21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Optical flow based Head Movement and Gesture Analyzer (OHMeGA)

Sujitha Martin, Cuong Tran and Mohan Trivedi
University of California San Diego
smartin@ucsd.edu, cutran@ucsd.edu, mtrivedi@ucsd.edu

Abstract

Automatically identifying and analyzing head gestures is useful in many situations like smart meeting rooms and intelligent driver assistance. In this paper, we show that head movements can be broken into its elemental forms (i.e. moving and fixation states) and combinations of these elemental forms give rise to various head gestures. Our approach which we term, Optical flow based Head Movement and Gesture Analyzer (OHMeGA), segments head gestures into moving and fixation states using optical flow tracking and intermittent head pose estimation. OHMeGA runs in real-time, is simple to implement and set up, is robust and is accurate. Furthermore, segmenting head gestures into its elemental forms gives access to higher level semantic information such as fixation time and rate of head motion. Experimental analysis shows promising results.

1. Research Motivation and Background

Humans express their state of mind through many modalities. While spoken words and written languages are powerful tools for expressing one's thoughts and intentions, in many situations complementary head movements prove to be very useful in understanding an individual's state of mind. For example, analyzing head gestures have given valuable insight into driver behavior [2] [3] [16], in meeting like scenarios [8] [13] [1], for surveillance [11], in human-machine interaction [14], and in the study of public displays [12], [5].

While the most accurate means of analyzing head gestures is using head pose estimation, however, it can be computationally intensive for the task at hand. We introduce a hybrid algorithm called OHMeGA, which segments head gesture sequences into understandable and logical elemental states (i.e. head movements and fixations) using Lucas-Kanade optical flow algorithm [6] and uses head pose estimation occasionally to remove any uncertainty in the spatial location of fixation. Detecting a sequence of elemental states not only identifies the type of head gesture, but also gives higher level semantic information such as fixation time and rate of change of head motion associated with the head gesture. As we will show in the following section, this approach is simple to implement and set up, and runs in real-time.

2. OHMeGA: Concept and Algorithm

Head gestures are composed of elemental states such as head movements and head fixations. These elemental states, when represented in a state machine as shown in Fig. 1 give real-time information on spatially where the head was previously fixated and in what direction the head is currently moving.

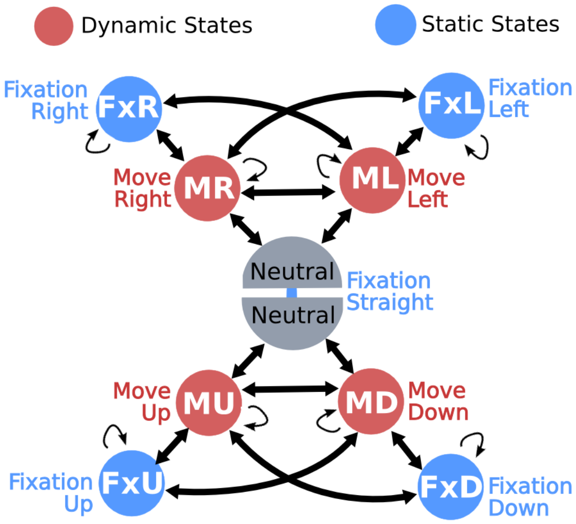

Figure 1. State diagram of OHMeGA for head gesture analysis.

The OHMeGA state diagram in Fig. 1 has two major parts: one part to represent horizontal movements and another part to represent vertical movements in the image plane. Using a frontal facing camera, head movements in the pitch and yaw rotation angles can be bro-

