Label Consistent Quadratic Surrogate Model for Visual Saliency Prediction

Yan Luo[1] Yongkang Wong[2] Qi Zhao[1]
[1]Department of Electrical and Computer Engineering, National University of Singapore. [2]Interactive & Digital Media Institute, National University of Singapore.

Recently, an increasing number of works have proposed to learn visual saliency by leveraging human fixations. However, the collection of human fixations is time consuming and the existing eye tracking datasets are generally small when compared with other domains. Thus, it contains a certain degree of dataset bias due to the large image variations (*e.g.,* outdoor scenes *vs.* emotion-evoking images). In the learning based saliency prediction literature, most models are trained and evaluated within the same dataset and cross dataset validation is not yet a common practice. Instead of directly applying model learned from another dataset in cross dataset fashion, it is better to transfer the prior knowledge obtained from one dataset to improve the training and prediction on another. In addition, since new datasets are built and shared in the community from time to time, it would be good not to retrain the entire model when new data are added.

To address these problems, we proposed a new learning based saliency model, namely Label Consistent Quadratic Surrogate algorithm, which employs an iterative online algorithm to learn a sparse dictionary with label consistent constraint. The advantages of the proposed model are three-folds: **(1)** the quadratic surrogate function guarantees convergence at each iteration, **(2)** the label consistent constraint enforces the predicted sparse code to be discriminative, and **(3)** the online properties enable the proposed algorithm to adapt existing model with new data without retraining.

As shown in Fig. 1, given the training samples, a discriminative sparse error term, $\|\mathbf{U}-\mathbf{L}\mathbf{X}\|_F^2$, and a classification error term, $\|\boldsymbol{v}^T-\boldsymbol{w}^T\mathbf{X}\|_2^2$, similar to [2, 3], are taken into account to approximate the discriminative sparse codes $\mathbf{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]\in\mathbb{R}^{k\times n}$ and to learn a sparse dictionary $\mathbf{D}$. The objective function in the dictionary learning problem for visual saliency prediction can be formulated as:

$$\begin{aligned}<\mathbf{D},\mathbf{L},\mathbf{X},\boldsymbol{w}>=\arg\min_{\mathbf{D},\mathbf{A},\mathbf{X},\boldsymbol{w}}&\|\mathbf{Z}-\mathbf{D}\mathbf{X}\|_F^2+\alpha\|\mathbf{U}-\mathbf{L}\mathbf{X}\|_F^2\\&+\beta\|\boldsymbol{v}^T-\boldsymbol{w}^T\mathbf{X}\|_2^2+\lambda\|\mathbf{X}\|_{1,1}\end{aligned}\tag{1}$$

where the coefficients α and β control the relative contribution of the discriminative sparse error term and classification error term, respectively. $\boldsymbol{v}$ is saliency labels from the human fixation ground truth and $\boldsymbol{w}$ is the classification weights to reconstruct the ground truth saliency labels. The matrix $\mathbf{U}\in\{0,1\}^{k\times n}$ is the discriminative sparse codes of features $\mathbf{Z}$ and $\mathbf{L}\in\mathbb{R}^{k\times k}$ is a linear transformation matrix to enforce original sparse codes in $\mathbf{X}$ to be more discriminative. Eq. (1) can be rewritten as:

$$<\tilde{\mathbf{D}},\mathbf{X}>=\arg\min_{\tilde{\mathbf{D}},\mathbf{X}}\|\tilde{\mathbf{Z}}-\tilde{\mathbf{D}}\mathbf{X}\|_F^2+\lambda\|\mathbf{X}\|_{1,1}\tag{2}$$

where $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{D}}$ are denoted as:

$$\tilde{\mathbf{Z}} = (\mathbf{Z}^T,\sqrt{\alpha}\mathbf{U}^T,\sqrt{\beta}\mathbf{v})^T\tag{3}$$
$$\tilde{\mathbf{D}} = (\mathbf{D}^T,\sqrt{\alpha}\mathbf{L}^T,\sqrt{\beta}\mathbf{w})^T\tag{4}$$

and λ is a regularization parameter.

