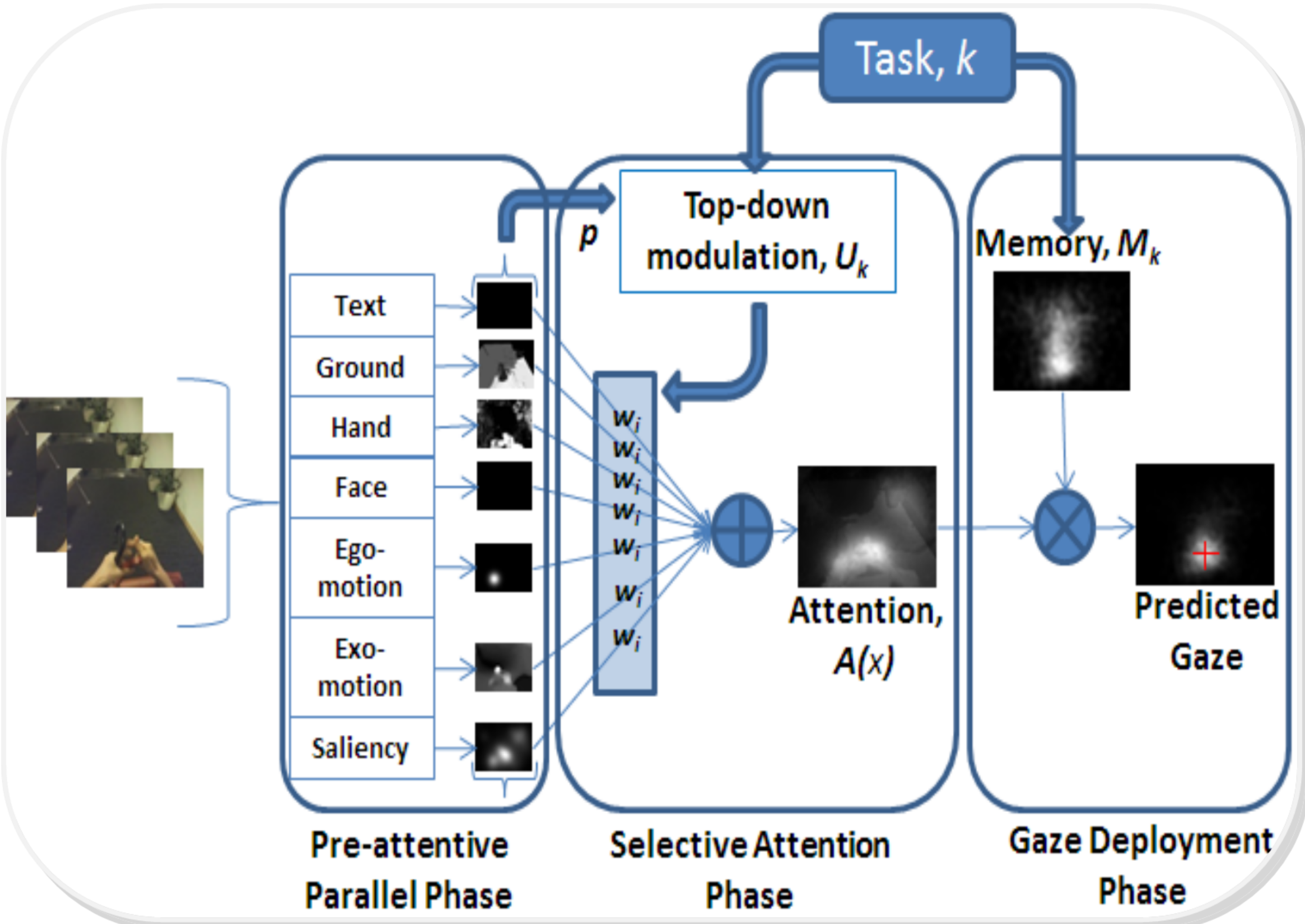
Multi-layer Linear Model For Top-down Modulation Of Visual Attention In Natural Egocentric Vision

KT Ma , Liyuan Li, Peilun Dai, Joo-Hwee Lim (I^2R), Chengyao Shen (*NUS*), Qi Zhao (*UMN*)

Abstract

Top-down attention plays an important role in guidance of human attention in real-world scenarios, but less efforts in computational modeing of visual attention has been put on it. Inspired by the mechanisms of top-down attention in human visual perception, we propose a multi-layer linear model of top-down attention to modulate bottom-up saliency maps actively.

Proposed architecture



1. **Pre-attentive Parallel Phase**: Multiple low-level features (saliency, motion etc.) and mid-level objects (text, face etc.) were processed to generate several physical salience maps.
2. **Selective Attention Phase**: The individual saliency were combined with weights computed from top-down modulation of different tasks and the standard deviation of each map to obtain an integrated attention map.
3. **Gaze Deployment Phase**: The integrated attention map and the selection history in long-term memory of each task are associated to deploy the gaze.

Pre-attentive Parallel Phase

Text: "Class-Specific Extremal Regions for Scene Text Detection" proposed by Luks Neumann and Jiri Matas is used.

Ground: Geometric Context algorithm developed by Hoiem et al. detects the ground plane and generate a saliency map.

Hand: Hand detection algorithm for ego-centric videos by Cheng Li and Kris Kitani generates a saliency map on hands.

Face: Haar feature-based cascade classifiers proposed by Viola and Jones is employed.

