A Bio-inspired Motion Sensitive Model and Its Application to Estimating Human Gaze Positions under Classified Driving Conditions

Jiawei Xu¹, Seop Hyeong Park², and Xiaqing Zhang³

Abstract— Human visual attention performs the spatio-temporal regularity, an inherent regularity in dynamic vision, for the natural visual tasks. However, recent computational visual attention models still have deficiency to reflect the spatio-temporal regularity thoroughly. Motivated by this, we propose a bio-inspired motion sensitive model to estimate human gaze positions in driving. The proposed model has four key advantages. First, inspired by two types of motion sensitive neurons, this model can distinguish motion cues from different directions spatially and temporally. Second, compared to conventional deep learning based models, this model does not rely on expensive training samples with gaze annotations. Third, the proposed model is based upon low-level visual signal processing without constructing a complex deep neural network architecture, which enables this model to be implemented with low-cost hardware. Fourth, inspired by the visual pathway in drosophila motion vision, the further visual signal processing on directional and depth motion sensitive map largely enhances this model’s competence in a similar way of human gaze positions in driving. To test this proposed model, we collect a new dataset from the perspective of egocentric vision that aims to systematically evaluate the performance of computational visual attentional models under real-life human driving tasks. The video clips in the dataset are categorised into ten classified driving conditions by ranking the motion cues from the simple to complex levels. The proposed model is evaluated by using the classified driving video clips in comparison with the human baseline of the gaze positions. Experimental results demonstrate that the proposed model can effectively estimate the human gaze positions in driving and consistently outperforms traditional visual attention models as well as the deep learning based model.

Keywords—human gaze position in driving, motion sensitive neurons, biological motion vision, visual signal processing

1. Introduction

The human visual system is capable of dealing with complex dynamic scenes because it effectively selects the most relevant visual cues. Computational visual attention models that perform functions similar to human visual attention can be applied to various fields including autonomous driving systems. The driver’s visual attention during actual driving is attracted by the dynamics of the driving environment. An efficient computational visual attention model for autonomous driving system should be able to estimate driver’s gaze positions in a specific moment. To achieve this level of efficiency, it is crucial to extract the key visual information that changes constantly. However, most of the computational visual attention models [4-11] proposed so far still have deficiencies in extracting the key visual information from dynamically changing scenes. This is primarily due to the complexity of processing motion cues under rapidly changing environments.

The vision-based perception technology has improved (for review, see [1]) over the past few decades. However, due to the complex motion stimuli, not many computational visual attention models are capable of efficiently selecting significant information [2]. The main reason why it is difficult to create a computational visual attention model originated from human attention is that we cannot fully understand the complex internal processes of the human visual system. Animal's vision has evolved to accurately and quickly extract and recognize key visual cues that are necessary for its survival. Amongst numerous methods, bio-inspired neural networks and bio-inspired computer vision technology provide inexpensive and feasible approaches to visual signal processing. From this perspective, it seems more promising to build a bio-inspired computational visual attention model that imitates the visual system of well-known insects than that of humans.

Recently, there have been approaches to incorporate deep neural networks into visual attention prediction [18, 19]. While these approaches show advanced performance in benchmark testing, inevitable limitations can be summarised as follows.

1. The spatio-temporal regularity: First, they do not reflect the spatio-temporal regularity of the human visual system. Although the long short-term memory (LSTM) networks learn the spatio-temporal relationship, the human spatio-temporal regularity reflects a high sensitivity on different spatio-temporal regularities for ego-centric vision as exploited by the psychologists [48], which is the key difference with LSTM networks. Alternatively, human vision are affected by the motion cues in spatio-temporal domain while recent LSTM still have difficulties in reflecting this characteristic when processing dynamic visual stimuli. To our best knowledge, recent computational visual attention prediction methods based on deep neural networks (DNNs) do not consider spatio-temporal regularity thoroughly. The computational visual attention model should not be purely based upon the

1. J. Xu is with the School of Computing Science, Newcastle University, Newcastle-upon-Tyne, NE4 5TG, United Kingdom. Email: jxulincoln@gmail.com.
2. S. Park is with the Department of Electronic Engineering, Hallym University, 200-702, South Korea. Email: spark@hallym.ac.kr
3. X. Zhang is with Department of Computer Science, Wenzhou University, Wenzhou 325035, China. Email: xqzhang@wzu.edu.cn
salient object recognition. This is consistent with the fact that human visual attention is not necessarily attracted by 'salient' objects as reported by the earlier research [48].

2. Generality to new data: Second, deep neural networks are not able to adapt to new data that have different characteristics from those used for a long-time training by collecting a big data. This means that new training process is required whenever a new video from a new driving environment is used.

3. Interpretation: Third, the deep neural networks contain many unexplainable units whilst our bio-inspired model are built upon the biological plausibility which is interpretable for future transformations.

Visual attention directs the driver's perception to the key visual information under dynamic world, it is therefore important to imitate human visual attention for the implementation of autonomous driving.

So far, there is no consensus on how human motion visual system works. A computational human motion visual model is thus lack of theoretical guarantee. We noticed that the visual system of some insects is similar to the human visual system, but is much simpler [3]. Based on this idea, we propose a bio-inspired computational motion-sensitive model that predicts the gaze position of the human driver while watching driving videos. The proposed model has four layers that form a hierarchical structure:

1) **Retinal layer** detects depth and directional motion cues from the videos. In order to build a simplified model for this layer, we modified the LGMD neuron and DSN models of locust visual system in [33-35].

2) **Laminar layer** in drosophila motion vision [43, 61] is implemented by the non-subsampled contourlet transform (NSCT) and used to decompose depth and directional motion maps into a low-frequency and a high-frequency map. Among these two maps, the low-frequency map is further processed because it is similar to the source image. This is plausible because experimental results in [50] shows that asymmetrical lateral processing of drosophila motion detectors by means of a low-pass filter improves velocity estimation.

3) **Medulla layer** decomposes the low-frequency map into the bright and dark channels. A morphological toggle operator (MTO) is used for the segmentation of the low-frequency map. This is inspired by the neurons in the medulla layer of drosophila motion vision. These neurons are responsive to the motion of brightness increments (ON path way) and the motion of decrements (OFF pathway) when perceiving different amounts of motion cues. We classified the driving conditions into the ten most common situations depending on the complexity of motion stimuli perceived by the driver. We used a genetic algorithm to optimize the coefficients of the bright and dark channels to mimic the adaptive brightness and darkness control in the medulla layer.

4) **Lobula complex** corresponds to the summary of human gaze position that targets the key visual information on the videos. This layer is inspired by the lobula complex in drosophila motion vision. Inspired by this bio-inspired characteristics, also considering NSCT are robust to the feature variance in both spatial and frequency domain, we then fuse the high-frequency map and the optimized low-frequency map using inverse NSCT. The predicted gaze position is centered at the peak value of the fused map based on the winner-take-all theory. The predicted gaze positions were extensively compared with human gaze positions obtained from the eye-tracking device.

