Analysis of the local statistics at the centre of fixation during visual scene exploration.

Andrea Carbone
Dept. of Computer and System Sciences Sapienza, Roma (IT)
Email: carbone@dis.uniroma1.it

Fiora Pirri
Dept. of Computer and System Sciences Sapienza, Roma (IT)
Email: pirri@dis.uniroma1.it

***Abstract*—We describe how to model the statistics of both the visual input selection and gaze orienting behaviour of a human observer undertaking a visual exploration task in a given visual scenario. Evidences from neuroscience research prove that complex visual systems have shaped their receptive fields and neural organisation in a continuous adaptation to the visual environment stimuli in which they evolved. This ecology is the basis of our investigation of the human visual behaviour from a real set of gaze tracked points of fixation acquired from a custom designed wearable device. By comparing these set of fovea-centred patches with a randomly chosen set of image patches, extracted from the entire observed scene, we aim at characterising the statistical properties and regularities of the selected visual input. Samples from a human observers are collected both in free-viewing and surveillance-like tasks. In this work we suggest that a generative model of the visual input emerging from the recorded scan-path can be used to model a set of feature detectors for the design of an artificial visual system.**

I. Introduction

Modern computer vision science is strongly interested in biological vision systems, following the assumption that perceptual (biological) systems in nature are designed through natural selection; evolved optimally in response to the distribution of natural visual cues perceived from the environment. Thus the knowledge about the statistical properties of natural images are crucial when incorporated in computational framework of visual processing. Visual images bring all the information that we need to perform a task. Human beings are very efficient in a lot of visual tasks, thanks to the ability of focusing those parts of the scene relevant to the task, ignoring most of the information which is sensed, but not perceived. Before understanding the nature of visual processing, one must understand the nature of the visual environment [6], [10], [11]. Most natural visual tasks involve selecting a certain amount of locations to fixate. The spatial distribution and the wavelenght sensitivity properties of the photo-receptors on the surface of the retina is highly non uniform. Indeed, the area of finer spatial resolution is concentrated around a small spot on the retina called *foveola* of around $1.2°$, surrounded by the *fovea* with an aperture of around $6°$. Despite the relatively small area occupied by the fovea on the retinal plane, it is mapped to a disproportionate area in the visual cortex with respect to peripheral retinal areas (*cortical magnification*). The visual scanning of the world - the *scanpath* - is performed in a very efficient way by programming a sequence of *saccades* on the visual array.

The strategy followed by humans in deploying the mechanism of visual attention has been subject of research in neuroscience, cognitive science and lately computer vision. It has inspired novel biologically based methods for image compression, visual search, navigation and all the areas of research in artificial systems where a preliminary selection of the area of interest, in a restricted portion of the input, helps in reducing the complexity of a generic further processing. The principle, underlying this approach, relies on a generic notion of *visual saliency* i.e. that the visual interestingness of the scene is a measurable entity encoding the task-relevant, context-based information embedded in the visual world. In general, the saliency has been modeled as a function on some feature space computed on the image. Several approaches have been presented challenging the problem of quantifying in a biological justified framework a measure of visual salience. For example, to cite only a few of the most popular: methods inspired to the Feature Integration Theory [31] engineered to model the competition between bottom-up cues such as local measure of centre-surround contrast on feature channels (i.e. orientation, color opponency, luminance) [15] [7], or tuned to specific visual search tasks [33] or accounting for a top-down bias towards current task, spatio-temporal locations or high level cues [32].

II. Contribution

The above considerations motivate our approach to the problem of characterising the visual behaviour of an observer. The goal of this work, is to investigate into the statistical nature of the visual environment in terms of high order statistics, from a set of image patches extracted at the point of fixation of a human observer undertaking a *surveillance/free-viewing* task. The step of our approach will be separately described in Sec.IV:

a. *Sampling the visual context*: to build a set of images carrying the information of the visual content linked to a specific scenario context, Sec. IV-A;
b. *Computing linear ICA* from a set of randomly selected patches samples from the database, Sec. IV-B;
c. *Collect observations*. This step realises the actual observation of the gaze behavior of person inspecting or freely viewing the scene, Sec. IV-C;
d. *Scan path projection on the context bases*. The gaze-centred patches computed in the second step are projected on the global visual environment ICA decomposition, Sec. IV-D;
e. *Scanpath variance Analysis* in which a dimensional reduction of the problem is performed driven by the variance distribution of the scan-path coefficient projected on the ICA visual context decomposition, Sec. IV-E .

