Analysis of sampling techniques for learning binarized statistical image features using fixations and salience

Hamed R.-Tavakoli, Esa Rahtu, Janne Heikkilä

Center for Machine Vision Research, University of Oulu, Finland
{hamed.rezazadegan, esa.rahtu, janne.heikkila}@ee.oulu.fi

Abstract. This paper studies the role of different sampling techniques in the process of learning *Binarized Statistical Image Features* (BSIF). It considers various sampling approaches including random sampling and selective sampling. The selective sampling utilizes either human eye tracking data or artificially generated fixations. To generate artificial fixations, this paper exploits salience models which apply to key point localization. Therefore, it proposes a framework grounded on the hypothesis that the most salient point conveys important information. Furthermore, it investigates possible performance gain by training BSIF filters on class specific data. To summarize, the contribution of this paper are as follows: 1) it studies different sampling strategies to learn BSIF filters, 2) it employs human fixations in the design of a binary operator, 3) it proposes an attention model to replicate human fixations, and 4) it studies the performance of learning application specific BSIF filters using attention modeling.

Keywords: Binary operators, visual attention, salience modeling.

1 Introduction

The research on image descriptors is a well-studied area in computer vision. In general, image descriptors describe the visual characteristic (e.g., shape, color, texture, motion) of the image. They are the building blocks of many vision related tasks such as image retrieval, recognition tasks (e.g., texture, object, face), action recognition, facial expression analysis, and etc.

Today, the computer vision domain is replete with image descriptors. Some descriptors are more generic, e.g., SIFT [10], SURF [1], BRIEF [2], DAISY [18] and their variants, compared to other operators such as LBP [12], LPQ [13] which are mostly developed for class specific applications (e.g., texture classifications, and face recognition). Nonetheless, they are somehow linked by a common framework of Filtering, Labeling and Statistics (FLS) which provides a unique implementation for LBP and SIFT like features [4].

Adopting the concepts of [4], one can write the LBP operator as the thresholded-quantized-mapped response of a series of multi-directional filter banks. While

traditionally the filters are hand tuned, intrigued to improve quality of filters, [9] proposed to learn the filters using image statistics in which the premise is that statistically learned filters convey image information better. Nonetheless, such an approach poses a new challenge by requiring effective training of the filters. Thus, this paper tries to seek a suitable answer by investigating the domain of salience modeling and visual attention. Initially, it exploits human fixations to train BSIF filters from natural image statistics in order to analyze possible relation between informative regions and training of filters.

Afterwards, motivated by the success of learning based methods, e.g. [16], in which a set of filters specific to a class category is learned, this paper explores learning the filters from application specific data sets and particular class categories. However, it faces a difficulty in using human fixations because there is no such a data set available. To compensate, it develops an attention model to replicate human fixations during the learning process.

Eventually, the performance of the sampling strategies is studied in several applications such as texture classification and face recognition. It will be demonstrated that learning of filters somehow benefits from selective sampling and the proposed framework for attention-based learning of filters improves the performance of face recognition.

1.1 Related work

This paper targets domain of binary patterns such as LBP [12]. Such operators treat the relation of each pixel and its surrounding as a binary code string. Consequently, an image is represented by the probability distribution of binary code strings obtained in terms of histograms. Thus, the paper adopts the binarized statistical image features (BSIF) to investigate the role of underlying data set information in the process of learning statistical representations.

BSIF binarizes the response of a set of statistically learned filters with a threshold at zero, in which each filter response is in correspondence with a different filter. The filters are learned by maximizing the statistical independence of the filter responses using *Independent Component Analysis* (ICA) [6].

In a few words, given an image I and a filter w_i of size $l \times l$, the filter response is

$$s_i = w_i * I, \tag{1}$$

where s_i is the response of the i-th filter, and $*$ is the convolution operator. For a specific pixel $\mathbf{x}$, BSIF derives a binarized filter response such that $b_{i,\mathbf{x}} = 1$ if $s_i > 0$ at $\mathbf{x}$, otherwise $b_{i,\mathbf{x}} = 0$. Thus, in presence of n filters a binary string of length n describes each pixel.

BSIF learns the filters using independent component analysis. To this end, it forms a training set of image patches by taking random samples from natural images. Afterwards, it employs a canonical preprocessing step and performs *Principal Component Analysis* (PCA) to obtain dimension-reduced whitened data samples. Eventually, it utilizes a standard ICA algorithm [6] to obtain a set of linearly defined filters.

