Learning visual saliency using topographic independent component analysis.

Stefic, D; Patras, I

LEARNING VISUAL SALIENCY USING TOPOGRAPHIC INDEPENDENT COMPONENT ANALYSIS

Daria Stefic, Ioannis Patras

School of Electronic Engineering and Computer Science
Queen Mary, University of London, Mile End Road, London E1 4NS, United Kingdom

ABSTRACT

Understanding the underlying mechanisms that drive human visual attention is a topic of immense interest. Most of the work is focused on extracting manually selected features that might resemble the human visual processing pathway and using a combination of those features to train a classifier that learns to predict where humans look. In contrast, we will learn the features in an unsupervised way using a generalization of Independent Component Analysis (ICA), namely the topographic Independent Component Analysis (tICA). We will show that those *learned* features in combination with linear SVM outperform the hand-crafted ones. In addition, we propose a novel optimization scheme, which jointly optimizes for linear SVM and tICA pooling weights and show that it further improves the results.

Index Terms— bottom-up saliency, independent component analysis, topography, supervised pooling

1. INTRODUCTION

Understanding the mechanism of human visual attention has attracted the interest of several research areas and several computational models for predicting saliency have been proposed and then validated on fixation maps obtained by recording human gaze[1, 2]. More recent approach in learning visual saliency uses that data in order to train classifiers that predict human fixations from image features [3, 4, 5, 6, 7, 8, 5]. A detailed review on recent advances in learning saliency from human data is given in [9].

A central issue in methods that learn to predict visual saliency is the choice of the appropriate features. The main idea behind early learning-based approaches [3, 4, 5] is to extract raw image patches and train an RBF SVM classifier to discriminate fixations and non-fixations. Other works [6, 8, 7, 10] are focused on designing features that are extracted from the patches in question in order to construct training samples. [6] uses feature vectors that consist of low (orientation filters), mid (horizon detection) and high level features (face detection). The latter is an important feature as it has been found that when faces are present in an image, our earliest fixations are usually on them and that inter-subject scanpath consistency on images with faces is higher than in images without them [11]. To boost the performance, [10] introduces more feature channels and combines them to learn different types of classifiers (whereas in [6] a simple linear SVM is used). Also, those models [6, 8, 7, 10] introduce central bias as it plays an important role in saliency prediction [8]. In [7, 8] the output of a face detector is used together with three low level features (color, intensity, orientation) and optimal weights for feature integration are found using linear, least square regression. Both central bias and the learned weights are used to construct the saliency maps. In a more recent work by the same authors, optimal weights (for same feature channels) were learned using nonlinear AdaBoost [7].

Some works support the idea that applying Independent Component Analysis leads to the emergence of oriented linear filters that resemble V1 simple-cell receptive fields [12, 13]. In order to explain the emergence of the V1 topography, in [12] a more generalized model, namely topographic ICA, is introduced (a detailed comparison between ICA and tICA is given in [13]). The main idea in this transform is to introduce nonlinearity in order to represent invariant features while considering the topographic ordering of basis vector so that vectors with stronger higher order correlations are close to each other in the topology. The motivation for such modeling comes from the fact that in the higher levels of the visual system, there are cells which respond to complex object parts irrespectively of their spatial location. This can not be well represented by linear feature detectors, but this topological ordering might lead to the emergence of complex cell properties where each neighborhood cell acts like a complex cell [12].

In contrast to other works that either use combinations of handcrafted features with linear[8, 6] or Adaboost classifiers[7], or raw image patches with RBF SVM [5, 3, 4] in this work we propose to *learn* the features and use them to feed the linear SVM. Inspired by the works of [12, 14] we do so by using a model that is argued to resemble the complex cells of the V1 human visual pathway, namely topographic Independent Component Analysis. In the second phase we will use a linear binary SVM criterion (as in [6]) and propose a novel scheme that jointly learns both the SVM weights and the weights of the pooling matrix in tICA. This

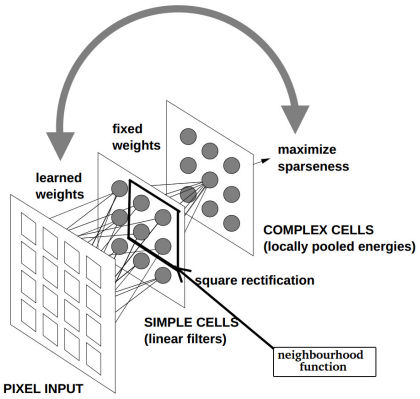

Fig. 1. tICA model[14]