The main contribution of the proposed model is summarised as follows.

First, unlike traditional optical flow methods, the proposed model is biologically inspired rather than empirical statistical models, such as DNNs. The model processes the motion cues by encoding the responses of two visual neurons: LGMD and DSNs. As shown in [22] and [25], they are capable of capturing depth and directional movements effectively.

Second, we refine the depth and directional motion cues by mimicking the visual signal pathways from drosophila motion vision [43, 61]. This crucial bio-inspired post-processing has proven this model to be effective through our extensive experiments.

The rest of this paper is organized as follows. Section II introduces related research in the literature and our contributions. The proposed bio-inspired motion sensitive model is explained in detail in section III. Section IV presents experimental results comparing the human baselines with the gaze positions from other visual attention models. Section V presents the conclusion of this paper.

2. Related Research

2.1 Visual Attention Models

Previous studies have made extensive contributions to the field of visual attention modeling by using top-down and/or bottom-up attention. Bottom-up attention is inspired by the human perception which selects salient features from natural scenes ([4]-[11]), while top-down attention uses task-relevant cognitive processing influenced by prior knowledge or active visual searching ([12]-[14]). Recently, a comprehensive survey [15] has reviewed state-of-the-art visual attention models. The earliest bottom-up visual attention model is based on low-level features inspired by the simplified human visual system [4]. Some models take the dynamics of a visual scene into consideration. For example, the attention based on information maximization (AIM) model [5] uses Shannon's self-information measure to calculate the saliency of image regions. The incremental coding length (ICL) approach [6] measures the respective entropy gain of each feature to maximize the entropy of the sample visual features in both dynamic and static scenes. Goferman et al. [7] proposed a context-aware saliency model which detects the feature contrast from color and intensity. In [8], Han et al. proposed an unsupervised method for extracting attentional objects from color images. Later, an unsupervised object extraction method was proposed to separate the target from the background [9], which was inspired by the contour delineation and texture discrimination based on the oriented edge features. In recent years, as sparse representation techniques have been widely used, an object-based visual saliency detection method has been proposed to distinguish a region of interest (ROI) from natural images [10]. In [11], Jia and Han fused the prior cues and the object appearance to detect salient objects. Moreover, several top-down visual attention models have been proposed in recent years. The top-down visual attention model proposed in [12] uses a dynamic Bayesian network (DBN) to infer the probability of the visually interesting objects and their corresponding spatial locations. A top-down and bottom-up integrated model was proposed to actively search for interesting objects for real-life
visual searching tasks [13]. The top-down visual saliency was learned jointly through the conditional random field with a discriminative dictionary [14]. Due to the dynamic environment, we focus on the spatio-temporal visual attention model with an emphasis on motion cues under real driving conditions.

2.2 Spatio-temporal Visual Attention Models Featured with Motion Cues

Recently, several studies have been published that integrates motion cues with spatio-temporal visual attention models. Zhai and Shah [16] proposed a spatio-temporal attention model that detects interesting objects and dynamic actions from a video sequence. Spatial and temporal saliency images are fused into one spatio-temporal saliency map controlled by motion contrast. Marat’s model [17] extracts two motion sensitive signals from the retina: information from parvocellular and magnocellular pathways. The signals are then split into elementary feature images and combined into a spatio-temporal saliency image. To the best of our knowledge, the extraction and fusion of the motion cues has not been fully integrated into the spatio-temporal visual attention models. Under some classified driving conditions, however, the researchers have demonstrated that the priority lies in motion cues rather than in salient objects in case of the egocentric movement [20]. Accordingly, this paper addresses egocentric vision to process key motion cues in a dynamic environment.

2.3 Motion Sensitive Neurons - LGMD and DSNs

It has been revealed that there are many different types of visual interneurons in animals’ visual systems that prefers to specific motion direction. This suggests that temporal cues are well coded in the visual processing systems of both vertebrates and invertebrates [21]. In the case of invertebrates, LGMD (lobular giant movement detector) neurons in locusts have been identified to show specific sensitivity to looming objects in a direct collision course [22, 23]. LGMD is sensitive to depth movement, which effectively responds to the objects approaching on a direct collision course, but produces little or no response to receding objects.

DSNs (directional sensitive neurons) are another type of specified visual neurons that have preference to the directional motion cues ([24]-[26]). Both LGMD and DSNs were discovered in animals ([27]-[29]), insects ([22], [23] and [30]) and vertebrates ([27]-[29] and [31]). DSNs have been proved to be involved in detecting looming objects as well [32]. Recent studies indicate that whole field direction selective neurons could be organized for extracting translational visual cues [33], and can be organized with a special framework for collision detection [34]. These directional and depth motion sensitive neurons that capture the rich spatio-temporal events in a dynamic world have been discovered in the visual pathways of many animal species. This may indicate that the high-level visual attention systems coping with dynamic visual events could be supported by these early visual processing mechanisms. Recent studies have shown that by integrating LGMD neurons and DSNs into a simple robotic platform one can effectively detect depth and lateral movements [35].

2.4 Non-subsampled Contourlet Transform

NSCT, which enhances and fuses the visual cues extracted from video frames, has been studied in previous studies ([36] and [37], for recent review, see [38]). NSCT is a multi-scale and multi-directional decomposition transform that consists of two filter banks: non-subsampled pyramid filter bank (NSPFB) and non-subsampled directional filter bank (NSDFB) [36]. First, NSPFB implements shift-invariant filtering to decompose the source image into low-pass and high-pass images. NSPFB catches the point discontinuities. Second, NSDFB decomposes the high-frequency images obtained by NSPFB to produce directional sub-images of the same size as the high-frequency images. The NSCT decomposition has two advantages. One is that it inherits multiscale and multidirectional properties from contourlet [37], and separates local characteristics of the source image in spatial and frequency domains. The other advantage is that it can provide a shift-invariant image representation using the non-subsampled filter banks. Therefore, by using NSCT, one can avoid the problems of spectrum aliasing and Gibbs phenomenon, and is able to extract the useful geometrical features from the source image. For each frame, NSCT decomposes the source image into low-frequency and high-frequency subband images. Notably, NSCT is inspired by the drosophila motion vision [43, 61], i.e., the pathway in which the motion cues are split into high-frequency and low-frequency visual signals and conveyed to the lamina area (the first neuropile layer).

2.5 Decomposition of the Low-frequency Image into Bright and Dark Channel

Recent findings reveal that the low-frequency coefficient image contains similar characteristics to the source image, and the gray value of the image changes smoothly in the spatial domain [41]. An MTO was chosen to decompose the low-frequency image into bright and dark channels. An MTO can suppress noise and pseudo-edges as reported in [42]. The decomposed bright and dark channels will be further adjusted to imitate drosophila motion vision [50] in its egocentric movement.