2 Fixations and BSIF

In order to learn the filters, BSIF requires several sampled image patches. It obtains them by randomly sampling image patches from natural images. Nonetheless, there are arguments and evidence that supports the fact that random sampling does not necessarily provide the best informative image patches. For instance [8] proposed taking image patch samples from the most salient regions to make descriptors in a recognition task and demonstrated the success of the attention based learning.

Intrigued to investigate application of informative regions in training of BSIF, the filters are learned using patches extracted around human fixation points on natural image statistics. The learning procedure is as follows: 1) The images are converted to grayscale, 2) The patches are selected around the fixation points of observers, 3) the DC-component (i.e., mean value) of each image patch is discarded, 4) The patches are dimension reduced and whitened, 5) the independent components are estimated. In mathematical terms, for an image patch, $\{x\}$, of size $l \times l$ centered at $\mathbf{x}$, one can apply ICA algorithm to estimate the independent components, i.e., the $n \times l^2$ filter matrix $\mathbf{W}$. The filter matrix includes n vectorized filters, w_i, of length l^2. Knowing that the all-in-one response of the filters on a patch can be formulated as $s = \mathbf{W}\{x\}$, one can write

$$s = \mathbf{U}\mathbf{z}, \tag{2}$$

where $z = \mathbf{V}\{x\}$, $\mathbf{U}$ is a $n \times n$ matrix which is estimated via ICA. The matrix $\mathbf{V}$ conveys the PCA whitening procedure which facilitates estimation of the orthogonal matrix $\mathbf{U}$ using the fact that $\mathbf{z} = \mathbf{U}^{-1}s$. Eventually, by estimating $\mathbf{V}$ and $\mathbf{U}$, it obtains $\mathbf{W} = \mathbf{U}\mathbf{V}$.

2.1 Fixations' replicate

In order to boost the performance of the operator, one may suggest learning the filters tuned for a specific data set, e.g. learning the filters from face images for a face recognition task. In this context, the aforementioned methodology for learning filters has one disadvantage which is the requirement of human fixations. Access to reordered fixations on class specific data is not always possible due to expensive gathering procedures. To compensate, this section introduces an artificial mechanism of fixation selection. The mechanism relies on a salience map, which is obtained using natural image statistics, and application of inhibition of return (IOR) procedure in selection of most salient region.

To compute the salience map, the proposed framework utilizes the filters learned from the previous step and intensity of an image. For each filter, it employs the *Saliency Using Natural statistics* (SUN) [21] to derive a conspicuousness map. SUN defines bottom-up salience as $P(F)^{-1}$ in which F indicates w_i learned as described before. It approximates $P(F)$ as the generalized Gaussian distribution (GGD) estimate of unidimensional distributions such that $P(F = f) = \prod_i P(f_i)$, where f_i is the i-th element in f, and

$$P(f_i) = \frac{\theta_i}{2\sigma_i \Gamma(\theta_i^{-1})} \exp\left(-\left|\frac{f_i}{\sigma_i}\right|^{\theta_i}\right), \tag{3}$$

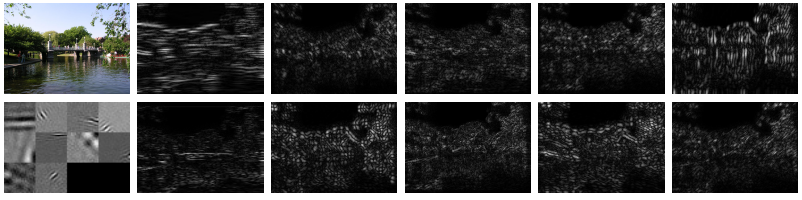


Fig. 1. ICA filter response, left column depicts an image and 10 ICA filters and on the right side the corresponding filter responses are visualized.

The discriminative power of ICA filters are even enhanced by nonlinear weighting of each dimension of f using GGD fit to their responses [15]. Fig. 1 depicts the conspicuousness maps obtained from 10 of the ICA filters. Traditionally these conspicuousness maps are combined with equal weights to derive a central saliency map (e.g. [21, 8]). Contrarily, the proposed framework treats them as features and employs linear *Support Vector Machines* (SVM) to combine the conspicuousness maps and intensity features to produce a salience map. To this end, it learns a linear SVM on a groundtruth consisting of human fixation density maps in which top 10% salient regions form positive set and top 10% non-salient regions form negative set. Thus, given a training set of n points with the feature input $x_i \in \mathcal{R}_n$ and the corresponding target label $y_i \in \{-1, +1\}$, the SVM is defined as a linear scoring function with a prediction rule such that