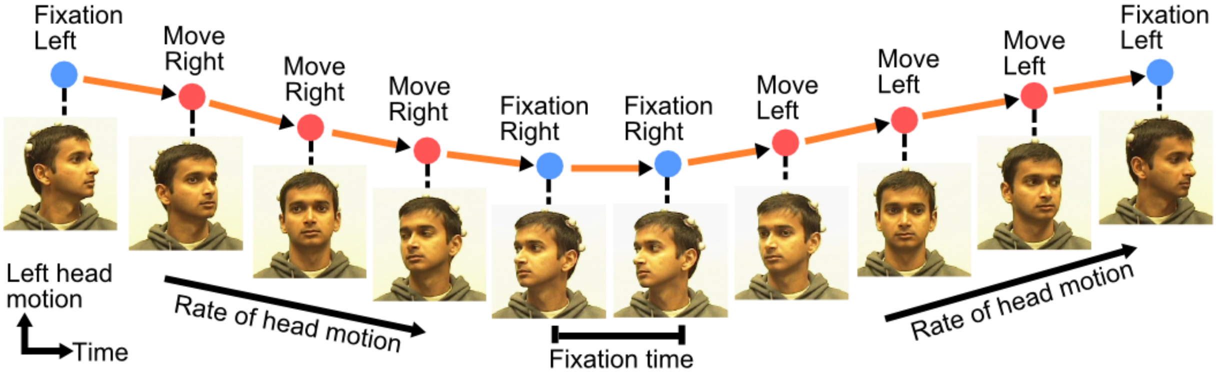

Figure 2. Typical head movements, gestures, fixations and temporal dynamics, which need to be analyzed.

ken into horizontal and vertical movements in the image plane. States associated with horizontal movements are move right (MR), move left (ML), fixation right (FxR), and fixation left (FxL) and states associated with vertical movements are move up (MU), move down (MD), fixation up (FxU) and fixation down (FxD). The remaining two neutral states are used to indicate that in its respective direction the fixation is neutral, and thus when both horizontal and vertical movements are in neutral state it indicates fixation straight. Fig. 2 shows an example of representing a typical head gesture in elemental states.

To begin, the OHMeGA is initialized into a fixation state using head pose estimation, which can be easily determined within few frames using many number of techniques [9], [17] [4] [10]. Once initialized, ideally, transitioning between any of the states in OHMeGA is a matter of knowing whether there is any motion in the image plane of a frontal facing camera and the direction of motion. Ideal in the sense that frame rate goes to infinity, there is no noise in camera sensors, and all motions detected by optical flow in the image plane are only due to head movements.

2.1 Noise and Other Practical Matters

In the real world, however, we have limited frame rates, there are noise due to camera vibrations, and motions detected by optical flow can be anything from movements in the background to body parts other than the head (i.e when moving hands to rub one's eyes as shown in Fig. 3b). While background noise can be mitigated by using techniques like face detection to only consider optical flow vectors in the head region and by using experimentally determined threshold on what is considered motion, noise in head motion due to occluding objects is an open-ended problem and will be further studied in future works.

Assuming optical flow vectors are computed only in the face region and thus represent true head motion, it is important to note that there is a correspondence between direction of flow vector and region of face for any given head motion as shown in Fig. 3a,c. Such situations arise due to rotational movements as oppose to in-plane translational movements as occurring in the prediction of driver behavior using foot gesture analysis [15]. One solution is to carefully select interest point regions over which to compute global flow vector.

Furthermore, out of plane head rotation induces certain amount of motion in both the horizontal and vertical direction of the image plane (Fig. 3d). By observation we know that head motion in the yaw rotation angle induces a strong horizontal motion in the image plane and to some extent the same applies for head motion in the pitch rotation angle with vertical motion in the image plane. In our current implementation of OHMeGA, we correspond horizontal and vertical motion in the image plane to head motions in the yaw and pitch rotation angles, respectively.

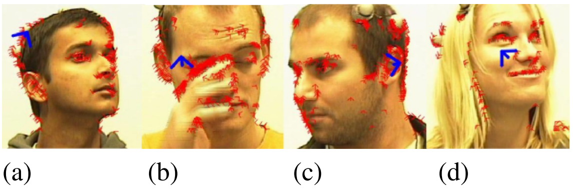
(a) (b) (c) (d)

Figure 3. Practical matters regarding optical flow vectors: (a) & (c) show correspondence between direction of flow vector and region of face for a given head motion (b) shows false head motion due to hand motion and (d) shows head motions in the pitch inducing both horizontal and vertical motions in the image plane.

Lastly, since motion vectors in the image plane can-

Table 1. Horizontal and vertical head movements detected by OHMeGA. Rows represent the predicted states and columns represent the ground truth.