2.6 The Biological Sensations from Drosophila Motion Vision

In real life, drosophila motion vision manifests an accurate and rapid natural instinct after perceiving the motion cues in the retina. Based on the anatomical and neurological findings, the first neuropile layer, named as the lamina layer, enables the drosophila to distinguish the signals at different frequencies. The second neuropile layer, the medulla layer, controls the brightness increments and decrements when perceiving different motion cues. Specifically, the ON and OFF pathways in this layer is crucial for drosophila’s innate bio-function necessary for egocentric flying. The terminal layer, the lobula complex, sums up the motion cues and controls its flying course [43, 61]. Recently, a computational structure for perceiving motion has been studied by modeling drosophila motion vision for contrast changes [44].

3. The Proposed Bio-inspired Motion Sensitive Model

Inspired by the biological discoveries [43, 44, 61] mentioned in the previous section, we propose a new bio-inspired motion sensitive model. In this model, we employ NSCT to mimic the lamina layer’s ability of frequency discrimination. We, then, apply MTO and genetic algorithm to optimize the bright and dark
channel control mechanism by ranking the visual cues perceived by the driver’s retina under classified driving conditions. The optimization mimics the medulla layer to control the bright and dark channels. Finally, we use inverse NSCT to fuse the maps and then we use the winner-take-all neural network theory [49, 59] to select the peak response of each fused map, which is the estimated gaze position. The real gaze positions from the human observers are summarized as the human baseline for comparison.

The ultimate aim of the proposed computational motion sensitive model is to predict the driver’s gaze positions while driving in real life. Figure 1 illustrates the block diagram of the proposed model. The model consists of four layers connected in series. The first layer extracts holistic (depth and directional) movement by encoding the responses of two visual neurons: LGMD and DSNs. The second layer decomposes holistic movement to high-frequency and low-frequency maps using NSCT. The third layer automatically adjusts the bright and dark channels using MTO and genetic algorithm. The last layer uses the reverse NCST to select the peak response that corresponds to the expected gaze position. The proposed model mimics the hierarchical processing pathway starting from the retinal layer (depth and directional movement extraction), to the laminar layer (frequency decomposition), and to the medulla layer (ON and OFF pathway to the lobula complex summation and output). The computed gaze position of each frame is then compared with the corresponding driver’s gaze position. Figure 2 shows an example of intermediate signals and the results generated by the proposed motion-sensitive model under single way single car driving condition.

![Fig. 1](image_url) The block diagram of the proposed bio-inspired motion sensitive model. The video stream flows into the retinal layer.

![Fig. 2](image_url) An example of the intermediate maps and the results generated by the proposed bio-inspired motion sensitive model to estimate human driver’s gaze position. The human driver is under single way single car driving condition.

### 3.1 Motion Sensitive Model of LGMD and DSNs

Previous studies have shown that LGMD is effective in detecting approaching events/objects [34] and that DSNs are sensitive to directional movements [35]. Unlike previous LGMD and DSNs integrated feed forward inhibition (FFI) to avoid false collision alert for the application of robotics, we remove this FFI module but biologically process the motion sensitive maps to estimate the eye gazing locations for building up the human-like visual attention model. Besides, considering the individual functions of these two types of neurons, we combined the LGMD model and DSNs model to construct a visual attention model that responds to objects moving in any directions.

#### 1) LGMD

The LGMD neural network model in Figure 3 is a slightly modified version of the model described in [39]. LGMD neural networks are known to be sensitive to the looming stimuli moving along the depth direction. This sensitivity to the looming stimuli was successfully applied to the robotics to alert the upcoming collision [35]. Recent finding has shown that LGMD enhances the intensity of the response towards fast-moving nearby objects [40]. There are four cell groups in the LGMD neural networks: photoreceptor P, excitatory E, inhibitory I and summing S. In this paper, we use the output of the S layer, i.e. the LGMD map, as a motion output map in the depth direction.
P layer: The input neuropile layer of LGMD neural networks is the photoreceptor P cells, which are mapped to the pixels of the input image in a one-to-one relationship. Let \( I_f(x,y) \) be the luminance of the f-th input frame captured by each photoreceptor cells, where \( x \) and \( y \) denote the coordinates of the cell, respectively. Each photoreceptor can be modelled as an infinite impulse response (IIR) filter. The output signal of the cell in this layer, \( P_f(x,y) \), is defined as:

\[
P_f(x,y) = \sum_{i,j} p_i P_{f-1}(x+i, y+j)w_f(i,j),
\]

where \( p_i \) is the change in luminance that corresponds with the pixel \((x,y)\) at frame \(f\). \( L_f \) and \( L_{f-1} \) are the luminance, subscript \(f\) denotes the current frame and \(f-1\) denotes the previous frame, and the persistence coefficient \( p_i \) is defined by \( p_i = (1 + e^{\mu i})^{-1} \), where \( \mu \in (-\infty, +\infty) \). We set \( \mu \) to 0 for simplicity.

\[ \text{Fig. 3. The neuromorphic structure of LGMD. Each of the P cells detects the change in luminance and transfers the signal to the counterpart cells in the E and I layer through the neural network (marked in blue). Lateral inhibition is indicated by the grey dashed lines, where the signal is delayed by one frame. Lateral inhibition in the I layer spreads to neighboring cells in the S layer without preferred direction. Excitation pathway is indicated with the yellow lines, where the signal is transferred without delay. If the sum of all excitation signals in the S layer is greater than the decay threshold, membrane potential is created on the LGMD cell membrane. This potential is transmitted to an LGMD cell (represented by the red lines).} \]

I and E layer: The output signal of each P cell is fed to both I and E type cells in the next layer. The E type cells pass on excitation to their retinotopical counterparts in the I layer with one frame delay. The intensity of the inhibition on one cell in the S layer is expressed as a weighted sum of the inhibition signals coming from the neighboring cells in the I layer. This can be expressed as:

\[
I_f(x,y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P_{f-1}(x+i, y+j)w_f(i,j),
\]

where \( I_f(x,y) \) is the inhibition at \((x,y)\) in the f-th frame, and \( w_f(i,j) \) is the local inhibition weight. \( mn \) is the inhibition area size. \( m \) and \( n \) are set to zeros for LGMD, \( w_f(i,j) \) ranges from 0.125 to 0.25.

S layer: The output signals of the E cells and the I cells are summed by the S cells using the following equation:

\[
S_f(x,y) = \left| P_f(x,y) - I_f(x,y) \right| W_f,
\]

where \( W_f \) is an adaptable global inhibition weight. Each cell in S layer delivers this signal to the next layer only if \( S_f(x,y) \) is greater than or equal to a threshold. This can be expressed as follows:

\[
f(x,y) = \begin{cases} 
S_f(x,y) & \text{if } S_f(x,y) \geq T_f, \\
0 & \text{if } S_f(x,y) < T_f,
\end{cases}
\]

where \( T_f \) is the decay threshold setting to 6 for LGMD structure. We set the threshold to a relatively low value because motion features may be a key visual cue that attracts the driver’s attention even at a low driving speed.

