Fixation Bank: Learning to Reweight Fixation Candidates

Jiaping Zhao, Christian Siagian, Laurent Itti
Department of Computer Science, University of Southern California.

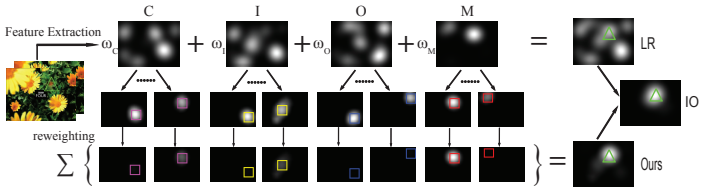

Figure 1: Location-dependent weighting: in the case of Linear Regression (LR), every pixel in a feature map receives the same weight $(\omega_{\mathcal{C}}, \omega_{\mathcal{I}}, \omega_{\mathcal{O}}, \omega_{\mathcal{M}})$. In contrast, our algorithm decomposes each map into up to $\mathcal{N}$ blobs (see ..... markings) and weights the contribution of each blob in a location-dependent manner according to the fixation bank. Our final output is a weighted sum of all blobs. Green triangle indicates the peak location in the human $\mathcal{IO}$ map.

Predicting where humans will fixate in a scene has many practical applications. Biologically-inspired saliency models decompose visual stimuli into feature maps across multiple scales, and then integrate different feature channels, e.g., in a linear, MAX, or MAP. However, to date there is no universally accepted feature integration mechanism. Here, we propose a new a data-driven solution: We first build a "fixation bank" by mining training samples, which maintains the association between local patterns of activation, in 4 feature channels (color, intensity, orientation, motion) around a given location, and corresponding human fixation density at that location. During testing, we decompose feature maps into blobs, extract local activation patterns around each blob, match those patterns against the fixation bank by group lasso, and determine weights of blobs based on reconstruction errors. Our final saliency map is the weighted sum of all blobs. Our system thus incorporates some amount of spatial and featural context information into the location-dependent weighting mechanism.

Fixation Candidates Generation: We treat each ℓ_1-normalized feature map $\mathcal{F}$ as a gaze probability distribution $\mathcal{P}_{\mathcal{F}}$. By sampling sufficient random points from $\mathcal{P}_{\mathcal{F}}$ and clustering them using mean-shift, we obtain $\mathcal{K}_{\mathcal{F}}$ clusters. Each cluster is approximated by a Gaussian blob with cluster center as mean and points covariance matrix as variance. Finally each blob on feature map $\mathcal{F}$ is treated as a fixation candidate.

Feature Map Decomposition: After extracting fixation candidates $b_k, k \in \{1,2,...,\mathcal{K}_{\mathcal{F}}\}$ from feature map $\mathcal{F}, \mathcal{F} \in \{\mathcal{CIOM}\}$, we decompose raw feature maps $\mathcal{CIOM}$ according to each blob b on $\mathcal{F}$: let $\hat{S}_b^{\mathcal{F}}$ be the decomposed feature map of blob b, which is a concatenation of 4 decayed feature maps, i.e., $\hat{S}_b^{\mathcal{F}} = [d_b^{\mathcal{C}}\ d_b^{\mathcal{I}}\ d_b^{\mathcal{O}}\ d_b^{\mathcal{M}}]^T$ with

$$d_b^f = \sum_{k=1}^{\mathcal{K}_f} \omega_{bk} \cdot g_k,\ f \in \{\mathcal{CIOM}\} \tag{1}$$

d_b^f is the a decayed map of channel f, w.r.t. reference blob b from channel $\mathcal{F}$, which is sum of $\mathcal{K}_f$ decayed blobs from f. In Eq.1, g_k is the κ^{th} blob from channel f and $\omega_{bk} = exp\{-\frac{1}{2\sigma^2}((x_b - x_k)^2 + (y_b - y_k)^2)\}$ is its weight w.r.t. reference blob b, where (x_k, y_k) and (x_b, y_b) are image plane coordinates of target blob g_k and reference blob b respectively, and σ controls decaying rate. Weight ω_{bk} is reversely proportional to the spatial proximity of two blob centers. The decomposed feature map $\hat{S}_b^{\mathcal{F}}$ of reference blob b is termed as signature of b, which describes local feature pattern around b, and is used to construct fixation bank during training and reweight blob b during testing. At the end, the raw feature maps $\mathcal{CIOM}$ are decomposed into $\sum_{\mathcal{F} \in \{\mathcal{CIOM}\}} \mathcal{K}_{\mathcal{F}}$ signatures $\hat{S}_b^{\mathcal{F}}$, each of which associates with blob b from channel $\mathcal{F}$.

Blocked Dictionary Construction: One dictionary $\mathcal{D}_{\mathcal{F}}$ is built for each feature channel $\mathcal{F}(\mathcal{F} \in \{\mathcal{CIOM}\})$, without loss of generality, we take channel-$\mathcal{C}$-associated dictionary $\mathcal{D}_{\mathcal{C}}$ construction as an example.