III. Related Works

The role that *central vision* [1] plays in visual processes is intimately linked to the understanding of the relationships between action, visual environment statistics and previous knowledge with the actual scan-path performed (i.e. the sequence of spatio-temporal fixations). Early experiment conducted by Yarbus [35] highlighted the influence of task on the pattern generated by an observer. Rothkopf and Ballard in [23] study the statistics at the point of gaze of human subjects involved in natural behaviours. The interesting novelty in their work is the choice of using data from a real observer, even if performing actions in a virtual visual environment. The authors argue that the active selection principle (depending on the ongoing task) is part of the process in shaping the regularities in the visual inputs. Saliency is therefore a measure of task-dependent importance. Bruce and Tsotsos [2], [3] propose a bottom up strategy relying on a definition of saliency aimed at maximising Shannon's information measure after ICA decomposition. They use a database of patches randomly sampled from a set of natural images. The saliency model is then validated against eye tracked data captured from laboratory experiment (recorded video and still images). In [21] the root mean square contrast is evaluated on a set of fixation points preformed by an observer looking at static natural (in this case natural landscapes) images. They derive a saliency model based on the minimisation of the total contrast entropy. Reinagel and Zador in [22] study the effect of visual sampling by analysing the correlations of contrast and grey-level correlation in the fovea and para-fovea[2] regions. In [34], the authors model the distribution of contrast and edges on gaze-centred image patches with a Weibull probability density function under the assumption that in a free-viewing context our gaze is drawn toward image regions which local statistics differs from the rest of the image. Tatler and Baddeley in [1] go through a deep discussion on determining what are the characteristics that most are likely influence the choice of the regions to fixate. They focus on local statistics on luminance, contrast and edges. The derived model, highlights a preference for high frequency edges. In [27] the authors observe the generic characteristic of the point of fixation conditioned to the magnitude of the saccade performed. Second order statistics regularities emerging in categories of natural images can be exploited as descriptors for classifying the kind of environment depicted [30].

IV. Experimental Scenario

Our work is closely related to the natural image statistics domain [13]. In literature, natural images or images of natural scenes are defined as: "*those that are likely to have similar statistical structure to that the visual system is adapted to during its evolution*" [8], [9], [18]. The term *natural* in our context may sound misleading as it generally refers to collection of images of natural landscapes. In our scope we consider natural images as those characterising the visual context of the observer. For example a collection of pictures of the internals of a building, or the visual landscape of a surveillance inspector.

[1] The high detailed visual information as projected on the neighbourhood of the centre of gaze during a fixation.

[2] Parafovea is the area sensed at minor resolution surrounding the fovea

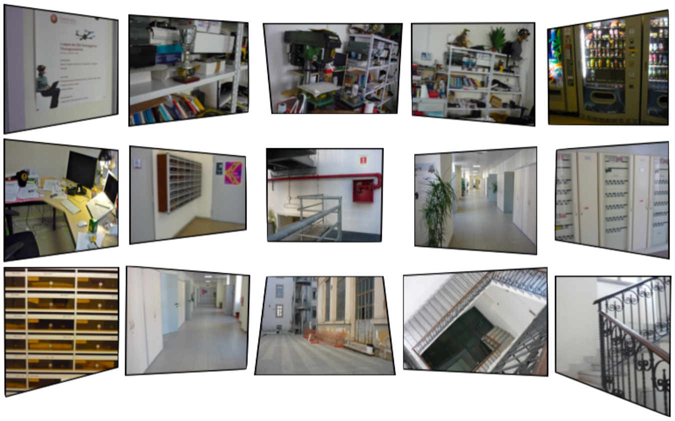

Figure 1. A collection of images sampled from the internals of the building hosting our department.

A. Sampling the visual context

The experimental setup differs from other relevant works in the field in the method of collecting data from the observer. To our best knowledge the closest methodology to our approach can be read in [24] where the fixations are collected from a real moving observer but immersed in a virtual environment which second order statistics resemble the one that can be measured in real environments[3]. The first step that we followed in order to model the experimental visual environment was to collect a set of views taken from the internals of a building. In this work we have chosen to take pictures of the *Department of Computer and System Sciences* building in Rome. We collected a set of 126 images representing the global content of visual information that people visiting or working in the building are likely to sense. The views contains sample images of different sub-contexts: corridors, rooms, laboratories, closets, doors. Pictures depicting the same sub-context were taken at different scales (i.e. from a closer or farest point of view) and different angles. Fig. 1 shows a subset of pictures selected from the database.