The LGMD cell: The membrane potential of the LGMD cell, \( U_f \), is the sum of all excitation in \( S \) cells as described by the following equation:

\[
U_f = \sum_{k=1}^{k} \sum_{l=1}^{l} |f(x,y)|
\]

where \( k \) and \( l \) are each set to 80 and 100, respectively. Then, \( U_f \) is converted to an output spike through the following sigmoid function:

\[
f = \frac{1}{1 + e^{-U_f/n_{cell}}}
\]

where \( n_{cell} \) is the total number of the cells in S layer, which is set to \( n_{cell} = kl = 8000 \) in this paper. It is noted that \( 0.5 \leq f \leq 1 \) since \( U_f \geq 0 \).

The motion sensitive map along the depth direction is spiky by the LGMD cell. A certain number of sequential spikes will trigger the LGMD cell. LGMD cells will then perceive motion cues in depth perception. We thus denote the summed spikes generated by the perceived depth motion sensitive map as \( f \) in equation (6).

2) DSNs

Figure 4 is an illustration of the neuromorphic structure of DSNs. DSNs perceive the directional motion cues in dynamic visual scenes [33]. Similar to LGMD model input layer, DSNs model have the same photoreceptor \( P \). DSNs also consist of the excitatory cells \( E \) and inhibitory cells \( I \). Four groups of summing cells \( (SL, SR, SU \text{ and } SD) \) can be found in the middle layer. Four-directional selective cells \( (L, R, U \text{ and } D) \) are the output layer in our DSN-based neural network. We only demonstrate the left inhibitory summing cell \( SL \) and left inhibitory cell \( L \) as an example.

Fig. 4. Illustration of the neuromorphic structure of DSNs. Five groups of cells form the DSN-based neural network. It consists of photoreceptor cells \( P \) sharing with LGMD based neural network, excitatory and inhibitory cells \( (E \& I) \), left inhibitory summing cells \( SL \), which only sum up the excitation and the leftward inhibition to the final L-cell. In a similar manner, the excitation and the rightward, upward and downward inhibition signals are gathered by SR, SU and SD cells, respectively. This figure shows an example of DSNs neuromorphic structure in the left inhibited cell L.

SL layer: The inhibition from an I cell is passed to its retinotopical counterpart’s neighboring cells in the next layer asymmetrically from one to eight cells away but with one image frame delay. The gathered strength of inhibition to the SL cell at the position \((x,y)\) in this layer, \( I_f(x,y) \), can be expressed as:

\[
I_f(x,y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P_{f-1}(x+i, y+j)w_f(i,j),
\]
where \( w^l(i,j) \) is the local inhibition weight set to 1.5 based on the earlier study on a DSNs model [35]. In equation (8), inhibitory signals can spread to left, right, up and down directions, but spreads mainly to the left. The signals that spread to the left are stronger than those spread to the right since \( m_L \) is greater than \( m_R \). It is unnecessary to use other three directional inhibitory signals which are smaller than the left directional inhibitory signals since the outputs from DSNs are integrated to extract the visual motion cues. To reduce unnecessary computation, we set \( n_L \) to 0 and \( m_L \) to 8. By doing so, we make the propagation of the inhibitory signals to the left direction. This results in the directional selectivity with a single non-preferred direction [35]. With strong inhibitory signals coming from the right side, the excitation caused by left translating movements is decreased or even cancelled [35]. The excitatory flow summed in the \( SL \) cell can be expressed as

\[
E^{SL}(x, y) = |P_L(x, y)| - |I_L(x, y)|W_L^L, \tag{8}
\]

where \( W_L^L \) indicates the global inhibition weight, which is set to 1.5 based upon the earlier DSNs model [35].

**L cell:** The excitatory signal greater than or equal to the threshold \( T_L \) in the \( SL \) cells is transferred to the left inhibitory cell \( L \). This summing operation is expressed as:

\[
\bar{E}^{SL}(x, y) = \begin{cases} 
E^{SL}(x, y) & \text{if } E^{SL}(x, y) \geq T_L, \\
0 & \text{if } E^{SL}(x, y) < T_L,
\end{cases} \tag{9}
\]

where threshold \( T_L \) is set to 12. The membrane potential of the left inhibitory cell \( L \) is described as:

\[
\lambda_L = \sum_{k=1}^{k} \sum_{l=1}^{l} |\bar{E}^{SL}(x, y)|, \tag{10}
\]

where \( k \) and \( l \) indicate the number of cells along the \( x \) and \( y \) direction, which are identical to those of LGMD structure. \( k \) and \( l \) are set to 80 and 100, respectively. The membrane potential of the \( L \) cell is transformed using a sigmoid function,

\[
g_L = \frac{1}{1 + e^{-\lambda_L/n_L}}, \tag{11}
\]

where \( n_L \) is the total number of cells in the \( SL \) layer which is set to \( n_L = kl = 8100 \) in this paper. It is noted that 0.5 ≤ \( g_L \) ≤ 1 since \( U_f \) ≥ 0. Like the spatiotemporal mechanism in the DSNs, the leftward excitatory signals would be decreased or cancelled by the leftward inhibitory signals in the \( L \) channel of the DSN neural network.

Similar to the map of the leftward inhibited DSN, the maps of the rightward inhibited DSN, upward inhibited DSN, and downward inhibited DSN, are further integrated to generate the fused motion direction map. Finally, the directional sensitive map is fused by L-DSN, R-DSN, U-DSN and D-DSN as follows:

\[
g = \frac{1}{4}(g^L + g^R + g^U + g^D), \tag{12}
\]

The fused map from four directional motion cues is denoted as \( g \). Together with the depth sensitive map \( f \), we further propose a bio-inspired motion cues processing method under various driving situations.

It is also worthy to note that the parameters, such as \( w_f(i,j) \) and \( T_f \) for LGMD, \( w_L(i,j) \), \( W_L^L \) and \( T_L \) for DSNs neuromorphic structure, only account for the output of depth and directional sensitive maps. For the final visual attention map, we mimic the visual signal processing mechanism of drosophila motion vision and exploit a genetic algorithm to automatically optimize the low-frequency maps, which is an adaptive procedure without training process. We will explain this method in the next subsection.

### 3.2 Motion Cues Refinement and Fusion Inspired by Drosophila Motion Vision

Drosophila motion vision is a low-cost yet effective system that inspired us to mimic its biological structure [43, 50, 61]. In biological sense, [61] comprehensively introduced the laminar layer, medulla layer and lobula complex from drosophila motion vision. First, the laminar is able to discern the high- and low-frequency visual signals, which motivated us to use the NSCT. Second, the ON and OFF pathways in the medulla layer enable the brightness increment and brightness decrement when perceiving the dynamic motion cues [43], which inspired us to apply the MTO. Third, the lobula complex sums up the sub-channel visual signals and decides its flying course. From this, we chose to use the inverse NSCT again and select the peak response according to the winner-take-all theory [49, 59]. Nevertheless, the latest research [50] showed that motion detectors in the ON and OFF pathways are asymmetrically processed, which enhances velocity estimation. Inspired by this, we refine the low-frequency signal rather than using symmetric processing on both pathways. We then use a genetic algorithm to automatically adjust the egocentric motion cues perceived by the driver’s retina.