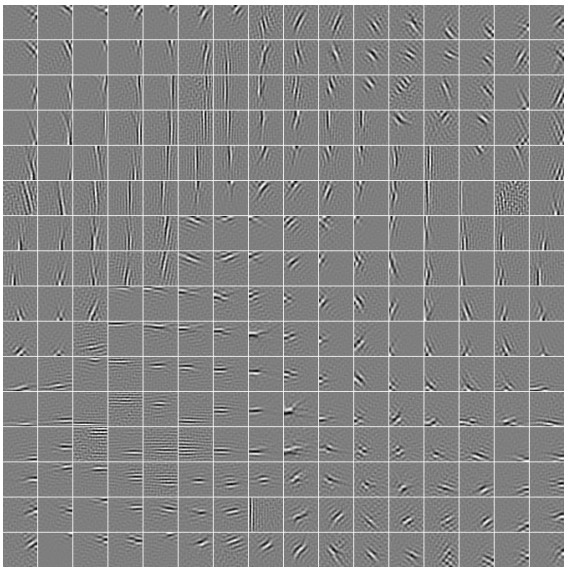

Fig. 2. tICA basis vectors

is in contrast to other works [12, 13, 14] in which the weights of the pooling matrix are kept fixed. To summarize, our contribution is two fold. First we show that our unsupervised feature learning method outperforms sophisticated state of the art methods that learn saliency using the combination of manually selected salient features, and second, we show that our novel scheme that jointly optimizes the classifier weights and the pooling layer weights further improves the results.

2. METHODOLOGY

2.1. Topographic independent component analysis

tICA can be described as a two-layered network with squared nonlinearity in the first, and square-root nonlinearity in the second layer of the network (Fig. 1). As proposed in [12], only the weights of the first layer $\mathbf{v}_j$ are learned and the second layer pooling matrix $\pi(i,j)$ is fixed. This matrix expresses the proximity of the features with indices i and j in a predefined underlying topography. For example, $\pi(i,j)$ is 1 if the feature j is in a 3x3 square neighbourhood of feature i; otherwise $\pi(i,j)$ is zero. This type of arrangement restricts features in the same clusters to have adaptive *size and weights* that depend on the nature of their features. With our joint optimization procedure described in 2.2 we are going to make the *weights* adaptive and dependent on the nature of the supervision criterion, i.e. we will change the non-zero elements of the pooling matrix, while keeping the size of the neighbourhood fixed, i.e. the zero elements stay zeros.

Assuming that we have observed a set of image patches $\mathbf{z}^t, t = 1, ..., T$ after whitening, tICA learns $\mathbf{v}_j, j = 1, ..., n_1$ by solving the following optimization problem:

$$\min_{\{\mathbf{v}_j|\ j=1,...,n_1\}} \sum_{i=1}^{n_2} \sum_{t=1}^{T} \sqrt{\sum_{j=1}^{n_1} \pi(i,j)(\mathbf{v}_j^T \mathbf{z}^t)^2} \qquad (1)$$

where n_1 is number of the neurons in the first layer and n_2 is the output dimensionality (i.e. number of neurons in the second layer after the pooling). In addition, the vectors $\mathbf{v}_j, j = 1, ..., n_1$ are constrained to be orthogonal. Minimization of this function is done by batch gradient descent.

Once the vectors $\{\mathbf{v}_j\}$ are estimated, given a whitened image patch $\mathbf{z}$, we can extract a feature $\mathbf{x} = [\mathbf{x}_1, ...\mathbf{x}_i, ..., \mathbf{x}_{n_2}] \in R^{n_2}$.The i-th element of $\mathbf{x}$ is given by:

$$\mathbf{x}_i = \sqrt{\sum_{j=1}^{n_1} \pi(i,j)(\mathbf{v}_i^T \mathbf{z})^2}. \qquad (2)$$

In Fig. 2 we show the basis vectors we obtained for tICA in order to illustrate the emergence of topographic organization. In contrast to Gabor-like linear features, in the case of tICA the location and orientation change smoothly in the topographic grid.

There were a couple of attempts to estimate both layers in this two layer model (i.e. learning the matrix π) using an energy-based approach in an unsupervised manner [15, 16]. However, in [16] there was no significant reduction in energy dependencies of the proposed model, and in [15] a novel estimation method for two layer network similar to the one of tICA was presented but no further analysis in terms of comparison to the original topographic ICA was given. In the section 2.2 we propose to learn the pooling matrix in a supervised manner using the classification criterion.