	MR	**ML**	**FxR**	**FxL**
MR	0.85	0.01	0.08	0.01
ML	0	0.85	0.01	0.08
FxR	0.02	0.10	0.88	0.09
FxL	0.13	0.04	0.03	0.83
Samples	994	982	355	320

	MU	**MD**	**FxU**	**FxD**
MU	0.90	0	0.06	0.02
MD	0	0.86	0	0.10
FxU	0.04	0.10	0.94	0
FxD	0.07	0.03	0	0.88
Samples	1229	1236	455	425

Table 2. Fixations, rate of motion and rate of change in motion detected by OHMeGA.

	Mean	Var
Overlap in fixation duration	0.91	0.02
Error in rate of motion	0.003	0.036
Error in rate of change in motion	0.001	0.029

not directly correspond to the amount of spatial movements in the real world, some ambiguity in the correct fixation state may arise. For example, if a driver is initially looking straight then turns his head rightward, pauses for a bit, and then continues to turn rightward, there is no ambiguity that the resulting state is fixation right. If, however, the driver now turns leftward and pauses, there is an uncertainty whether the driver is fixated straight or fixated left. For such cases, OHMeGA uses head pose estimation to remove any uncertainty.

2.2 Motion Analysis using Optical Flow

To estimate head motion used for state transitions, optical flow vectors are computed over sparse interest points in the image using the Lucas-Kanade algorithm [6]. Interest points can be easily found using methods like Harris corner detection and Fröstner corner detection. The global flow vector is then computed based on majority vote from the computed optical flow vectors. Lastly, the global flow vectors are averaged over a few frames to mitigate sporadic noise. A sample at the output of optical flow tracking as applied to a video sequence containing head motions with selected image frames are shown in Figure 4.

3. Experimental Evaluation

The data for the evaluation of our approach is taken from a subset of the dataset used for the evaluation of HyHOPE [9]. Evaluation is done on three subjects, who performed various head movements in the pitch and yaw rotation angles. Ground truth head pose, as collected from a motion capture system, is used to derive ground truth head motion in the pitch and yaw rotation angle. As mentioned earlier, in our current implementation, we correspond horizontal and vertical motion in the image plane to head motions in the yaw and pitch rotation angles, respectively. Therefore, data with head movements in the yaw (pitch) rotation angle is used to evaluate the part of OHMeGA corresponding to horizontal (vertical) motion. Furthermore, in our current implementation, we use ground truth head pose to relieve uncertainty in the spatial location of fixation.

As mentioned earlier, there is a concern for computing global flow vector due to correspondence between direction of optical flow vector and region of face for a given head motion. In our current implementation, we manually annotate a generous region around the nose and use that region for computing global flow vector. In Fig. 4, the red bounding box on the images indicate typical regions used for computing said global flow vector. The confusion matrix for the evaluation of OHMeGA over horizontal and vertical head movements is given in Table. 1. Note that, no accuracy is reported for Neutral state because no fixation occurred exactly at neutral in either direction. Specific details in implementing OHMeGA, such as transitioning between states is similar to the simpler version of OHMeGA presented in [7]. In future studies, we will address the problem of accurately representing head motions in the pitch rotation angle using optical flow vectors in both the horizontal and vertical motions in the image plane.

Apart from state level classification accuracy, we also show that higher level semantic information such as fixation time and rate of head motion can be derived from elemental state sequences of a head gesture with high level of accuracy in Table 2. For comparing fixation times, we first note that our approach detected 98% of all continuous fixation time events with a duration of at least 2 frames. The detected fixation durations overlapped the corresponding ground truth fixation durations on an average of 91%.

Regarding the other semantic information, rate of head motion for predicted motion and ground truth motion was computed when both showed signs of motion in the same direction and normalized with their respec-