Dozens of map fusion methods have been proposed during recent years (for reviews, see [38]). In this study, the motion cues are considered to be the key factors for the proposed model performance under different driving tasks. We propose an NSCT-based MTO fusion method to improve the motion image fusion. Specifically, the overall fusion strategy consists of the following main steps: (1) The MTO in the NSCT domain is applied to effectively extract bright and dark motion cues from the low-level frequency channel; (2) A genetic algorithm is adopted to optimize the fused low-frequency coefficients; and (3) The fused image is reconstructed by the inverse NSCT. The morphological sequential toggle operator based on the NSCT decomposition is to extract the bright and dark cues from the motion maps at low-frequency domain. These extracted motion cues from the bright and dark channels should be adjusted based on different speed conditions. For example, in parking situation, the signal from bright (ON) channel should be enlarged because the motion cues are sparse whilst signal from dark (OFF) channel should be reduced because the motion cues is sparse because of the low speed. On the other hand, in driving situation, the motion cues are abundant because of the high speed, thus the signal from bright (ON) channel should be reduced and the signal from dark (OFF) channel should be enlarged. Further, the genetic algorithm proves its efficiency to optimize the low-frequency coefficients before the final fusion step. After classifying the driving situations into 10 categories based on the strength of the received visual stimuli, we used a genetic algorithm to optimize the weight coefficients of the bright and dark motion maps. This way, the proposed visual attention model can effectively select most informative visual cues. To measure the performance of the proposed model, a reference gaze position was selected by an eye-tracking device. The proposed motion sensitive model is as follows.

#### 1) NSCT on the input LGMD and DSNs maps
As mentioned in Section II, NSCT has proved its efficiency to decompose the input images (here, LGMD and DSNs) into high- and low-frequency images. Biologically, the laminar layer in drosophila motion vision can effectively distinguish different types of visual signals, such as high- and low-frequency visual stimuli in the spatio-temporal domain [43, 61]. We followed the released Matlab toolbox [36] to decompose the input LGMD and DSNs maps to the high and low frequency maps with the original mathematical operations as defined in [36].

The low-frequency subband image has similar characteristics to those of the source image and its gray values change smoothly [41]. Inspired by the asymmetrical characteristics of drosophila motion detectors in fly visual circuitry [50], we adopt a strategy to process the low frequency map rather than processing both low- and high-frequency maps. This strategy is intended to mimic the asymmetric ON and OFF motion detectors of drosophila vision. This asymmetric processing reduces computational burden [41].

After decomposing the depth motion sensitive map \( f \) and the directional motion sensitive map \( g \) by using the NSCT, we can obtain the low-frequency subband coefficients denoted as \( \{C_L/C^r_L \} \) and the high-frequency subband coefficients denoted as \( \{C_I/C^r_I \} \). The low-frequency subband coefficients \( \{C_L/C^r_L \} \) correspond to the low-frequency LGMD and DSNs maps, can be denoted by \( f_L \) and \( g_L \), respectively. The next step uses the MTO to mimic the asymmetrical drosophila motion vision is explained below.

2) Morphological toggle operator in low-frequency maps

MTOs have been widely used in image processing. The dilation and erosion operation on \( f_L \) with a structuring element \( B(u,v) \) are defined as

\[
 f_L \bigoplus B = \max_{u,v} (f_L(x - u, y - v) + B(u,v)),
\]

and

\[
 f_L \bigodot B = \min_{u,v} (f_L(x + u, y + v) - B(u,v)),
\]

respectively.

On the basis of dilation and erosion, the opening and closing operators on \( f \) with the structuring element \( B(u,v) \) can be expressed as

\[
 f_L \circ B = (f_L \bigodot B) \bigoplus B,
\]

and

\[
 f_L \bullet B = (f_L \bigoplus B) \bigodot B,
\]

respectively.

Applying the opening and closing operations, the white and black top-hat transforms of \( f_L \) with a structuring element \( B(u,v) \) can be rewritten as

\[
 WTH(f_L) = f_L - f_L \circ B,
\]

and

\[
 BTH(f_L) = f_L \bigodot B - f_L ,
\]

respectively.

The white and black image regions can be effectively extracted by the top-hat transform. Then, the top-hat based contrast operator can be defined as:

\[
 THO_B = f_L + WTH(f_L) - BTH(f_L),
\]

where subscript \( B \) in \( THO_B \) is a ‘disk-shaped’ structuring element with a 5-pixel diameter is used for the MTO in our model.

On the basis of multi-scale dilation and erosion, the toggle contrast operator, which is based on the primitives and defined rules, can be defined as

\[
 TO_B = \begin{cases} 
 f_L \bigoplus B, & \text{if } f \bigoplus B - f < f - f \bigodot B \\
 f_L \bigodot B, & \text{if } f \bigoplus B - f > f - f \bigodot B \\
 f_L, & \text{elsewhere}
\end{cases}
\]

Then, the MTO in [42] can be defined as

\[
 MTO_B = TO_B(THO_B) .
\]

The above MTO is also used in the low-frequency DSNs map \( g_L \) in the same manner.

3) Extraction of bright and dark motion cues from low-frequency channel

As indicated in Figure 2, the proposed model further decomposes low-frequency signals into bright and dark channels to mimic the brightness and darkness adjustment of drosophila motion vision ON and OFF pathways.

Let \( BIF_{L,B}^f \) and \( DIF_{L,B}^f \) be bright and dark image features from motion cues that are decomposed from a low-frequency channel, respectively. Then, this decomposition process can be described as follows. For the low-frequency coefficients of the LGMD depth map \( f_L \), the bright motion cues are obtained by comparing the gray values of the \( MTO_B \) and \( BIF_{L,B}^f \) of each pixel. The bright motion cues are identified by a situation in which \( MTO_B \) has a gray value greater than \( C_L^f \). In the same manner, the dark motion cues are identified by comparing the gray values of \( MTO_B \) and \( DIF_{L,B}^f \) of each pixel. The bright and dark motion cues are defined as follows:

\[
 BIF_{L,B}^f = MTO_B - C_L^f ,
\]

and

\[
 DIF_{L,b}^f = C_L^f - MTO_B ,
\]

respectively.

Similar to the MTO operation of the LGMD map, the bright and dark motion cues extracted from the low-frequency coefficients of the DSNs map \( g_L \) are defined as follows:

\[
 BIF_{L,B}^g = MTO_B - C_L^g ,
\]

and

\[
 DIF_{L,b}^g = C_L^g - MTO_B ,
\]

respectively.