2.2. Supervised pooling

Typically, learning saliency is posed as a binary classification problem over some handcrafted features [3, 4, 5, 6, 8, 7]. In

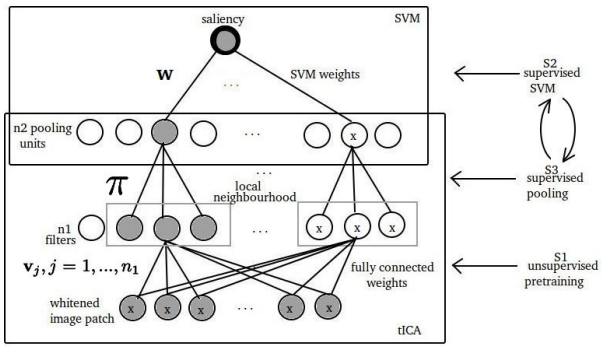

Fig. 3. Illustration of our proposed architecture. Due to clarity, weights are shown only for the marked neurons.

this framework, let $\{\mathbf{x}^t, y^t\}_{t=1}^T$ denote a set of training vectors $\mathbf{x}^t \in R^{n_2}$ representing the extracted features of image patches (see (2)) and their corresponding labels $y^t \in \{-1, 1\}$. $y^t = 1$ if the position where $\mathbf{x}^t$ is extracted is a fixation point and $y^t = -1$ otherwise (fixation point is a point which users that were presented with the image in question fixated their gaze; for details see section 3 or [6, 18, 5]). Instead of solving the standard SVM problem where only the weights of the SVM are learned, we solve the following minimization problem:

$$\min_{\mathbf{w},b,\pi} f(\mathbf{x}; \mathbf{w}, b, \pi), \tag{3}$$

where

$$f(\mathbf{x}; \mathbf{w}, b, \pi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{t=1}^{T}(\max(0, 1 - y^t(\mathbf{w}^T\mathbf{x}^t + b)))^2, \tag{4}$$

where π is the pooling matrix and $\mathbf{x}$ is a function of π (see (2)). We solve this by following a coordinate descent iterative optimization procedure which iterates between steps S2 and S3 (see Fig. 3). Each of those two steps is a convex subproblem, and solved with respect to a subset of the unknown variable.

The full learning algorithm, which consist of three steps, is given below:

S1. Pretrain the lowest layer of the architecture according to the tICA criterion (see (1)), i.e. in an unsupervised manner.

S2. Minimize with respect to SVM parameters, that is:

$$\min_{\mathbf{w},b} f(\mathbf{x}; \mathbf{w}, b, \pi). \tag{5}$$

This is a standard SVM problem that we solve using LIBLINEAR library[19].

S3. Minimize with respect to the pooling matrix π, that is:

$$\min_{\pi} f(\mathbf{x}; \mathbf{w}, b, \pi). \tag{6}$$

Following the idea of training an SVM in the primal form [20], this optimization is solved using batch gradient descent. In our optimization procedure, one gradient step for updating π consist of four substeps:

1. Take a step following the gradient of (6) with respect to π, that is:

$$\pi^{k+1} \leftarrow \pi^k - \mu \begin{cases} 0 & \text{if } y^t(\mathbf{w}^T\mathbf{x}^t + b) \geq 1 \\ \frac{\partial f(\mathbf{x};\mathbf{w},b,\pi)}{\partial \pi} & \text{if } y^t(\mathbf{w}^T\mathbf{x}^t + b) < 1 \end{cases} \tag{7}$$

   where μ is the learning rate.

2. Constrain π to have non-negative elements, that is:
   $\pi_j^{k+1} \leftarrow \pi_j^{k+1} + \min_l(\pi_j^{k+1}(l)), j = 1, ..., n_2,$
   where l is an index to the elements of the vector π_j^{k+1}.

3. Preserve the neigbourhood size, i.e. setting the elements outside of the neighbourhood to zero, that is:
   $\pi_j^{k+1} \leftarrow \pi_j^{k+1} \circ \pi_0, j = 1, ..., n_2,$
   where π_0 is the fixed pooling matrix as proposed in [14] and $\circ$ denotes elementwise multiplication.