For a training frame, suppose there are $\mathcal{K}_{\mathcal{C}}$ Gaussian blobs b_i on channel $\mathcal{C}$, let be $\hat{S}_{b_i}$ the decomposed feature map of b_i. If the peak location $\mathcal{P}_{b_i}$ of b_i is within some distance ξ to the peak location $\mathcal{P}_{\mathcal{IO}}$ of $\mathcal{IO}$ map, i.e., $\|\mathcal{P}_{b_i} - \mathcal{P}_{\mathcal{IO}}\|_2 \leq \xi$, then $\hat{S}_{b_i}$ is treated as a positive exemplar and assigned to the 1st block $\mathcal{D}_{\mathcal{C}}^P$; while when $\|\mathcal{P}_{b_i} - \mathcal{P}_{\mathcal{IO}}\|_2 \geq \tau, \tau > 0 \wedge \tau > \xi$, it is a negative exemplar and assigned to the 2nd block $\mathcal{D}_{\mathcal{C}}^N$. This assignment process iterates over all training frames. Finally, we build a channel-$\mathcal{C}$-associated dictionary $\mathcal{D}_{\mathcal{C}}$, $\mathcal{D}_{\mathcal{C}} = [\mathcal{D}_{\mathcal{C}}^P \mid \mathcal{D}_{\mathcal{C}}^N]$.

Each blocked dictionary $\mathcal{D}_{\mathcal{F}}$ has two big blocks, and we further divide training exemplars in each big block into smaller blocks by their peak locations. In our case, the image plane is cut into $M \times N$ non-overlapping cells, each with size $s \times s$. When the peak location of an exemplar falls into cell i, then it is assigned to sub-block i, $i = \{1,2,...,M \times N\}$. Finally, each blocked dictionary $\mathcal{D}_{\mathcal{F}}$ has $2 \times M \times N$ blocks.

Gaussian Blob Reweighting: For a test frame, after extracting decomposed feature maps of Gaussian blobs, we formulate reweighting of each blob as a group lasso problem. The final gaze density map is a weighted sum of all blobs.

For each Gaussian blob on feature maps of a test frame, to calculate its contributing weight to the final saliency map, we first solve a group lasso problem and then define its weight as a function of reconstruction errors from the positive and negative groups. Given a blob b from channel $\mathcal{F}$ with decomposed feature map $\hat{S}_b^{\mathcal{F}}$, we solve the problem:

$$\min_{\beta}\{\frac{1}{2} \parallel \mathcal{D}_{\mathcal{F}} \cdot \beta - \hat{S}_b^F \parallel_2^2 + \lambda_1 \sum_{g=1}^{G} L_g \parallel \beta_g \parallel_2 + \lambda_2 \parallel \beta \parallel_1\} \tag{2}$$

Where β_g is the coefficients of g^{th} group, $\beta = (\beta_1, \beta_2, ..., \beta_G)$, $G = 2 \times M \times N$ is the entire coefficient vector, $L_g = \sqrt{|\beta_g|}$accounts for varying group sizes, and λ_1 and λ_2 are controlling parameters making balance between reconstruction and sparsity.

We define weight of blob b as the ratio between negative and positive reconstruction errors:

$$\omega_b^{\mathcal{F}} = \varepsilon^N(\hat{S}_b^{\mathcal{F}}) / \left(\varepsilon^P(\hat{S}_b^{\mathcal{F}}) + \varepsilon\right) \tag{3}$$

where $\varepsilon^N(\hat{S}_b^{\mathcal{F}}) = \parallel \mathcal{D}_{\mathcal{F}}^N \cdot \beta^N - \hat{S}_b^{\mathcal{F}} \parallel_2$, $\varepsilon^P(\hat{S}_b^{\mathcal{F}}) = \parallel \mathcal{D}_{\mathcal{F}}^P \cdot \beta^P - \hat{S}_b^{\mathcal{F}} \parallel_2$ and ε is a small constant to avoid singularity. β^P and β^N are coefficients of positive and negative groups respectively.

The finally gaze density map $\mathbf{S}_f$ of frame f is a weighted sum of all Gaussian blobs:

$$\mathbf{S}_f = \sum_{\mathcal{F} \in \{\mathcal{CIOM}\}} \sum_{k=1}^{\mathcal{K}_{\mathcal{F}}} \omega_{b_k}^{\mathcal{F}} \cdot g_{b_k}^{\mathcal{F}} \tag{4}$$

Where $g_{b_k}^{\mathcal{F}}$is the k^{th} Gaussian blob b_k from channel $\mathcal{F}$, $\omega_{b_k}^{\mathcal{F}}$ is its weight defined in Eq.3, and $\mathcal{K}_{\mathcal{F}}$ is the number of Gaussian blobs on channel $\mathcal{F}$.

Acknowledgements: This work was supported by the National Science Foundation (grant numbers CCF-1317433 and CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and W911NF-12-1-0433), and the Office of Naval Research (N00014-13-1-0563). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.