The fusion of the bright and dark motion cues of the LGMD and DSNs maps is performed in two steps. The first step is to fuse the bright and dark motion cues extracted from the LGMD and DSNs low-frequency maps. The fused bright motion map can be written as follows:

\[
 FBIF_{L,b} = \begin{cases} 
 BIF_{L,B}^f & \text{if } BIF_{L,B}^f \geq BIF_{L,B}^g \\
 BIF_{L,B}^g & \text{elsewhere}
\end{cases}
\]
The fused dark motion map extracted from LGMD and DSNs in the low-frequency domain is obtained as follows:

\[
F_{LB}^{DIF} = \begin{cases} 
DIF_{LB}^f & \text{if } DIF_{LB}^f \geq DIF_{LB}^g \\
DIF_{LB}^g & \text{elsewhere}
\end{cases}
\]  

(27)

In the second step, we further optimize the low-frequency fusion map based on the fused dark and bright motion map in equation (27), which is inspired by the medulla layer of drosophila motion vision.

4) Optimization of the low-frequency coefficients using a genetic algorithm

In drosophila motion vision, the brightness increment and decrement are controlled by the parallel ON and OFF pathways in the medulla layer. This biological visual function endows accurate motion sensitivity under natural egocentric movement [43, 61]. Here, the ON and OFF channels indicate brightness increment and brightness decrement (i.e., darkness increment) for drosophila motion vision, respectively. To automatically control the brightness and darkness adjustment, we optimize the coefficients of both channels as described below.

Fused low-frequency coefficients are calculated by combining the low-frequency LGMD map \( C_L^f \), the low-frequency DSNs map \( C_L^g \) and the above fused bright and dark maps in the low-frequency domain. Then, the objective function of the fused low-frequency map can be formulated as

\[
C_L^{f} = (C_L^f + C_L^g) / 2 + \alpha \cdot F_{LB}^{BIF} - \beta \cdot F_{LB}^{DIF}
\]

(28)

where \( \alpha \) and \( \beta \) represent the weight parameters of the bright and dark motion cues, respectively, and \( \alpha, \beta \in [0, 1] \). As stated in related research [43, 61], drosophila motion vision actively adjusts brightness increment and decrement for its egocentric vision. We then apply a genetic algorithm to optimize the low-frequency weight parameters \( \alpha \) and \( \beta \). Genetic algorithms are adaptive heuristic search algorithms based on evolutionary ideas of natural selection and genetics [45]. Recently, genetic algorithms are also applied to generate experimental visual stimuli for the purpose of fast searching [46]. As stated earlier, inspired by the motion adaption in biological vision, we classify the driving conditions into 10 levels, from sparse to abundant perceived motion cues. For each driving condition, we set its egocentric speed as the initial individuals in the genetic algorithm to find the optimal parameters for the objective function as described in Table I.

<table>
<thead>
<tr>
<th>Table I</th>
<th>PSEUDO CODE OF GENETIC ALGORITHM FOR OPTIMIZING ( \alpha ) AND ( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEED_RANGE = [0, 200]: % Input speed range for each driving condition</td>
<td></td>
</tr>
<tr>
<td>INITIAL_STATE = 10: % Classified driving conditions</td>
<td></td>
</tr>
<tr>
<td>Nvar = 2: % Number of optimized coefficients ( \alpha ) and ( \beta )</td>
<td></td>
</tr>
<tr>
<td>GGAP = 0.9: % Generation gap</td>
<td></td>
</tr>
<tr>
<td>XR = 0.7: % Crossover rate</td>
<td></td>
</tr>
<tr>
<td>MUTR = 0.05: % Mutation rate</td>
<td></td>
</tr>
<tr>
<td>MAXGEN = 100: % No. of generations</td>
<td></td>
</tr>
<tr>
<td>Initialize population</td>
<td></td>
</tr>
<tr>
<td>Begin generational loop ( while Gen &lt; MAXGEN )</td>
<td></td>
</tr>
<tr>
<td>Assign fitness values to entire population</td>
<td></td>
</tr>
</tbody>
</table>

The optimal weight parameters \( \alpha \) and \( \beta \) for each classified condition are given in Table II. The table shows that in the case of condition where motion cues are sparse, for example, parking or traffic light condition, \( \alpha \) is relatively large while \( \beta \) is relatively small. On the other hand, in the case of condition where motion cues are abundant, \( \alpha \) is relatively small while \( \beta \) is relatively large. These values are in line with the biological mechanism of the drosophila motion vision. The bright motion map requires enhancement when motion cues are relatively sparse, which corresponds to enhanced signals from bright channel and decreased signals from dark channel.

<table>
<thead>
<tr>
<th>Table II</th>
<th>( \alpha ) AND ( \beta ) OPTIMIZED BY GENETIC ALGORITHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified Driving Conditions</td>
<td>Input Video</td>
</tr>
<tr>
<td>1. Parking</td>
<td>PIC0004</td>
</tr>
<tr>
<td></td>
<td>PIC0670</td>
</tr>
<tr>
<td></td>
<td>PICT0023</td>
</tr>
<tr>
<td>2. Single way single car</td>
<td>PIC0014</td>
</tr>
<tr>
<td></td>
<td>PIC0023</td>
</tr>
<tr>
<td></td>
<td>PICT0545</td>
</tr>
<tr>
<td>3. Single way pedestrian</td>
<td>PIC0002</td>
</tr>
<tr>
<td></td>
<td>PICT0014</td>
</tr>
<tr>
<td></td>
<td>PICT0021</td>
</tr>
<tr>
<td>4. Single way traffic follow</td>
<td>PIC0008</td>
</tr>
<tr>
<td></td>
<td>PIC0025</td>
</tr>
<tr>
<td></td>
<td>PIC0743</td>
</tr>
<tr>
<td>5. Single way traffic opposite</td>
<td>PIC0021</td>
</tr>
<tr>
<td></td>
<td>PIC0098</td>
</tr>
<tr>
<td></td>
<td>PICT0325</td>
</tr>
<tr>
<td>6. Single way follow opposite</td>
<td>PIC0008</td>
</tr>
<tr>
<td></td>
<td>PICT0008</td>
</tr>
<tr>
<td></td>
<td>PICT0046</td>
</tr>
<tr>
<td>7. Motorway follow</td>
<td>PIC0641</td>
</tr>
<tr>
<td></td>
<td>PIC0657</td>
</tr>
<tr>
<td></td>
<td>PICT0657</td>
</tr>
<tr>
<td>8. Motorway follow opposite</td>
<td>PIC0638</td>
</tr>
<tr>
<td></td>
<td>PIC0870</td>
</tr>
<tr>
<td></td>
<td>PIC0870</td>
</tr>
<tr>
<td>9. Traffic light</td>
<td>PIC0097</td>
</tr>
<tr>
<td></td>
<td>PICT0097</td>
</tr>
<tr>
<td></td>
<td>PICT0670</td>
</tr>
<tr>
<td>10. Junction</td>
<td>PICT0622</td>
</tr>
<tr>
<td></td>
<td>PICT0005</td>
</tr>
<tr>
<td></td>
<td>PICT0641</td>
</tr>
</tbody>
</table>