B. Sparse Coding and ICA

A large amount of literature deals with the concepts of sparseness, efficient coding and blind source separation. These three aspects are intimately related to each other [20]. Optimal coding is equivalent to the problem of finding a set of independent (thus uncorrelated) sources. Sparseness is a statistical property meaning that a random variable takes both small and large values more often than a normal density with the same variance. A sparse code, then represents data with a minimum number of active units. The typical shape of a sparse probability density distribution shows a peaked profile around zero and long heavy tails. The sparseness of the response of the cortical cells to the visual input [5] , suggests the adoption of a computational framework suitable to discover the latent factors that represent the basis of an alternative space for encoding properly the visual data. The aim is to use a generative model of the data. A generative model of observed data (the visual input) is thus generated as transformation of some simple original variables. The original variables are called *latent* or *hidden* since they cannot be observed directly. The generative model we use in this work is the linear independent component analysis or ICA [14]. The linear independent component analysis models linear relations between pixels. In this model, any (greylevel) image patch $\mathcal{I}(x,y)$ can be expressed as a linear combination of a basis vector $\mathcal{B}_i$ (sometimes called the *mixing matrix*):

$$\mathcal{I}(x,y) = \sum_{i=1}^{n} \mathcal{B}_i(x,y) s_i \tag{1}$$

where the s_i are called the ICA coefficients that vary from patch to patch. The s_i can be computed from the image by inverting the mixing matrix:

$$s_i = \sum_{x,y} \mathcal{W}_i(x,y)\, \mathcal{I}(x,y) \tag{2}$$

The $\mathcal{W}_i$ are called *features* or *coefficients* (because of the simple linear operation between coefficients s_i and features $\mathcal{W}_i$). An example of computed ICA bases and features can be seen in Fig.2 and Fig.3 respectively. The s_i are scalar values sparsely distributed (non Gaussian).

The coefficients $\mathcal{W}_i$ resemble the organisation of the *simple cells* in the primary visual cortex V1 [17], [19] (i.e. a set of oriented, localised, bandpass filters). The linear ICA computations presented in this work were realised with the FastICA package [12].

C. Sampling the Gaze

The sampling of the gaze is realised through a custom device that we here briefly introduce. A longer and detailed description may be found in [16]. The *Gaze-Machine* Fig. 5 is made of a helmet upon which sensors are embedded. A stereo rig and an inertial platform are aligned along a stripe mounted on the helmet. Two more cameras, that is, two microcameras *C-mos* are mounted on a stiff prop and point at the pupils. Each eye-camera provides two infrared LEDs. All cameras were pre-calibrated using the well known Zhang camera calibration algorithm for intrinsic parameters determination and lens distortion correction. Extrinsic and rectification

[3] The so called $1/f$ noise.

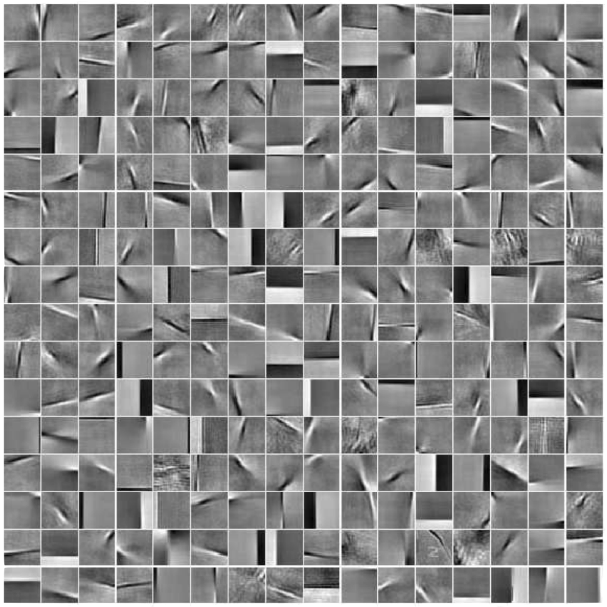

Figure 2. The set of linear ICA bases computed from a set of 25000 random patches sampled from the global visual context database.

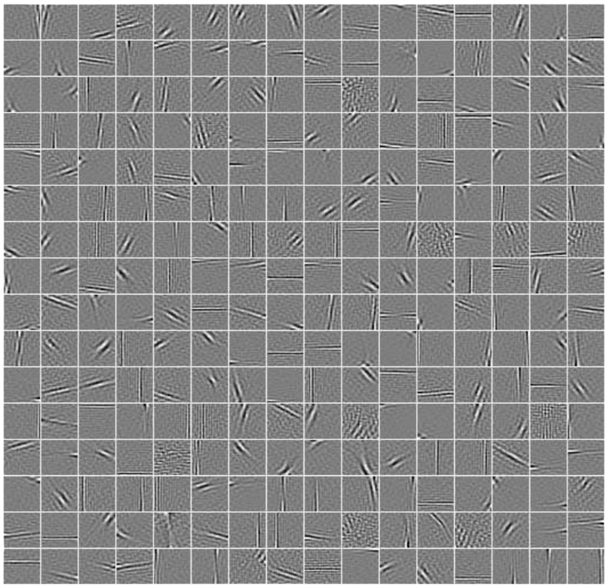

Figure 3. The set of linear ICA features computed from a set of 25000 random patches sampled from the global visual context database.