$$\hat{y}(x_i) = \text{sign}(\omega^T x_i + \beta), \tag{4}$$

where β is the bias and ω is a weight vector. The weight vector ω is obtained via a minimization problem as follows

$$\begin{aligned} \min_{\omega}. \quad & \frac{1}{2}\omega^T\omega + \lambda \sum_{i=1}^{n} \zeta_i \\ \text{s.t.} \quad & \hat{y}(\omega^T x_i + b) = 1 - \zeta_i \\ & \zeta_i \geq 0 \quad i = 1 \ldots n \end{aligned} \tag{5}$$

where λ is a smoothing regularization parameter balancing the trade-off between error and margin. Consequently, the saliency map is defined as the score obtained by combining the features using ω. In other words, for a feature vector of f, the saliency Sal is defined as $Sal = \omega^T f$. Fig. 2 depicts saliency maps produced using the described technique.

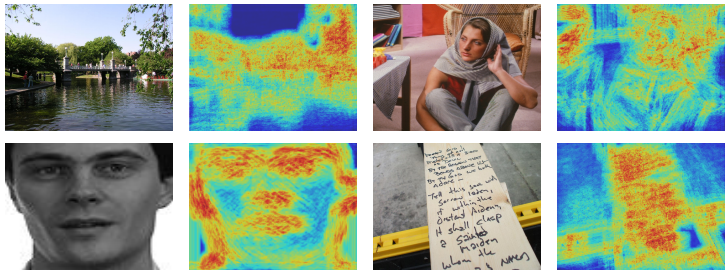


Fig. 2. Salience maps produced using the described technique. As depicted, more salient regions are somehow meaningful to human, e.g. eyes and mouth regions of the face.

To select fixations, the proposed method applies an inhibition of return (IOR) like mechanism. As depicted in Fig. 3, it implements an iterative scheme consisted of 1) it picks randomly among the salient locations, 2) it attenuates the salience map response at the selected fixation proportional to a Gaussian kernel. The procedure is repeated until enough number of fixations are obtained which replicate the human fixations. Fig. 4 visualizes samples taken by such a process, as depicted, samples taken using artificial fixations are concentrated on more meaningful parts of the image compared to random samples.

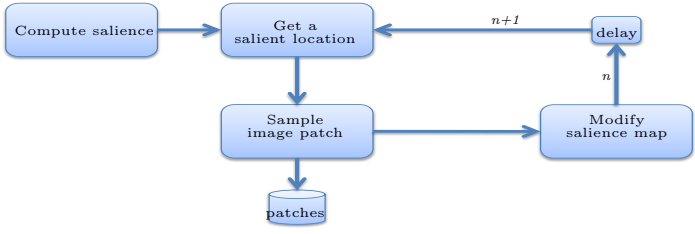


Fig. 3. Sampling image patches using an artificial fixation generation mechanism. For an image, a saliency map is generated and fixations are taken by considering the salient locations. Each time, a location among salient locations is selected randomly and its corresponding image patch is extracted. Afterwards, the salience map is modified and the current fixation location is attenuated to reflect its selection. The process continues over time until enough samples are taken

3 Experimental Analysis

This section assesses the aforementioned scenarios. The analysis covers experiments on texture and faces. Initially, it discusses the texture classification experiments. Afterwards, it continues with the experiments on face recognition which is followed by a discussion.

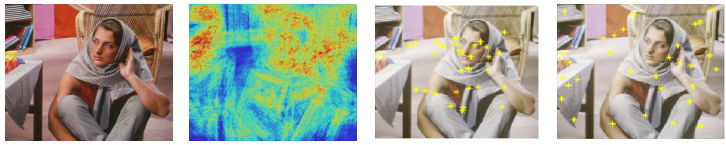


Fig. 4. Sampling using artificial fixations, from left to right, original image, saliency map, samples taken using artificial fixations, random sampling.

3.1 Texture classification

The texture experiments assess two sampling strategies for learning the BISF filters. It compares filters learned from patches taken randomly with the filters learned from patches centered on human fixations. The filters are learned using natural images provided by MIT [7] database. It consists of 1,003 images along with the eye movement statistics, particularly fixations, of 15 viewers at a distance of 48cm. The image set includes natural indoor and outdoor images; each image is presented for 3 second. In order to learn the filters, the images are converted to gray-scale and 500,000 image patches are sampled either randomly or using human fixations. The image patches are of the sizes 3×3, 5×5, and 7×7, as bigger patch sizes are demonstrated not to perform well on the textures [9, 20], which are learned at different number of bit levels (i.e. ICA filters) ranging from 5 to 11.