5) Fusion of high-frequency and low-frequency motion cue maps

The lobula complex next to the medulla layer in drosophila is a terminal layer that integrates motion stimuli and selects key visual cues by using winner-take-all neural network theory [49, 59]. As illustrated in the lobula complex block in Figure 2, the final fusion map \( F \) is reconstructed by the fused high frequency \( C_h^f \) and the fused optimized low frequency \( C_l^f \) using the reversed NCST. The coefficients \( C_h^f \) of the final fused map \( F \) can be expressed as
\[ C^F = \left( C_{\text{hit}}^F + C_{\text{miss}}^F \right) / 2. \]  

The human fovea is responsible for providing sharp central vision. Fixation is an important role within the visual information processing chain. This paper aims to apply the visual attention model to perception-oriented autonomous driving. The most salient visual information should be processed at a specific moment. Winner-take-all theory is adopted to find key information that attracts to human attention in each frame. The peak value of the fused map \( F \) is represented by the calculated gaze positions (eye fixations) in this paper. In the next subsection, we introduce the similarity scores to compare the calculated gaze positions with the human ground truth defined in [48].

### 4. Experiments and Performance Evaluation

This section presents ground truth from humans’ eye gazing locations and the similarity measurement, then evaluates the performance of the proposed model using the basic motion stimulus types of real-life driving videos. The performance of the proposed visual attention system is compared with those of the following seven state-of-the-art visual attention models: Itti’s bottom-up model for rapid scene analysis [4], saliency based on information maximization [5], saliency based on spectral residual [6], visual attention in spatiotemporal cues [17], deep visual attention prediction [18], top-down visual attention model using dynamic Bayesian network (DBN) [12] and attentional-awareness saliency model [60].

#### 4.1 Ground Truth of Human Baseline and Similarity Measurement

We adopted a frame-based similarity measure to quantify the differences between the computed gaze positions with the gaze positions of the human observers. As described in early research [48], for each frame, the gaze positions of 35 participants were collected using the ViSaGe Eye Tracking Toolbox and their mean value was used as a ground truth gaze position value of the human eye. We used ‘leave-one-out’ protocol to calculate the similarity score among the participants (~0.83 in our case, which is superior to those of previous computational studies). Such score system has commonly been adopted for computational modelling research [51, 55].

The closer the calculated gaze positions are to the ground truth gaze positions, the less the difference will be. The similarity score \( S \), which is a measure of the reliability of the calculated gaze positions, is defined as follows.

First, given a video sequence consisting of \( Z \) frames, let \( \phi_{k,f} = (x_{k,f}, y_{k,f}) \) be the coordinate of the measured gaze position of the \( k \)-th observer in the \( f \)-th frame with \( 1 \leq k \leq K \) and \( 1 \leq f \leq Z \). 35 observers participated in the experiment, so \( K \) is set to 35. Let \( \bar{\phi}_f = (\bar{x}_f, \bar{y}_f) \) be the coordinate of the ground truth gaze position in the \( f \)-th frame.

Then \( \bar{\phi}_f \) is defined as the centroid of the gaze positions of the observers, which is expressed as

\[ \bar{\phi}_f = \frac{1}{K} \sum_{k=1}^{K} \phi_{k,f} = \frac{1}{K} \left( \sum_{k=1}^{K} x_{k,f}, \sum_{k=1}^{K} y_{k,f} \right). \]

All-except-one method is used to compute the between-observer agreement for each gaze position. For the \( f \)-th frame, the similarity scores of the \( k \)-th observer can be written as follows:

\[ s_{k,f} = \frac{\bar{d}_{k,f}}{d_{k,f} + \bar{d}_{k,f}}, \]

where

\[ d_{k,f} = |\bar{\phi}_f - \phi_{k,f}|. \]

The similarity score for the given video clip with \( Z \) frames is

\[ s_k = \frac{1}{Z} \sum_{f=1}^{Z} s_{k,f}, \]

and the average human baseline similarity score for each video is expressed by

\[ S_{\text{human}} = \frac{1}{K} \sum_{k=1}^{K} s_k. \]

When we calculate the similarity score of another visual model for the \( f \)-th frame, we consider the visual model [4-6, 12-13, 17] as one of the observers. Then, the similarity score of the \( f \)-th frame of a visual attention model \( m \), \( S_{m,f} \), is expressed as

\[ S_{m,f} = \frac{\bar{d}_{m,f}}{d_{m,f} + \bar{d}_{m,f}}, \]

where

\[ d_{m,f} = |\phi_{m,f} - \phi_{m,f}|. \]

In the same manner as in equation (35), the similarity score of a visual attention model \( m \), \( S_{m,f} \), for the given video clip with \( Z \) frames is calculated as follows:

\[ S_m = \frac{1}{Z} \sum_{f=1}^{Z} S_{m,f}. \]

The similarity score of the proposed visual attention model and the other models to be compared are calculated in this way. In the performance comparison experiment, video clips classified into 10 conditions were used and 3 video clips were used for each classified driving condition. It is noted that \( 0 < S_{m,f} \leq 1 \), where 0 and 1 represent no similarity at all and perfect match, respectively. A detailed comparison of the resulting performance across models is given in Table III.
Normalized Scanpath Saliency (NSS) and Area under Curve (AUC) or Receiver Operating Characteristics (ROC) are alternative measures popularly used in the literature to compare the predictive power of visual attention models [51]. However, NSS is the average of the response values of human gaze position without considering the spatial phase [52], and the computation of AUC or ROC is more suitable for a static image analysis rather than dynamic video analysis (e.g., unable to capture dynamic changes between frames, see [53] and [54]). In this paper, similarity score is used as a measure to compare the performance of the proposed model and recent state-of-the-art spatio-temporal visual attention models. For each test video clip, estimated gaze position was computed by each of these models using the online available codes with default settings provided by the developers of the models.

4.2 Experiments Setup

There are several benchmark databases that have been released for the purpose of experimental evaluation of the visual attention models’ performance [55-58]. However, few of these benchmark databases are designed to test the performance of computational visual attention models on classified driving conditions. Considering the fact that multiple motion stimuli are a key factor in building the visual attention model, we classified the total 30 video clips in the database into 10 different driving conditions. For each driving condition, we selected three most representative video clips with 3-seconds length. Thus, there are two kinds of video sources used in evaluating the performance of visual attention models, and our driving video database was uploaded for public use.

The video clips used in the experiment were captured at a resolution of 960 by 800 pixels at 30 frames per second on the Panasonic camcorder HX-DC3. The camcorder was fixed on a tripod to ensure image quality to capture the front visual field of the driver viewing perspective.

All experiments were carried out on a Windows 7 platform with a PC (CPU: Intel® Core™2 Quad Q8400 2.66 GHz, RAM: 4GB). Recent state-of-the-art visual attention models are also compared with the proposed model in terms of the predicted gaze position on this driving videos.