parameters for stereo camera were computed too, and standard stereo correlation and triangulation algorithms used for scene depth estimation. An inertial sensor is attached to the system, to correct errors due to involuntary movements occurred during the calibration stage. The scene camera frames are suitably paired with the eye-camera frames for pupil tracking. The data stream acquired from these instruments is collected at a frame rate of 15 Hz and it includes the right and left images of the scene, the cumulative time, the right and left images

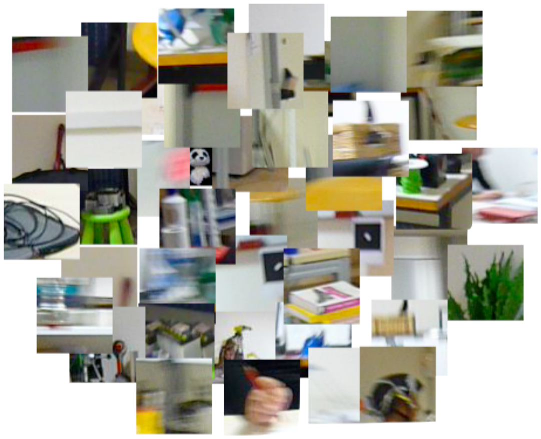

Figure 4. A subset of patches belonging the scan-path performed.

of the eyes, and the head angles, in degrees, accounting for the three rotations: the *pitch* (the chin up and down), the *roll* (the head inclined towards the shoulders) and the *yaw* (the head rotation left and right), obtained by the inertial system. In order to correctly locate the point of regard of the user, an eye-camera calibration phase, for each of the two camera-eye system, is required to estimate the multi-view relationships between the real and virtual (modeling the eye) image planes.

Each gaze tracked sample $g^{(i)}$ are acquired on each frame. The i_{th} gaze sample is defined as the n-uple:

$$g^{(i)} = \langle p^{(i)}, t^{(i)}, f^{(i)} \rangle \tag{3}$$

where $p^{(i)}$ are the $(x^{(i)}, y^{(i)})$ image plane coordinates of the gaze point, $t^{(i)}$ is the timestamp (in milliseconds) and $f^{(i)}$ the frame index. The full set of gaze samples defined as:

$$\mathcal{G} = \{g^{(1)}, g^{(2)}, \ldots, g^{(k)}\} \tag{4}$$

being k the number of samples taken. As we are interested in analysing the information sampled at the centre of gaze during a fixation we proceed to filter out from the set $\mathcal{G}$ all those samples that are likely to belongs to a saccade (the rapid eye movement between two consecutive fixations). We realise this filtering as a non-parametric clustering problem. We borrow from Duchowski the definition of fixation as a sustained *persistence* of the line of sight in time and space [4]. In practice, a fixation is the centre of a spatial and temporal aggregation of samples in a given neighbourhood. We use the *mean shift* algorithm on the feature space spanned by the vectors in $\mathcal{G}$ (except the frame index information which is not useful to cluster together samples belonging to the same fixation). A similar approach has been presented in [26]. The output of the mean shift is set of samples described by the tuple $f = \langle c, V \rangle$ where the c are the centres

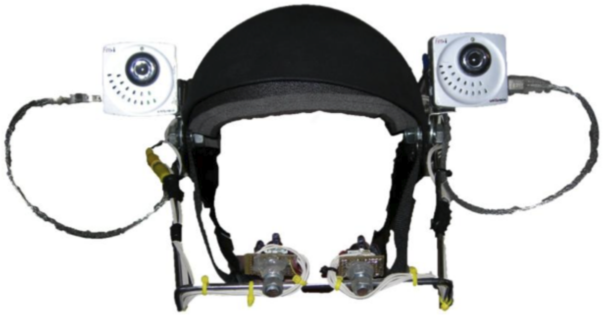

Figure 5. The gaze-machine used to acquire the scanpath.