Ego-motion: An average global motion vector is computed along the boundaries of the Large Displacement Optical Flow (LDOF) flow field. This motion vector is used to build an ego-motion saliency map.

Exo-motion: First, LDOF is applied to compute the flow field of two consecutive frames. Then, by subtracting global motion vector from the flow field, the absolute values of the remaining components are normalized as the exo-motion saliency map.

Saliency: GVBS is employed for low-level saliency

Selective Attention Phase

Fused attention map is the weighted sum of the PPP maps.

$$A(x) = \sum_{i=1}^{L} w_i A_i(x)$$

For different top-down task U_k, w_i are different.

$$w_i = \sum_{j=1}^{L} u_j^i p_j + u_0^i = u_i^T p$$

$$where$$

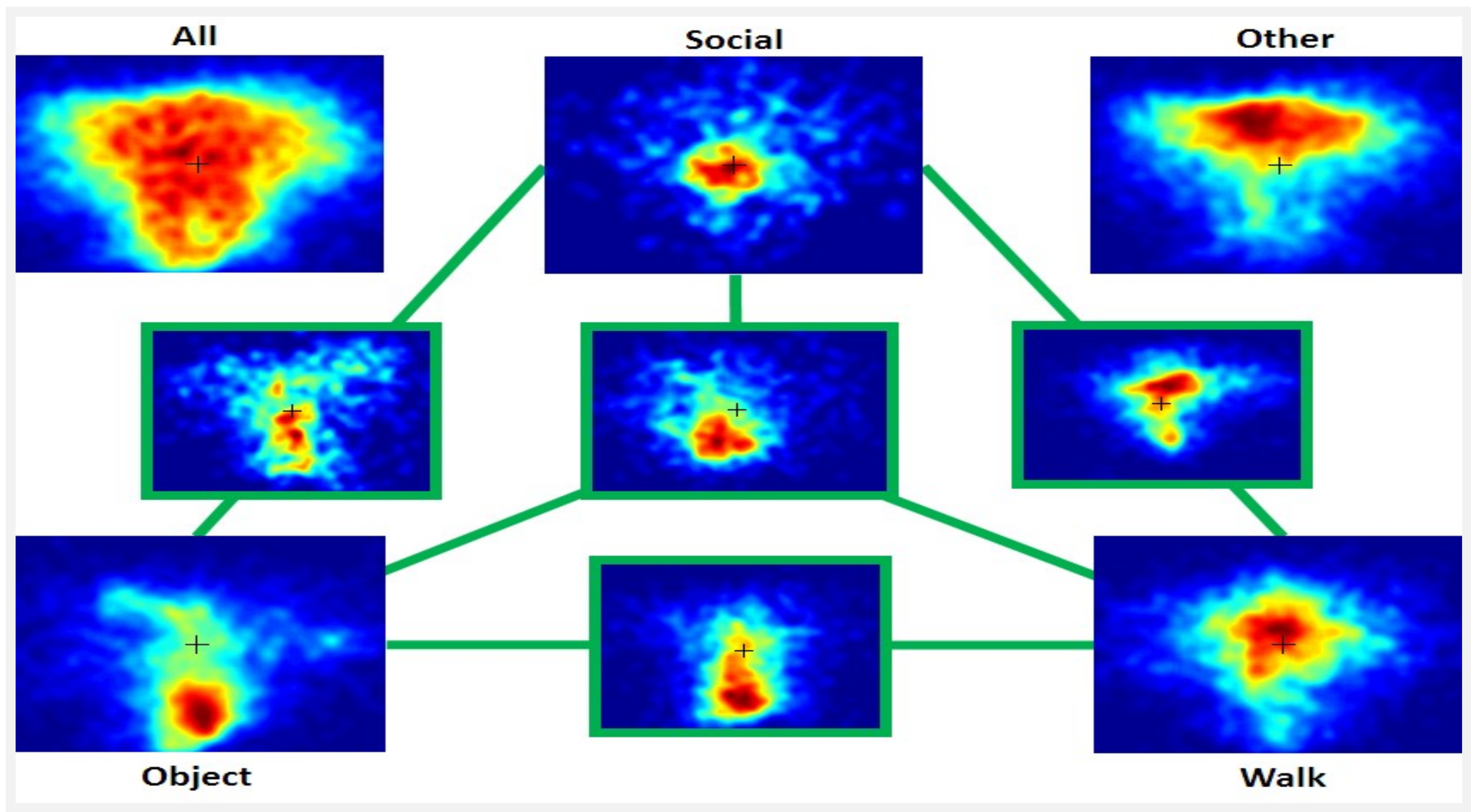
$$p_j = Dev(A_j(x))$$

With matrix representation, we solve for U_k *with the Least Square Regression as*

$$U_k = WP^T(PP^T + \alpha I)$$

where I is the identity matrix and α is the regulation parameter.

Gaze Deployment Phase



Experimental Results

Models	AAE (Smaller better)	AUC (Bigger better)
Boolean Map Saliency	17.8	0.620
Graph-Based Visual Saliency	15.6	0.642
ITTI/Koch	16.9	0.626
SALICON	15.6	0.653
Normalized Sum	16.2	0.593
Normalized Max	28.7	0.577
Linear Regression	16.3	0.593
kNN	16.7	0.512
Centre bias	12.8	0.509
Our Model	**12.3**	**0.677**

Acknowledgement

This work is supported by the Reverse Engineering Visual Intelligence for cognitiVe Enhancement (REVIVE) programme funded by Joint Council Office of A*STAR (Grant: 1335h00098).