimation is dependant on an initialisation stage as we will cover later. Lets assume that we know the user's spatial attention of the previous frame and it is located at pixel $c^{t-1} = [x^{t-1}, y^{t-1}]$, we need to estimate the change of user's attention at the current frame Δc^t regarding the head motion H^t. Therefore, the spatial attention of current frame is $c^t = c^{t-1} + \Delta c^t$ and Δc^t is obtained from

$$\Delta c^t = \Phi(F), \quad (2)$$

where Φ is an attention estimation model and $F = (H^t, c^{t-1})$ is a head motion vector of the current frame H^t accompanying with its previous spatial attention position $c^{(t-1)}$. The attention estimator can be considered as a conditional expectation given a set of samples $\mathbf{H}$ and its known previous spatial attention $\mathbf{c}^{t-1}$ and their corresponding response values $\Delta\mathbf{c}$ as $\Phi = \mathrm{E}(\Delta\mathbf{c}|\mathbf{F})$.

We choose the Nadaraya-Watson estimator to estimate the change of user's attention locations in the scene. Given a sample of bivariate data $\{x_1, y_1\}, \ldots, \{x_n, y_n\}$, a so called non-parametric estimator of $\tilde{y}'(x')$ is written as follows

$$\tilde{y}'(x') = \frac{\Sigma_{i=1}^{n} K_\alpha(x_i - x')y_i}{\Sigma_{i=1}^{n} K_\alpha(x_i - x')}. \quad (3)$$

Each kernel K has its centre located at a sample x_i with a bandwidth α. The observed response $\tilde{y}'$ at location x' is a locally weighted average of its neighbour. Since the estimator is a non-parametric function and approximates $\tilde{y}'$ as a locally weighted average, the estimator does not need any further parameters, and the function is assumed continuous.

To simplify, we generate two independent estimators for horizontal and vertical directions and then construct n training pairs $\left\{\Delta c_j, (H_j, c_j^{t-1})\right\}$. The head motion parameters can be randomly sampled from the user's head motion during the training stage. Therefore, the estimation of the total shift in spatial attention at the current frame Δc_k is formulated as

$$\Delta c_k = \frac{\Sigma_{j=1}^{n} d_{jk} \Delta c_j}{\Sigma_{j=1}^{n} d_{jk}}, \quad (4)$$

where d_{jk} is the kernel function and is defined as a Gaussian density function $d_{jk} = \exp\left(-\sum_{j=1}^{n} \frac{F_j - F_k}{\alpha}\right)$ and $F_j - F_k$ is the Euclidean distance between current sample k and learned example j. The bandwidth α controls the shape of the regression. Therefore, the output of sample F_k is a normalised weighted sum of its nearest neighbour samples.

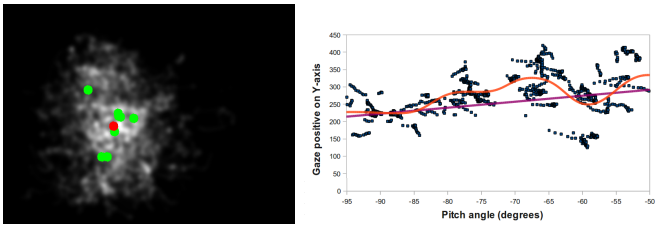

Figure 2: (L) True eye-gaze distribution for 8 users operating on multiple tasks and their centre of gaze (green dots) and overall centre of gaze (red dot). (R) Function mapping the relative head's pitch angle against gaze elevation using Kernel Regression (red) and a linear regression via RANSAC (magenta).

Spatial attention initialization using head orientation

The model Φ in Equation 2 estimates the change in spatial attention. This model requires initialisation, i.e. an estimate of c^{t-1}. To address this, we propose an initialization method as follows.

The underlying approach we take is based on two observations from analyzing eye gaze wearable video of daily living activities. The first observation is that gaze is relatively tightly concentrated on the 2D image coordinates of the HMC as can be seen in Fig 2L where gaze fixations are shown for all gaze locations for various activities. The centre of gaze distribution across all users is illustrated by a red dot and the centres' of gaze distribution for each individual are shown in green. To calculate a fixation, we use the method reported in [4], i.e. angular velocity less than 100 degrees/s. As can be seen, while fixations are not on a single location as expected, they are also not evenly spread across the field of view (and are biased towards the left as we use a HMC worn on the right side).

Our second observation is that the eye-gaze distribution, as illustrated in Fig 2L, appears to align along a vertical line and that there appears to be a consistent relationship between the head pitch angle and the location of the gaze driven by the principal location of hand activity. That is, for manipulating objects on a table, the user's gaze focuses on a lower part of the scene but not too low, while operating the upper buttons on a printer, the gaze is higher up but below its maximum possible range. This relationship is shown in Fig 2R where we

Algorithm 1: Estimating attention from head motion.

input : The user's head motion data.
output: Temporal ($T_{attention}$) and spatial (c^t) attention.