As shown in Figure 5, the main goal of our proposed model is to minimize the error between the ground truth gaze positions of drivers and the key visual information, which are the estimated gaze position estimated by the proposed algorithm. We define the error of the visual attention model $m$, $E_m$, as

$$E_m = s_{human} - s_m$$ (39)

where $s_{human}$ and $s_m$ are defined in equation (35) and (38), respectively. By adopting the bio-inspired methods, such as using morphological toggle operator (MTO) to mimic the ON and OFF pathways’ decomposition in medulla layer, and genetic algorithm (GA) to adjust the coefficients of ON and OFF pathways in drosophila motion vision [43, 61], the proposed model minimizes the errors defined in equation (39).

In Figure 6, the gaze positions calculated by the proposed model are sequentially listed by 10 classified driving conditions: Parking, Single way single car, Single way pedestrian, Single way traffic follow, Single way traffic opposite, Single way follow opposite, Motorway follow, Motorway follow opposite, Traffic light and Junction from top to bottom. Each row represents one frame arbitrarily selected in the video of each driving condition. The five images from left to right in each row represent the original image, LGMD depth map, DSNs direction map, fused low-frequency map and the visual attention map with the overlapping the calculated gaze positions, respectively.

To further validate the effectiveness of the medulla layer of our model (the third block in Figure 1), which mimics the ON and OFF pathways in the medulla layer, we remove MTO and GA separately, as shown in Table III. This is to compare the performance of our model with that of other state-of-the-art spatio-temporal visual attention models. The models include Itti’s saliency model [4], Bruce’s information maximization [5], spectral residue saliency approach [6], visual attention in spatiotemporal cues [17], deep visual attention prediction [18], and top-down visual attention model [12]. It is apparent that our approach yields a higher similarity score than other models. In addition, the model without MTO or GA yields lower similarity score than the proposed model which has MTO and GA in it.

We further compare the computational time of the models to evaluate their suitability for real-time perception-oriented autonomous driving. Table IV shows that the computational time of the proposed model is less than that of [5, 6, 12] and [18], but greater than that of [4] and [17]. However, it should be noted that the source codes of [17] are written in C++.

4.3 Discussions and Improvements

Real-life driving conditions stimulate different types of motion cues, including global optical flow that is disrupted by movement of independently moving objects. In fact, motion stimuli are very complex visual signals that is defined by various visual cues (luminance, color, contrast, texture, etc) and can be categorized into several different subtypes (the first order motion, the second order motion, and optical flow). Such complexity might be part of the reasons why motion has not been successfully integrated into the saliency map computation like other visual signals. Since motion cues are of great importance in spatio-temporal visual attention models for perception-oriented autonomous driving, we need to find a way to incorporate motion cues into visual attention model.

---

1 Visual Stimuli in Driving: https://sites.google.com/view/xu-jiawei/database
1. Parking (PICT0023)

2. Single way single car (PICT0545)

3. Single way pedestrian (PICT0014)

4. Single way traffic follow (PICT0025)

5. Single way traffic opposite (PICT0025)

6. Single way follow opposite (PICT008)

7. Motorway follow (PICT0641)

8. Motorway follow opposite (PICT0870)

9. Traffic light (PICT0670)

10. Junction (PICT0641)
Our proposed model uses two bio-inspired neural networks to extract depth and directional motion cues, and then mimics drosophila motion vision to refine most significant motion cues. It is notable that the proposed method yields the smallest error values. We believe that this bio-inspired model contains significance compared to traditional optical flow-based methods. Another advantage of the proposed model is that it requires relatively low hardware implementation cost for perception-oriented autonomous driving.

5. Conclusion

We proposed a novel motion sensitive model by mimicking biological processing of visual cues to estimate human gaze positions under real-life driving conditions. The first stage effectively extracts depth and directional motion cues by encoding depth and directional sensitive neurons – the LGMD neuron and the DSNs of insect vision system. The second stage imitates the visual signal pathways inspired by the drosophila...
motion vision. The simulation results show that the proposed model can effectively select key visual events under real-life driving conditions. The simulation results also show that this model yields smallest estimation errors among different models. As the first biologically plausible motion sensitive model to imitate human visual attention under real-life driving conditions, we believe that this model provides a solid foundation for perception-oriented autonomous driving.

This bio-inspired model addresses spatio-temporal regularity into the complex artificial vision system. It is also notable that this biologically inspired model, unlike deep neural network approaches, does not require additional training process whenever a new task is given. In this sense, this model is task-independent, which makes our model suitable for implementation in vision-based robots.

Acknowledgments

We thank Prof. Ling Shao in Inception Institute of Artificial Intelligence (IIAI) and Dr. Yang Long in Newcastle University for their insightful suggestions to improve the paper quality. The intermediate step without FPGA acceleration (Xilinx Zynq-7000 boards) of this bio-inspired model is uploaded.

References


---

*https://www.youtube.com/watch?v=DHbON4gVqss*


Jiawei Xu is a research associate in Newcastle University, UK in 2017. He was a visiting postdoc of the National Institutes of Health (NIH), USA in 2016. He received the Ph.D. (2012-2015) degree with the topic of Human Visual Attention in Driving and its Bio-inspired Computational Modeling, in University of Lincoln, UK. He obtained a Master’s degree (2009-2011) in Electronic Engineering, Hallym University, South Korea after passing the Korean language test in 2008. He received a Bachelor degree (2003-2007) in Automotive Engineering, Shanghai University of Engineering Science and Technology, Shanghai, China.

Seop Hyeong Park received the B.S., M.A., and Ph.D. degrees from Seoul National University, Korea, in 1984, 1986, and 1990, respectively, all in electrical engineering. From 1990 to 1992, he was with the HDTV Development Center, Korean Academy of Industrial Technology, where he worked on the design and implementation of an HDTV decoder. From 1992 to 1998, he was with Korea Telecom, where he worked on digital video compression, multimedia service management software, video on demand and video conferencing service over ATM networks. In 1993, he was a visiting researcher at the NTT Human Interface Laboratory, Yokosuka, Japan, where he worked on post-processing of compressed HDTV video signals. He joined Hallym University, Gangwon-do, Korea, in 1998, where he is currently a Professor in the Department of Electronic Engineering. In 2004, he was a visiting scholar at the University of California, Santa Barbara. From 2006 to 2009, he was the dean of College of Information and Electronics, Hallym University. In 2011, He was a visiting professor at Mongolia International University. His research interests are signal processing including speech, video and ultrasound signals, multimedia communication, and machine learning.

Xiaoqin Zhang received the B.Sc. degree in electronic information science and technology from Central South University, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China, in 2010. He is currently a Professor with Wenzhou University, China. He has authored or co-authored over 80 papers in international and national journals and international conferences. His research interests are in pattern recognition, computer vision, and machine learning.