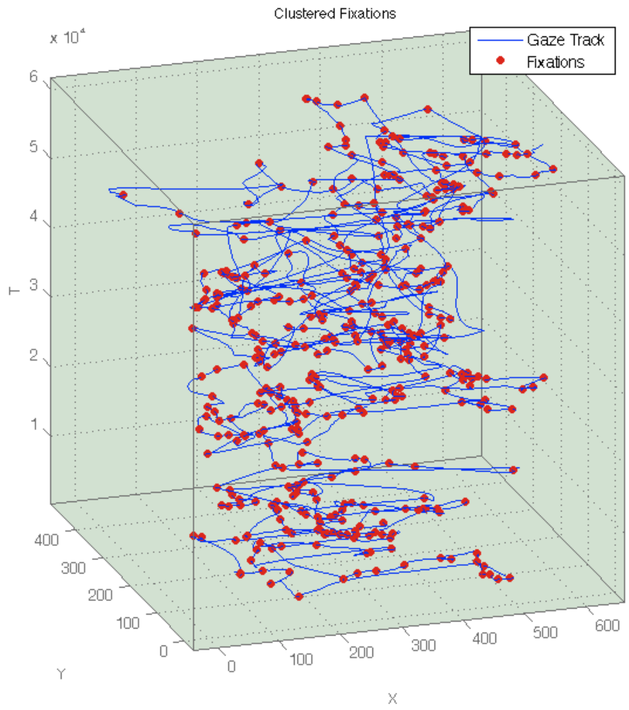

Figure 6. In the figure the mean-shift clustered fixation points (in red) superimposed on the plot of the full gaze track (continuous line). X, Y axes refers to the spatial image coordinates. The T axis represents the timestamp in milliseconds of the gaze sample.

(x, y, t) resulting from the mean shift and V is the patch centred in c. Therefore the full scanpath sequence:

$$\mathcal{F} = \{f^{(1)}, f^{(2)}, \dots, f^{(l)}\} \tag{5}$$

contains only samples classified as fixation points. Results are shown in Fig.6 and Fig.7.

Fig. 4 shows a subset of gaze-centred patches from a scan-path.

D. Scan path projection on the context bases

We have now a model of the visual context $\mathcal{C}$, defined as $\mathcal{C} = \langle \mathcal{B}, \mathcal{W} \rangle$ and the scanpath $\mathcal{F}$. In our experiment we are using (16×16) greylevel patches,

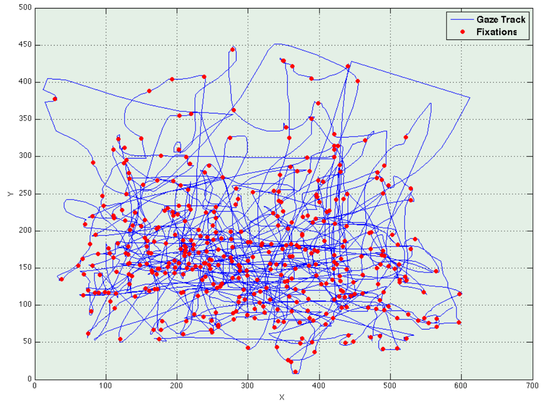

Figure 7. Same data as shown in Fig. 6 projected on the X, Y plane, discarding the time dimension.

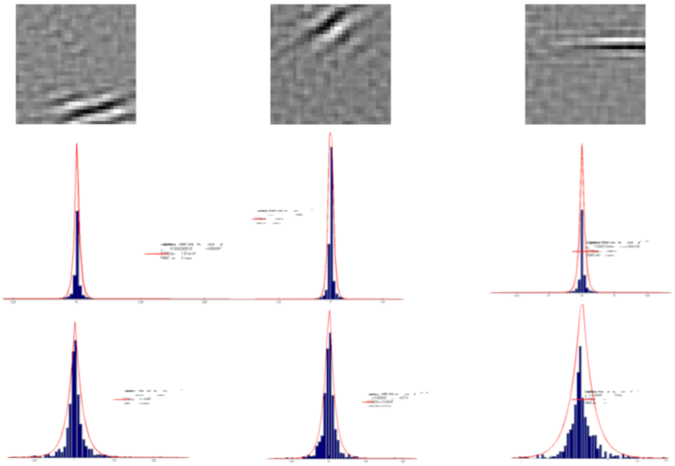

Figure 8. Observing the variance of the ICA channels coefficients. The first row, we show 3 of the 256 $\mathcal{W}$ (Fig.3). The second and third rows show the histograms of the coefficients (of the corresponding features) computed respectively on the context and the scan-path. The profile of the ML estimated Laplacian distribution is superimposed on the normalised histograms.

$\mathcal{B}, \mathcal{W} \in \mathbb{R}^{256 \times 256}$ containing on the $i_t h$ column the vectorised base B_i and feature $\mathcal{W}_i$ patches.