Given IMU data, construct the head motion vector H;
if $a \leqslant \tau$ *and* $\omega \leqslant \nu$ **then**
  $T_{attention}$ = stationary

while $T_{attention}$ = stationary **do**
  if *This frame is the first frame of* $T_{attention}$ **then**
    Determine the attention starting point;
  else
    Estimate Δc_k from F_k using Eq. 4;
    The current spatial attention $c^t = c^{t-1} + \Delta c_k^t$;
    Assign $c^{t-1} = c^t$ for the next coming frame;

collect the vertical head orientation against the location of fixations (black dots) and apply the Kernel regression in Equation 3 to obtain the mapping function (red line). We compare the Kernel regression to a linear regression calculated using RANSAC.

Our method thus builds on these two observations to estimate the starting point of user's attention based on the expected location of gaze, and the experimentally obtained mapping of head pitch angles relative to the ground plane and gaze location. This approach for computing the starting point for estimation is simple and fast to calculate. Algorithm 1 summarises the user's attention estimation method using the head motion cues.

EXPERIMENTS AND RESULTS

As described before, we modified an ASL's Mobile Eye XG gaze tracker by attaching a Google Glass to it as illustrated in Fig 1a. This allows us to record the user's current activities and gaze positions along with the user's head motion.

Eight volunteers were asked to wear the bundled device. Time synchronization was done by asking users to look at a stopwatch displayed on a computer screen at the start of the trial while we record the first four seconds of the videos on the ASL camera and the Google Glass camera. After this, Google Glass only recorded IMU data. Both devices operate at 30 fps and we verified the time synchronisation error is constrained to one image frame.

It should be noted that volunteers performing the activities were unaware of the variables we use so that head position or behaviour should not have been biased. The volunteers then started performing tasks directed by an investigator. The tasks included normal daily activities such as grabbing objects, opening doors, writing on a whiteboard, etc. Since any current wearable gaze tracking hardware is imperfect and fails to work in various real conditions we used a wand with a ping-pong ball attached to its tip to point at the location where we instructed the volunteers to carry the tasks (Fig 1b). The ping-pong ball thus served as an additional continuously available ground-truth for the user's spatial attention. The wand is also useful to calibrate the system when no gaze

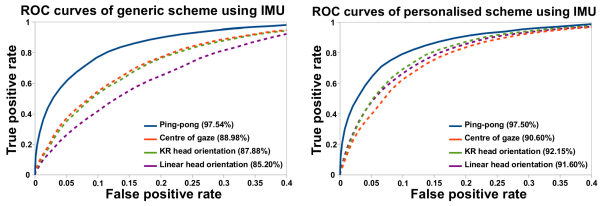

Figure 3: ROC curves and AUC scores (in brackets) for estimated spatial attention in generic (left) and personalised (right) schemes.

tracker is attached to it. Our visuo-IMU dataset is published in archival form. Please contact authors for details.

Temporal attention results

As described before, the spatial attention process starts estimating the user's attention location if $T_{attention}$ returns the stationary state. Therefore, to be able to identify the state of $T_{attention}$, we need the optimal threshold values for τ and ν. For this we used an ROC analysis. We manually labelled frames of the recorded videos indicating the temporal attention ground-truth as *stationary* or *in motion*. The threshold τ and ν were then varied and we chose the optimal threshold values from the ROC curve as $\tau = 3.0$ m/s^2 and $\nu = 0.5$ rad/s. In addition, the AUC score is 81.44% and the false positive rate and true positive rate are 0.144 and 0.79, respectively.

Spatial attention results and Discussion

We use three measures to evaluate the performance of the prediction of users spatial attention: Area Under (ROC) Curve (AUC), Average Angular Error (AAE) and attention fixation Precision and Recall results. The accompanying video illustrates the results visually. AUC measures the consistency between the predicted attention and the ground-truth. AAE measures the angular distance between the predicted spatial attention and the ping-pong ball position and the Precision and Recall measure how well the attention area overlaps with the area around the groundtruth. Recall the spatial attention is only considered when the temporal attention is detected.

We use three different starting points: the Kernel regression (KR) method, the linear regression method, and the centre of user's gaze (CoG) distribution. To determine the user's KR and linear mapping functions, the user's gaze fixations accompanying the head pitch angles are acquired from the training data when the $T_{attention}$ is identified as *stationary*. The KR and linear mapping functions are then constructed using the method described before and the CoG is the average (x,y) positions of all collected fixations. The inclusion of the ping-pong ball as starting point in the results is to have the "ideal case" baseline performance for estimating the user's spatial attention. But note that to be of more value, our reported results are not relative to this "ideal case" but instead reported in absolute terms.