4. Normalize π:
   $\pi_j^{k+1} \leftarrow \frac{\pi_j^{k+1}}{\|\pi_j^{k+1}\|}, j = 1, ..., n_2.$

The algorithm converged after few iterations between steps S2 and S3 resulting in an energy decrease of 3-5%, depending on the dataset.The number of iterations in substep S3 was experimentally set to 100 for all datasets. The size of the neighbourhood is proposed to be set to 3x3, however, in our experiments we have found that 5x5 gives better results and also better improvement after the joint optimization (since the capacity for learning the weights is larger in the larger neighborhood). Other parameters were chosen either as proposed in the literature [14] (n_1 and n_2 in tICA network) either by cross validation (C value of SVM). Also, we have found that using a single image patch size (41x41) was sufficient, while in the other works that we compare our results to, combination of features on multiple scales are used.

3. EXPERIMENTAL RESULTS

We validate our method on three datasets that contain human eye fixations that were recorded while the subjects observed images of natural scenes: the dataset of Judd *et al.* [6] (MIT dataset), the dataset from Bruce and Tsotsos [18] (Toronto dataset) and the dataset used in [5, 3, 4] (Kienzle dataset). Subjects observed images in a *free-viewing* manner, i.e. without any given task, which suggest using the bottom up saliency detection approach using, in our case, learned low level features. In all of our experiments we follow the same

training and testing protocols as in the other literature that we compare our results to, except that in our case sampling fixations in 1% of most salient areas and non-fixations in lowest 70% of salient areas worked best, as opposed to 20/70 used in the other literature (the results we obtained with 20/70 were also better when compared to the ones from the literature, but we report our best ones obtained with 1/70 sampling).

Real valued saliency maps are obtained by computing per pixel SVM responses $\mathbf{w}^T\mathbf{x} + b$) where $\mathbf{x}$ are the features calculated as described in previous section on a patch centered around the point in question (see (2)). Then, as proposed in [6], the maps are smoothed with a Gaussian filter. Following the literature we report our results by means of the AUC (Area Under the ROC Curve).

In Table 1 we compare our results to the state of the art in the learning saliency paradigm without the central bias. Wherever possible, i.e. for MIT and Toronto dataset, we compare our results to the methods that use a linear classifier. For the Kienzle dataset for which there are no such results, we compare to all of the reported results. In [8, 7] the feature vector consists of two color, one intensity, four orientation and a face channel. Our results outperform the ones using linear integration with optimal weights learned using linear, least square regression, even though they use as feature the output of a high level face detector. The latter has shown to be an important feature for saliency prediction (in [7] it was found that the face channel is the most informative for the MIT dataset). On the Kienzle dataset the only reported results are for the model of Itti *et al.* [21] which uses the same channels but without learning their weights using human ground truth data, and for the model of [5] in which centre surround patterns are learned with an RBF SVM on raw image patches. Our model outperforms both and achieves state-of-the-art on this dataset, which is the most challenging since it contains scenes taken in the nature that are without any particular regions of interest, as opposed to other two datasets. In the same table we show that our proposed joint optimization procedure (learned pooling) gives consistently better results than using linear SVM on top of tICA features - the improvements are 0.9, 1.0 and 0.1% for MIT, Toronto and Kienzle dataset, respectively. We also observe the increase in the classification performance (both for training and testing sets), the results for which we omit due to space restrictions.

In order to give some insight to the advantages and limitations of the proposed method we illustrate some examples of our saliency maps in comparison to the human ground truth in Fig. 4. The maps were tresholded in a way that the 10% highest values in the map are considered salient. In the first two rows we show some representative examples of our saliency maps to illustrate how our method can usually predict human fixations very well in scenes that obtain clear shapes and objects like cars, buildings and traffic signs. In rows 3 and 4 we show maps that do not match to the human ground truth since they contain some semantic content (letters or faces).

Table 1. Comparison to the state-of-the-art

	MIT	Toronto	Kienzle
Linear integration [21, 11]	0.776	0.828	-
Linear integration with optimal weights [8]	0.792	0.834	-
Itti *et al.*[21]	-	0.828	0.620
Center-surround patterns [5]	-	-	0.640
Ours (tICA + linear SVM)	0.803	0.850	0.654
Ours (learned pooling)	**0.812**	**0.860**	**0.655**