We have sampled the scan path with $N \geq 300$ highly foveated images (size 16×16), from this set we obtain a new feature vector of size $N \times 256$. That is, the feature vector is defined as follows. Let $V_j(x, y)$ be the j scan-path patch and $\mathcal{W}_i(x, y)$ the i-th inverted *mixing matrix*, $i = 1, \dots, 256$ and $j = 1, \dots, N$, then

$$sp_{ij} = \sum_{xy} \mathcal{W}_i(x, y) V_j(x, y) \qquad sp_{ij} \in \mathbb{R} \tag{6}$$

The sp_{ij} are the scan-path coefficients which are the weights of the bases obtained from the dot product of W_i and V_j. The resulting matrix Sp is thus a $\mathbb{R}^{N \times 256}$ matrix formed by N observations (the scan-path patches) encoded in the ICA $\mathbb{R}^{256}$ feature space .

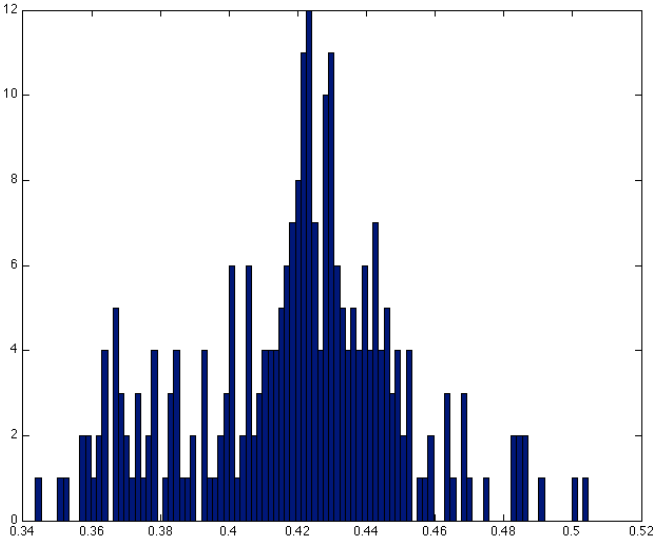

Figure 9. Average context variance distribution: $\sigma^2_{\mathcal{C}} = 0.4193$

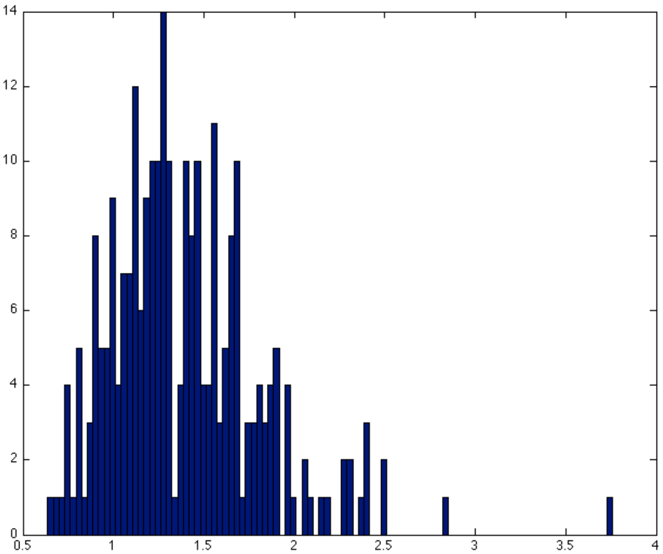

Figure 10. Average scanpath variance distribution: $\sigma^2_{S_p} = 1.3923$

Fig. 8 shows the plot of three feature channels[4] and the corresponding distribution of the coefficients encoding the context (2nd row) and the scan-path (3rd row). The coefficient distribution are then modeled with a *L*aplacian distribution:

$$P(u|\lambda) = \exp\left(\left|\frac{u}{\lambda}\right|\right) \qquad (7)$$

with λ being the variance. The examples shown in Fig. 8 highlight a different distribution of the variance between the context and the scan-path. The measured mean of the 256 context coefficients variance is around $\sigma^2_{\mathcal{C}} = 0.4193$ while the mean of the scan-path coefficients variance is $\sigma^2_{S_p} = 1.3923$ (see figures Fig. 9 and Fig. 10). The scan-path distribution has still a sparse activation, but with higher variances w.r.t. the context. We will describe in the next section our proposal for achieving both a dimension reduction of the feature space and the estimation of the probability density function characterising the scan-path.

E. Scanpath variance Analysis

As shown in the previous section, the scan path seems to be generated by less sparse data with higher variance, with respect to the context. A natural question is how the hidden context emerges from the scan path, what is the contribution of unexpressed dependences in the scan path image formation.