Given a set of training samples $\tilde{\mathbf{Z}}=[\tilde{\boldsymbol{z}}_1,\ldots,\tilde{\boldsymbol{z}}_n]$ where $\tilde{\boldsymbol{z}}_i\in p(\tilde{\boldsymbol{z}})$, one sample $\tilde{\boldsymbol{z}}_t$ is drawn from $\tilde{\mathbf{Z}}$, at iteration t, to compute the decomposition of $\tilde{\boldsymbol{z}}_t$, $\boldsymbol{x}_t$, with the dictionary learned in the previous iteration, $\tilde{\mathbf{D}}_{t-1}$, using LARS algorithm [1]

$$\boldsymbol{x}_t=\arg\min_{\boldsymbol{x}\in\mathbb{R}^k}\frac{1}{2}\|\tilde{\boldsymbol{z}}_{t-1}-\tilde{\mathbf{D}}_{t-1}\boldsymbol{x}\|_2^2+\lambda\|\boldsymbol{x}\|_1\tag{5}$$

The computed $\boldsymbol{x}_t$ will be used to update the knowledge matrices $\mathbf{Q}$ and $\mathbf{H}$ via

$$\begin{aligned}\mathbf{Q}_t&\leftarrow\mathbf{Q}_{t-1}+\boldsymbol{x}_t\boldsymbol{x}_t^T\\\mathbf{H}_t&\leftarrow\mathbf{H}_{t-1}+\tilde{\boldsymbol{z}}_t\boldsymbol{x}_t^T\end{aligned}\tag{6}$$

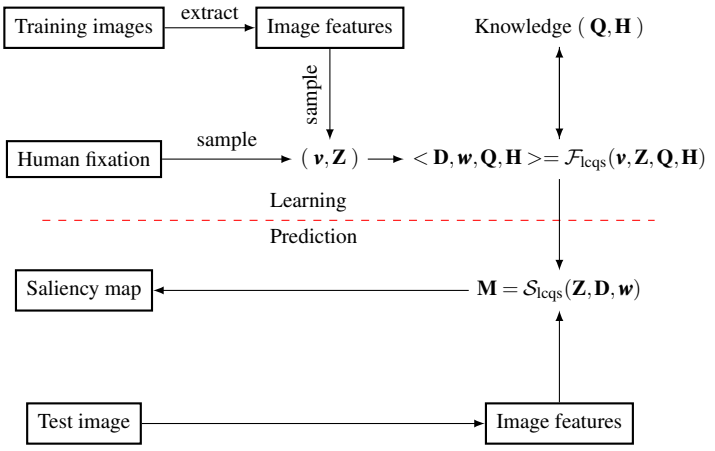

Figure 1: An overview of the LCQS saliency model.

where $\mathbf{Q}_0$ and $\mathbf{H}_0$ are both zero matrices if there is no prior information. At the meantime, the objective function in Eq. (2) can be rewritten in an iterative fashion

$$\begin{aligned}\tilde{\mathbf{D}}_t &= \arg\min_{\tilde{\mathbf{D}}\in\mathcal{C}}\frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{2}\|\tilde{\boldsymbol{z}}_i-\tilde{\mathbf{D}}\boldsymbol{x}_i\|_2^2+\lambda\|\boldsymbol{x}_i\|_1\right)\\&=\arg\min_{\tilde{\mathbf{D}}\in\mathcal{C}}\frac{1}{t}\left(\frac{1}{2}\mathrm{Tr}(\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}\mathbf{Q}_t)-\mathrm{Tr}(\tilde{\mathbf{D}}^T\mathbf{H}_t)\right).\end{aligned}\tag{7}$$

In the dictionary update process, the block-coordinate descent method is applied with $\mathbf{D}_{t-1}$ as warm restarts. The update procedure does not require any parameter to control the learning rate. In addition, it does not store the training samples and sparse codes from the previous iterations, but only the thesaurus matrices $\mathbf{Q}_t=[\mathbf{q}_{1,t},\ldots,\mathbf{q}_{k,t}]$ and $\mathbf{H}_t=[\mathbf{h}_{1,t},\ldots,\mathbf{h}_{k,t}]$. In each iteration, each basis in $\tilde{\mathbf{D}}$ is sequentially updated, *i.e.,* updating the j-th basis $\mathbf{d}_j$ at a time while freezing the other ones under the constraint $\mathbf{d}_j^T\mathbf{d}_j\le 1$. Specifically, $\mathbf{d}_j$ is updated to optimize for Eq. (7)

$$\begin{aligned}\mathbf{y}_j&\leftarrow\frac{1}{\mathbf{Q}_{jj}}(\mathbf{h}_j-\tilde{\mathbf{D}}\mathbf{q}_j)+\mathbf{d}_j\\\mathbf{d}_j&\leftarrow\frac{1}{max(\|\mathbf{y}_j\|_2,1)}\mathbf{y}_j\end{aligned}\tag{8}$$

In the dictionary update process, each basis in $\tilde{\mathbf{D}}$ undergoes the update until a convergence criteria is satisfied [4].

The proposed model is evaluated on 3 benchmark eye tracking datasets and it shows promising performances.

[1] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[2] Ming Jiang, Mingli Song, and Qi Zhao. Leveraging human fixations in sparse coding: Learning a discriminative dictionary for saliency prediction. In *IEEE International Conference on Systems, Man., and Cybernetics*, pages 2126–2133, 2013.

[3] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Label consistent K-SVD: learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, 2013.

[4] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.