To perform texture analysis, this study utilizes CUReT [3], Outex [11] datasets. The Columbia-Utrecht (CUReT) dataset consists of 61 texture classes, each observed with almost 205 viewing and illumination combinations (more than 12,000 images in total). The categories include a variety of surfaces such as specular, diffuse, isotropic, and etc. The Outex database consists of several test suits. This study utilizes test suits TC_00002 and TC_00012. Each of them consists of 24 classes of texture, while TC_00002 has no rotation and contains only one illuminant, TC_00012 has three illuminants and considers 7 rotation orientations[1].

The classification procedure is chosen to be consistent with the protocols used in [9]. In other words, texture classification is carried out using nearest neighbor classifier in which the distance measure is $\mathcal{X}^2$ using l_1-normalized feature histograms. To classify the CUReT textures, the images are grouped into non-overlapping train and test sets and the procedure is repeated 100 times as described in [19]. The Outex experiments utilizes the provided partitions of [11].

Fig 5 depicts the results of the two differently trained filters on the CUReT database. There seems to be a small difference between the two sampling approaches. Nonetheless, the filters learned using the fixation sampled images perform marginally better than randomly learned ones meanwhile achieving maximum accuracy of 96.6.

Fig. 6 visualizes the results of the Outex database. As depicted in 6(a), similarly the Outex TC_00002 results indicate slight improvement in training the

[1] Please see: http://www.outex.oulu.fi/index.php?page=classification for detailed information on test suits.

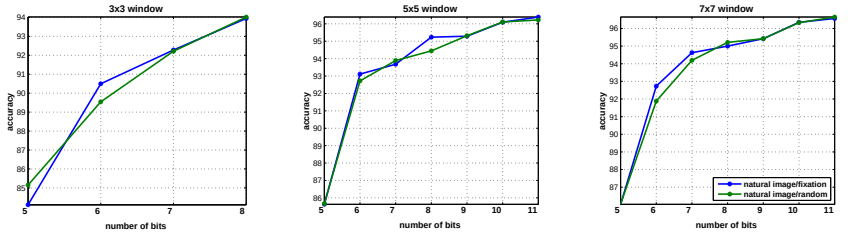
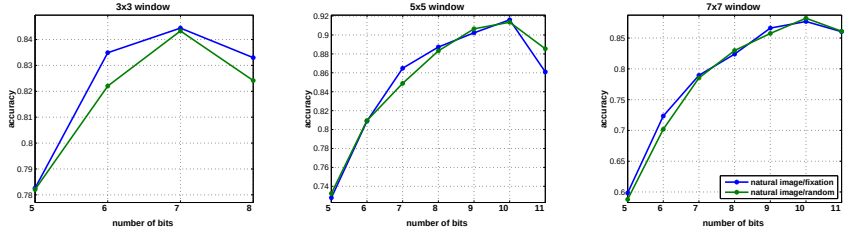
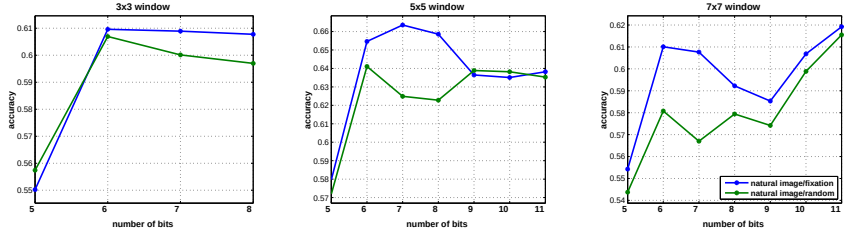


Fig. 5. CUReT, performance analysis of filters trained randomly compared to filters trained on human fixations.



(a) TC_00002



(b) TC_00012

Fig. 6. Outex, performance analysis of filters trained randomly compared to filters trained on human fixations.

filters using patches sampled at fixation points. On the other hand, the performance analysis of Outex TC_00012, showed in 6(b), reveals a 4% performance improvement using fixations to train the BSIF filters (5×5-7 bits performing 66.4% vs. 5×5-6 bits performing 64%).

Comparing the results on Outex TC_00012[2] with TC_00002 and CUReT conveys that *the selective training of filters boosts the performance of operator in handling data carrying more information*. Thus, this study motivates the assessment of sampling strategies on more complicated scenarios and data. Intrigued to have a better understanding, this paper performs a series of analysis on face recognition task.