A latent variable model more oriented to variance, like the PCA would allow to verify the effective impact of the directions of maximal variance of the scan-path projection on the context. In particular probabilistic PCA is very suitable to study data structure by a combination of local linear principal component projection. The real difficulty is to determine the number of latent components and the number of possible mixtures.

Probabilistic PCA [28], [29] is here used to build the mixtures of principal component analysers in which, however, the number of components are randomly chosen. PPCA defines a probability model relating two sets of variables a D-dimensional vector of observations and a ρ-dimensional vector of unobserved variables. It is particular suited for high dimensional data whenever, as in our case, we need to study the variance of the projection of the scan-path. To visualise the data we have chosen a three dimensional projection for the observations $\mathbf{Y}_{tt} = 1^T \subset R^D$, of the whole generated dataset Sp, and thus the remaining dimensions, say ρ are left to the latent variables $\mathbf{z}_{tt} = 1^T \subset R^{\rho}$.

Given a Gaussian noise model $\mathcal{N}(0, \sigma^2 \mathcal{I}_\rho)$ for the latent variable $\mathbf{z}$, the marginal distribution of $\mathbf{Y}$ is Gaussian, namely:

$$\begin{aligned} &P(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \sigma^2 \mathcal{I}_\rho), \\ &P(\mathbf{Y}|\mathbf{z}) = \mathcal{N}(\mathbf{Y}|\mathcal{A}\mathbf{z} + \mu, \sigma^2 \mathcal{I}_\rho), \\ &p(\mathbf{Y}) = \mathcal{N}(\mathbf{Y}|\mu, \mathcal{A}\mathcal{A}^\top \sigma^2 \mathcal{I}_\rho). \end{aligned} \qquad (8)$$

Here $\mathcal{I}_\rho$ is the identity $\rho \times \rho$ matrix. The maximum-likelihood estimation of the $\rho \times \rho$ matrix $\mathcal{A}$ relating latent variables and observations is [28]:

$$\mathcal{A}_{ML} = \mathbf{U}_{i\rho}(\Lambda_{i\rho} - \sigma^2_i \mathcal{I}_\rho)^{1/2} \mathcal{I}_\rho. \qquad (9)$$

Here $\mathbf{U}_{i\rho}$ is the $D \times \rho$ vector formed by the eigenvectors corresponding to the greatest eigenvalue $\lambda_{i1}, \dots \lambda_{i\rho}$, of

[4] The number of the feature channel chosen is due exclusively to visualisation limits.

the original sample variance of S_p, Λ_ρ is the $\rho \times \rho$ diagonal matrix of the eigenvalues and σ is the average variance per discarded dimension. The ML estimation of the variance is [28]:

$$\sigma^2_{ML} = \frac{1}{D-q} \sum_{j=q+1}^{D} \lambda_j. \tag{10}$$

The mixture of PPCA is thus defined for the model $(\mathcal{A}_k, c_k, \mu_k, \sigma^2_k)$, $k = 1, \ldots, M$, and M the number of possible components:

$$f(\mathbf{Y}_t) = \sum_{k=1}^{M} c_k \mathcal{N}(\mathbf{Y}_t | \mu_k, \mathcal{A}_k \mathcal{A}_k^\top + \sigma^2_k \mathcal{I}_\rho), \;\; t = 1, \ldots, T; \tag{11}$$

Here c_k is the mixture proportion. The parameters of the mixture of PPCA can be approximated by the reestimation procedure given in [25], [28], [29], in which the EM approach is taken to maximise the log-likelihood of the complete-data $\mathcal{L}_C = \sum_{t=1}^{T} \sum_{k=1}^{M} w_{tk} \ln\{c_k p(\mathbf{Y}_t, \mathbf{z}_{tk})\}$ and the reestimation steps are given in [29]. Given that the search for an opimal number of components is still to be concluded, the mixture of PPCA allows to compute the variance of the latent variables and of the model:

$$C = \sigma^2 \mathcal{I} + \mathcal{A}\mathcal{A}^\top$$

and the posterior covariance is

$$\sigma^2(\sigma^2 \mathcal{I} + \mathcal{A}^\top \mathcal{A})^{-1}$$

From this, which is the variance of the latent data, given the observed, we have been able to evaluate the dependency of the scanpath from the context.

V. Conclusion

We showed an approach aimed at modeling the visual selection of an generic observer from a real scan-path performed in a given environment. The rich information content from the visual environment is encoded in a set of feature bases which capture the linear correlations between the images. Then we project the actual scan-path onto the feature space representing the context. The last step is to estimate a reduced problem in which the subspace spanned by the scan-path is modeled as a mixture of latent factors by PPCA. This approach is appealing, because puts together the visual context surrounding the observer and the learning of a selection scheme from a recorded scan-path. This approach seems valid, though a clear way for determining the number of mixtures according to the generated scan-path is still to be defined..