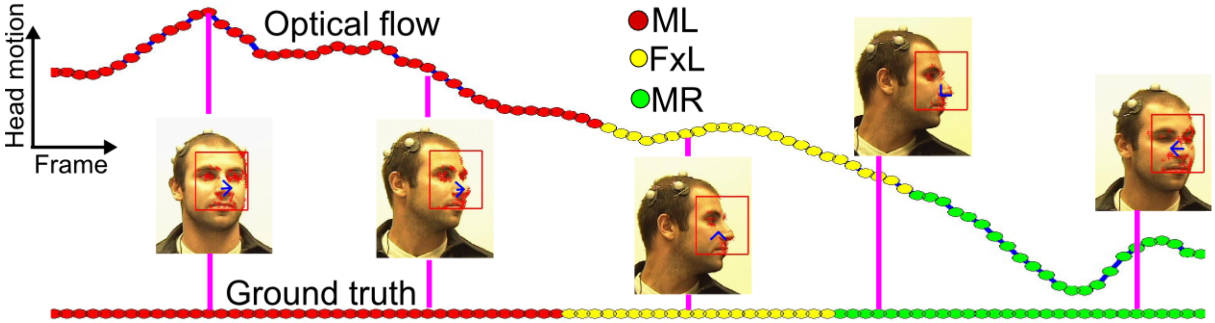

Figure 4. Annotation of how to interpret optical flow head tracking with sampled images.

tive maximum values. Rate of change in head motion was similarly calculated and the results are given in Table 2.

4. Concluding remarks

In this paper we have introduced a new, simplistic, robust and accurate way to detect and analyze head gestures. We presented our analysis in a controlled setting where subjects were asked to perform head movements in the pitch and yaw rotation angles. Experimental analysis shows promising results of an average of 88% accuracy in fixation state and an average of 87% accuracy in move state level classification. Future work will be in the direction of optimizing global optical flow vector calculation - currently optimal for in plane movements - for out of plane rotations, using feature detection more compliant with faces and applying OHMeGA to naturalistic environments.

References

[1] A. Doshi and M. Trivedi. Head and gaze dynamics in visual attention and context learning. In *CVPR Workshops 2009.*, pages 77 –84, june 2009.

[2] A. Doshi and M. Trivedi. On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *Intelligent Transportation Systems, IEEE Transactions on*, 10(3):453 –462, sept. 2009.

[3] A. Doshi and M. Trivedi. Attention estimation by simultaneous observation of viewer and view. In *CVPR Workshops 2010*, pages 21 –27, june 2010.

[4] K. Huang and M. Trivedi. Robust real-time detection, tracking, and pose estimation of faces in video streams. In *ICPR 2004.*, volume 3, pages 965 – 968 Vol.3, aug. 2004.

[5] A. Lablack and C. Djeraba. Analysis of human behaviour in front of a target scene. In *ICPR 2008.*, pages 1 –4, dec. 2008.

[6] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging Understanding Worshop*, pages 121–130, 1981.

[7] S. Martin, C. Tran, and M. M. Trivedi. Optical flow based head movement and gesture analysis in automotive environment. In *IEEE International Conference on Intelligent Transportation Systems-ITSC*, Sept. 2012.

[8] E. Murphy-Chutorian and M. Trivedi. 3d tracking and dynamic analysis of human head movements and attentional targets. In *ICDSC 2008.*, pages 1 –8, sept. 2008.

[9] E. Murphy-Chutorian and M. Trivedi. Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 512 –517, june 2008.

[10] E. Murphy-Chutorian and M. Trivedi. Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness. *Intelligent Transportation Systems, IEEE Transactions on*, 11(2):300 –311, june 2010.

[11] K. Sankaranarayanan, M.-C. Chang, and N. Krahnstoever. Tracking gaze direction from far-field surveillance cameras. In *ACV Workshop 2011*, pages 519 –526, jan. 2011.

[12] A. Sippl, C. Holzmann, D. Zachhuber, and A. Ferscha. *Real-Time Gaze Tracking for Public Displays*, volume 6439 of *Lecture Notes in Computer Science*, pages 167–176. Ambient Intelligence, 2010.

[13] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling people's focus of attention. In *Modelling People, 1999. Proceedings. IEEE International Workshop on*, pages 79 – 86, 1999.

[14] K. Toyama. "look, ma - no hands!" hands-free cursor control with real-time 3d face tracking. In *Proceedings of the 1998 workshop on Perceptive user interfaces*, pages 49–54, 1998.

[15] C. Tran, A. Doshi, and M. M. Trivedi. Modeling and prediction of driver behavior by foot gesture analysis. *Computer Vision and Image Understanding*, 116(3):435 – 445, 2012.

[16] M. Trivedi and S. Cheng. Holistic sensing and active displays for intelligent driver support systems. *Computer*, 40(5):60 –68, may 2007.

[17] J. Wu and M. M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41(3):1138 – 1158, 2008.