[2] TC_00012 is difficult because it contains several rotations and illuminants.

3.2 Face recognition

This section considers face recognition task in order to study the role of sampling in training of BSIF in a more challenging task. It extends the sampling mechanism by incorporating faces in the learning process. To learn the filters from face images, it adopts a cropped version of the Labeled Faces in the Wild (LFW) [5], recognized as LFWcrop [17][3]. It consists of more than 13,000 images of faces which are cropped to prevent the recognition by getting advantage of the background information. However, it does not have any eye tracking data available. Therefore, the artificial fixation generation scheme mentioned above is employed in order to select the location of each image patch selectively. Eventually, this section analyzes a set of 4 different filters: the filters learned on face data using random sampling and artificial fixation selection mechanism and the two filters applied in the texture analysis study. The same parameters and configurations applied in the learning of the filters on face data.

The experiments are carried out on the FERET database [14] using the frontal profile images. The images are partitioned into gallery (*fa*) and *fb* probe images. The gallery consists of 1196 images, and the probe consists of 1195 images with varying facial expressions. Fig. 7 depicts some of the face images. It is expected that the performance will be somehow related to the amount of information the filters would be able to encode and the data of the experiments.

Fig. 7. Sample images from FERET data base.

The recognition procedure initially crops the images using the location of subjects' eyes to have the complete frontal face in the center of frame. Afterwards, the images are normalized to a canonical size of 128×128. It divides the face image into 8×8 non-overlapping rectangular regions and computes the BSIF descriptor independently for each segment. Concatenation of l_1 normalized descriptors makes an image descriptor. The classification uses nearest neighbor and $\mathcal{X}^2$ distance measure.

Fig. 8 depicts the results of the face recognition task. The 7×7 filters with 12 coding bits achieve the performance of 94.23%. The comparison of curves somehow expresses that *the number of coding bits (i.e. information) has a direct relationship with using selective sampling approach in the learning process of ICA filters*. It is worth-noting that while learning small filters does not benefit from training on class specific data, bigger windows and higher number of bits

[3] Download link: http://conradsanderson.id.au/lfwcrop/

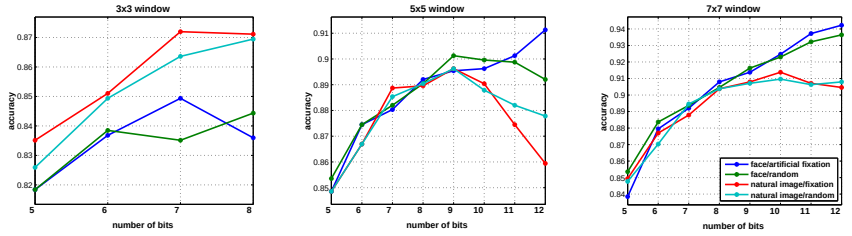


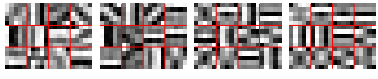
Fig. 8. Face recognition and sampling strategies using different window sizes.

get advantage of such data. Nonetheless, the behavior of curves raises some questions which this study tries to address in the next section.

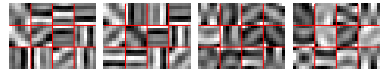
3.3 Discussion

The variation in the curves depicting performance of texture and face recognition raises some questions. **Why there is a marginal contribution in adapting selective sampling for texture?** The texture often consists of simple repeating patterns which makes them difficult to discriminate. Nonetheless, the learning of such simple structures are somehow easily doable by having enough number of samples via ICA. As depicted in Fig. 4, the filters learned from natural image statistics consists of similar structures which are probably enough to represent the textures. Nonetheless, the significance of selective sampling becomes apparent in applying a rotation variant operator (i.e. BSIF) to rotated texture samples; referring to Fig. 6(b), one realizes that selectively trained filters perform 4% better than randomly trained filters in the task of the recognition of TC_00012 textures.

(a) 3×3 Filters of 7 bits



(b) 5×5 Filters of 12 bits



(c) 7×7 Filters of 12 bits

Fig. 9. Visualization of the ICA filters learned using different sampling strategies, from left to right: randomly learned from natural images, learned from fixations using natural images, learned randomly from face images, and learned using artificial fixations from face images.