References

[1] R. Baddeley and B. Tatler, "High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis." *Vision Research*, vol. 46, pp. 2824–2833, Jan 2006.

[2] N. Bruce and J. K. Tsotsos, "An information theoretic model of saliency and visual search." *Lecture Notes in Computer Science*, vol. 4840, p. 171, 2007.

[3] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach." *J. Vis.*, vol. 9, no. 3, pp. 1–24, 2009.

[4] A. Duchowski, *Eye Tracking Methodology: Theory and Practice*. Springer, 2007.

[5] D. Field, "What is the goal of sensory coding?" *Neural Computation*, Jan 1994.

[6] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells." *J Opt Soc Am A*, vol. 4, no. 12, pp. 2379–2394, Dec 1987.

[7] S. Frintrop, M. Klodt, and E. Rome, "A real-time visual attention system using integral images." *Proc. of ICVS*, 2007.

[8] W. S. Geisler, "Visual perception and the statistical properties of natural scenes." *Annu. Rev. Psychol.*, vol. 59, pp. 167–192, 2008.

[9] W. S. Geisler and D. Ringach, "Natural systems analysis." *Visual neuroscience*, vol. 26, no. 1, pp. 1–3, 2009.

[10] J. J. Gibson, *Perception of Visual World.*, 1966.

[11] ——, *The Senses Considered as Perceptual Systems.*, Jun 1983.

[12] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis." *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[13] A. Hyvarinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics A Probabilistic Approach to Early Computational Vision.*, 2009, vol. 39.

[14] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications." *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[15] 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 Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] S. Marra and F. Pirri, "Eyes and cameras calibration for 3d world gaze detection." 2008, pp. 216–227.

[17] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." *Nature*, vol. 381, no. 6583, pp. 607–609, Jun 1996.

[18] ——, "Natural image statistics and efficient coding." *Network: Computation in Neural Systems*, vol. 7, no. 2, pp. 333–339, 1996.

[19] B. A. Olshausen and D. Fteld, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Research*, Jan 1997.

[20] A. Pece, "The problem of sparse image coding." *Journal of Mathematical Imaging and Vision*, Jan 2002.

[21] R. Raj, W. Geisler, R. Frazor, and A. Bovik, "Natural contrast statistics and the selection of visual fixations." *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 3, pp. III – 1152–5, Sep 2005.

[22] P. Reinagel and A. Zador, "Natural scene statistics at the centre of gaze." *Network: Computation in Neural Systems*, vol. 10, no. 4, pp. 341–350, 1999.

[23] C. A. Rothkopf and D. H. Ballard, "Image statistics at the point of gaze during human navigation." *Visual neuroscience*, vol. 26, no. 01, pp. 81–92, 2009.

[24] C. A. Rothkopf, D. Ballard, and M. Hayhoe, "Task and context determine where you look." *Journal of Vision*, vol. 7, no. 14, p. 12, 2007.

[25] S. T. Roweis, “Em algorithms for pca and spca,” in *NIPS*, 1997.

[26] A. Santella and D. DeCarlo, “Robust clustering of eye movement recordings for quantification of visual interest.” *ETRA ’04: Proceedings of the 2004 symposium on Eye tracking research & applications*, Mar 2004.

[27] B. Tatler, R. Baddeley, and B. Vincent, “The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task.” *Vision Research*, vol. 46, no. 12, pp. 1857–1862, 2006.

[28] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” *Journal of the Royal Statistical Society, Series B*, vol. 61, pp. 611–622, 1999.

[29] ——, “Mixtures of probabilistic principal component analysers,” *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[30] A. Torralba and A. Oliva, “Statistics of natural image categories.” *Network: Computation in Neural Systems*, vol. 14, no. 3, pp. 391–412, 2003.

[31] A. M. Treisman and G. Gelade, “A feature-integration theory of attention.” *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[32] J. Tsotsos, S. Culhane, W. K. Wai, Y. Lai, N. Davis, and F. Nuflo, “Modeling visual attention via selective tuning.” *Artificial intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.

[33] J. M. Wolfe, K. R. Cave, and S. L. Franzel, “Guided search: an alternative to the feature integration model for visual search.” *Journal of experimental psychology. Human perception and performance*, vol. 15, no. 3, pp. 419–433, Aug 1989.

[34] V. Yanulevskaya, J.-M. Geusebroek, J. B. C. Marsman, and F. W. Cornelissen, “Natural image statistics differ for fixated vs. non-fixated regions.” vol. 37, 2008.

[35] A. L. Yarbus, *Eye Movements and Vision.*, 1967.