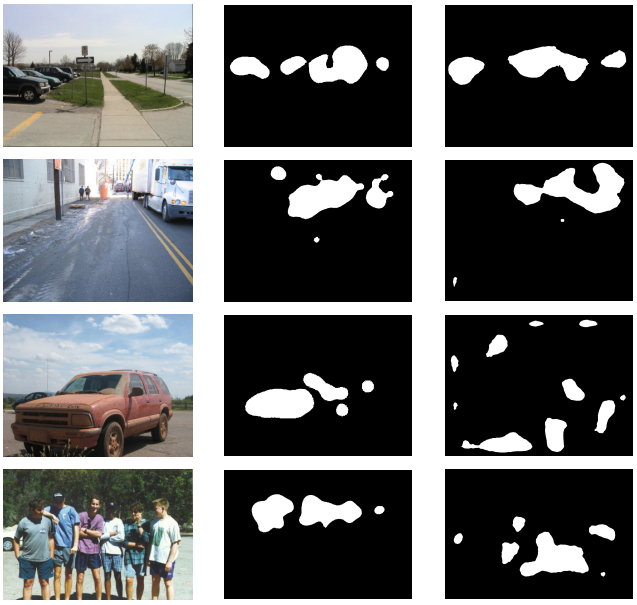

Fig. 4. (a) stimuli image, (b) human ground truth saliency map, (c) our saliency map

4. CONCLUSIONS

In this paper we propose a method that learns bottom-up visual saliency by learning features instead of using manually selected ones. By doing so, we do not handcraft them, rather we infer them from training images. We show that the performance of our model using only these learned low level features outperformed approaches where higher level features such as face detectors were used in combination with linear SVM or the one where an RBF SVM on raw patches was used, and we achieve the state of the art results on the Kienzle dataset. Finally, we show that our optimization scheme that optimizes the tICA pooling matrix weights jointly with SVM weights further improves the results.

5. REFERENCES

[1] A. Borji, H. Tavakoli, D. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *Proceedings of International Conference on Computer Vision*, 2013.

[2] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Learning*, vol. 35, no. 1, pp. 185–206, 2013.

[3] W. Kienzle, F. A. Wichmann, B. Scholkopf, and F. O. Matthias, "A nonparametric approach to bottom-up visual saliency," in *Proceedings of Neural Information Processing Systems Conference*, 2007.

[4] W. Kienzle, F. A. Wichmann, B. Scholkopf, and F. O. Matthias, "Learning an interest operator from human eye movements," in *In Beyond Patches Workshop, International Conference on Computer Vision and Pattern Recognition*, 2006.

[5] W. Kienzle, F. O. Matthias, B. Scholkopf, and F. A. Wichmann, "Center-surround patterns emerge as optimal predictors for human saccade targets," *Journal of Vision*, vol. 9, no. 5, pp. 1–15, 2009.

[6] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of International Conference on Computer Vision*, 2009.

[7] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using adaboost," *Journal of Vision*, vol. 12, no. 6, pp. 1–15, 2012.

[8] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, pp. 1–15, 2011.

[9] Q. Zhao and C. Koch, "Learning saliency based visual attention: A review," *Signal Processing*, vol. 93, no. 6, pp. 1401–1407, June 2012.

[10] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proceedings of International Conference on Computer Vision and Pattern Recognition*, 2012.

[11] M. Cerf, J. Harel, W. Einhauser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proceedings of Neural Processing Systems Conference*, 2007.

[12] A. Hyvarinen, P. Hozer, and M. Inki, "Topographic ica as s model of v1 receptive fields," in *Proceedings of International Joint Conference on Neural Networks*, 2000.

[13] R. Mutihac and M. M. van Hulle, "Statistics of feature extraction by topographic independent component analysis from natural images," in *Proceedings of International Conference on Electronics Control and Signal Processing*, 2003.

[14] A. Hyvarinen, Y. Hurri, and P. Hoyer, *Natural Image statistics*, Springer, ISBN: 978-1-84882-491-1, 209.

[15] U. Koster and A. Hyvarinen, "A two-layer ica-like model estimated by score matching," in *Proceedings of International Conference on Artifical Neural Networks*, 2007.

[16] U. Koster, T. Lindgren, and M. Gutmann, "Learning natural image structure with a horizontal product model," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation*, 2009.

[17] S. Sukhbaatar, T. Makino, and K. Aihara, "Auto-pooling: Learning to improve invariance of image features from image sequences," in *Proceedings of International Conference on Learning Representations*, 2013.

[18] N. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Proceedings of Neural Information Processing Systems Conference*, 2005.

[19] R. Fan, K. Chang, C. Hsein, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[20] O. Chapelle, "Training a support vector machine in the primal," *Neural Computation*, vol. 19, no. 5, 2007.

[21] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Learning*, pp. 1254–1259, 1998.