The sampling strategies and learning are not limited to selective sampling. Face recognition included filters learned on class specific trained filters, i.e. filters

learned on faces. **Is there any benefit in training the filters on class specific data sets?** As depicted in Fig. 8, the maximum face recognition rate is achieved using the filters which are trained on class specific data and convey more information. In other words, learning ICA filters from class specific data becomes useful as the amount of information required to perform a task increases. To find grounds for such a behavior, Fig. 4 visualizes the ICA filters learned using various sampling techniques and data. As shown, in case of small filters of size 3×3, the filters learned on natural statistics present a structure similar to gradient filters, which effectively encodes almost any data in such a small neighborhood. Consequently, natural image statistics somehow perform better for that specific scale. Contrarily, as the filter size and number of bits increase, one can observe – e.g., from Fig. 9(c) – that the filters trained from class specific data are absolutely different, probably reflecting the underlying data better. The performance even slightly improves using the artificial fixation based sampling meanwhile it seems that more complicated filters are learned using fixation.

What is the verdict? The conducted experiments reveals that there is a relation between the amount of information conveyed by the BSIF filters, sampling strategies and learning filters from class specific data. The conclusion is that *depending on the amount of information embedded in the data, reaching the optimum operation point may benefit from learning on class specific data and selective sampling.*

4 Conclusion

This study examined different sampling strategies for learning ICA filters used by BSIF operator. These strategies include random sampling and selective sampling. The study employed two techniques for taking the samples selectively, first it utilized fixation points on natural image statistics, second it developed an artificial fixation generation scheme to replicate human fixations in the process of learning the filters.

To generate artificial fixations, it proposed an attention model. The attention model derives a salience map using natural image statistics responses and linear support vector machine. Afterwards, it implements an inhibition of return mechanism to replicate the human fixations. Consequently, the proposed locations of image patches are more concentrated on meaningful areas of the image. The mechanism is particularly applied in the process of learning ICA filters for the task of face recognition. Eventually, the proposed mechanism is applied to replicate human fixations in the process of learning from face data.

The experiments suggest that using selective sampling and class specific data in learning the filters affects the performance of the BSIF operator. Nonetheless, the improvement is somehow dependent on the assigned task because it is affected by the the amount of information required to represent the image.

Acknowledgment

This work was supported by the Infotech Oulu doctoral program. The first author thanks Dr. Juho Kannala for sharing his code.

References

1. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), 346 – 359 (2008), similarity Matching in Computer Vision and Multimedia
2. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: BRIEF: Computing a Local Binary Descriptor Very Fast. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(7), 1281–1298 (2012)
3. Dana, K., Ginneken, B., Nayar, S., Koenderink, J.: Reflectance and texture of real world surfaces. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 151–157 (1997)
4. He, C., Ahonen, T., Pietikainen, M.: A bayesian local binary pattern texture descriptor. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. pp. 1–4 (Dec 2008)
5. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
6. Hyvärinen, A., Hurri, J., Hoyer, P.O.: Natural Image Statistics A probabilistic approach to early computational vision. Springer (2009)
7. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV) (2009)
8. Kanan, C., Cottrell, G.: Robust classification of objects, faces, and flowers using natural image statistics. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 2472 –2479 (june 2010)
9. Kannala, J., Rahtu, E.: Bsif: binarized statistical image features. In: ICPR (2012)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
11. Ojala, T., Mäenpää, T., Pietikäinen, M., Viertola, J., Kyllönen, J., Huovinen, S.: Outex - new framework for empirical evaluation of texture analysis algorithms. In: 16th International Conference on Pattern Recognition (2002)
12. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 971–987 (2002)
13. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) Image and Signal Processing, Lecture Notes in Computer Science, vol. 5099, pp. 236–243. Springer Berlin Heidelberg (2008)
14. Phillips, P., Wechsler, H., Huang, J., Rauss, P.J.: The feret database and evaluation procedure for face-recognition algorithms. Image and Vision Computing 16(5), 295 – 306 (1998)
15. Shan, H., Cottrell, G.: Looking around the backyard helps to recognize faces and digits. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1 –8 (2008)

16. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
17. Tistarelli, M., Nixon, M.S. (eds.): Advances in Biometrics, Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings, Lecture Notes in Computer Science, vol. 5558. Springer (2009)
18. Tola, E., Lepetit, V., Fua, P.: DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(5), 815–830 (May 2010)
19. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. Int. J. Comput. Vision 62(1-2), 61–81 (Apr 2005)
20. Ylioinas, J., Kannala, J., Hadid, A., Pietikäinen, M.: Learning local image descriptors using binary decision trees. In: WACV (2014)
21. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision 8(7) (2008)