We trained and tested the head motion model using the leave-one-out cross-validation method. We separated the training and testing to two schemes – generic (**G**) and personalised (**P**) testing schemes. The generic testing scheme is a test of

	Starting point mode				
	Ping-pong	Eye gaze	CoG	KR	Linear
G	8.09	11.10	12.61	12.70	13.98
P	7.17	10.51	10.76	10.33	10.18

Table 1: AAE (degrees) for three starting point methods in generic (**G**) and personalised (**P**) schemes.

Figure 4: Spatial attention results. Ground-truth (blue) and estimated (magenta) using Kernel regression.

one participant against the remaining while the personalised testing scheme is a test of one sub sequence to all of the sub sequences of that volunteer. We chose $N = 3$ and $\alpha = 0.2$ as that gave the best prediction performance. Table 1 shows the AAE results of the generic and personalised scheme and Fig 3 shows the comparison of ROC curves and AUC.

Using gaze as a starting point results in similar performance to the three IMU-based methods. Though gaze, being not very reliable to sense, had 23% of frames with missing data. The other three starting point methods have no missing data and give a similar AUC score at 88.98% for the CoG starting point, 87.88% for the KR starting point, and 85.20% for the linear regression starting point when training data from all volunteers (generic scheme). However, when considering individualised mapping functions (personalised scheme), the results improve overall as the KR head orientation achieves AUC scores of 92.15%, which is slightly better than CoG at 90.60% and the linear regression model at 91.60%. Fig 5 (left) shows why the individual mapping functions perform better for the case of KR and by association, the linear regression and CoG methods compared to the average generic function Fig 5 (right). Note that this is not a real obstacle for our approach as such individual calibration is relatively simple and requires the user to gaze at different scene objects relative to his/her location.

This evaluation indicates there is small angular error difference between using any of the initialization methods but to fully test spatial attention fixation, the AAE is insufficient as it does not encode how accurate the attention is spatially stable during a period of time. We thus used a Precision and Recall approach where we define two areas of 200×200 pixels which cover most of the objects or parts interacted with. They are centered at the ground-truth and predicted spatial attention positions. If these areas overlap using the PASCAL overlap criteria at 30%, and this overlap lasts 10 frames consecutively, a positive fixation is declared. Ten frames are akin to the 300ms used in eye-gaze for fixation determination [3]. Table 2 shows the results that indicate better performance for the personalised Kernel Regression over the other methods.

CONCLUSIONS

Estimating the where and when wearers are attending using non visual or gaze sensors is challenging but our results are

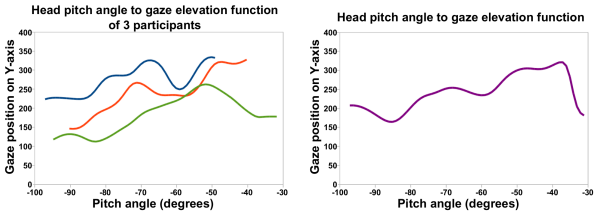

Figure 5: Function that maps head's pitch angle vs gaze elevation in 3 individuals (left) compared to the overall generic function for all individuals (right).

		Starting point mode			
		ping-pong	CoG	KR	Linear
G	Precision	1.00	0.61	0.50	0.48
	Recall	0.82	0.56	0.47	0.44
P	Precision	1.00	0.63	0.71	0.65
	Recall	0.82	0.57	0.65	0.60

Table 2: Precision and recall for spatial attention.

encouraging as using widely available IMU sensors, allows to approximate the estimation of these important tasks.

Our data-driven approach uses regression for estimating spatial attention, the *where* the wearer is looking at, and a principled ROC-evaluated approach to select thresholds for the *when* the user is stationary. The method achieves high temporal (> 0.8 TPR) and good spatial accuracy (< 10 Deg) as well as spatial attention stability ($> 5\%$ over alternative methods considered). Results are better with personalised training which is attained after a simple calibration process.

REFERENCES

1. Bulling, A., Roggen, D., and Tröster, G. Wearable EOG goggles: Seamless sensing and context-awareness in everyday environments. *Journal of Ambient Intelligence and Smart Environments 1*, 2 (May 2009), 157–171.
2. Ishimaru, S., Kunze, K., Uema, Y., Kise, K., Inami, M., and Tanaka, K. Smarter eyewear: using commercial eog glasses for activity recognition. In *UbiComp14 Adjunct* (2014).
3. Land, M., Mennie, N., and Rusted, J. The roles of vision and eye movements in the control of activities of daily living. *Perception 28*, 11 (1999), 1311–1328.
4. Leelasawassuk, T., and Mayol-Cuevas, W. W. 3D from looking. In *ISWC*, ACM Press (2013), 105.
5. Li, Y., Fathi, A., and Rehg, J. Learning to predict gaze in egocentric video. In *ICCV* (2013).
6. Ward, J. A., Lukowicz, P., Troster, G., and Starner, T. E. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE PAMI 28*, 10 (2006), 1553–1567.
7. Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A., and Hiraki, K. Attention prediction in egocentric video using motion and visual saliency. In *Adv in Image and Video Tech*. 2012.