LOW RESOLUTION IMAGE SAMPLING FOR PATTERN MATCHING*

R. Brunelli
ITC-irst
Povo di Trento, ITALY
brunelli@itc.it

Abstract The paper presents a simulated mobile system that learns to solve the ego-location task in a known environment, in a supervised way, using a very low resolution sampling of the optical array and RBF approximation techniques. The impact of the number of sensors, of their layout, in particular of Sobol sequences with respect to regular grids for a progressively refined sampling of images, and of the complexity of response of each sensing unit has been investigated in an attempt to simplify as much as possible the architecture of the image processing module retaining good localization ability.

Keywords: Appearance based recognition, statistical object representation, machine learning, radial basis function networks, graphical simulator.

1. Introduction

The investigation of simple architectures capable of visual competence holds promise of further insight into the fundamental mechanisms underlying animal perception. It is from this perspective that the problem of associating a view of a known environment to the position of the observer and its gaze direction has been approached. The specific variation of the task analyzed requires the development of a vision processing module for museum environments (see Figure 1) capable of understanding which area of a painting the user is targeting the camera at, in order to provide contextualized information to the visitor. The proposed solution is based on the learning-by-example paradigm: several images, annotated with the required parameters are fed to the system that learns the appropriate mapping from them to painting coordinates. The results obtained on image sampling and learning architectures can be generalized to the development of image retrieval systems. Specifically, more compact image

*Research partially funded by Provincia Autonoma di Trento under Project PEACH.

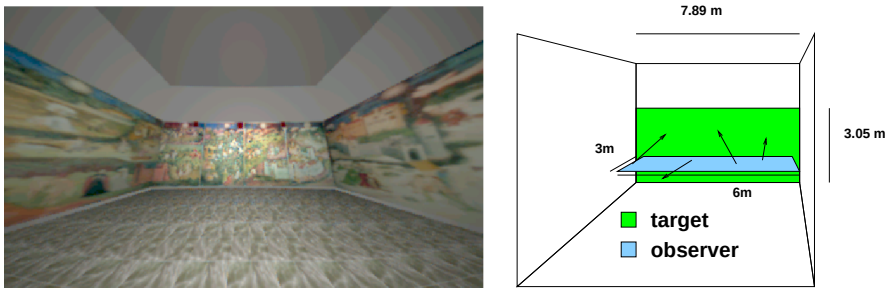

Figure 1. A wide angle view of the simulated environment where experiments are performed (left) and the four degrees of freedom of the observing camera (horizontal motion in a fixed height plane and bidimensional pointing without roll at the painting).

signatures can be devised which prove useful for very large databases applications. The organization of the paper follows a three stage workflow presenting *the imaging*, *the sensing*, and *the learning* processes, along which data are generated, gathered and interpreted by the system developed. Conclusions and future work are covered in the final section.

2. The imaging process

A major problem in the development of systems based on learning is the availability of a sufficient number of training examples. In order to gather enough data to support the exploration of different learning architectures and different sensing strategies, a flexible graphical engine has been realized following the ideas proposed in Bertamini et al., 2003. The tool supports the efficient simulation of several characteristics of real optical systems (distortion, vignetting, digital artifacts, limited resolution and focusing ability) relying on a set of post-processing functions acting on a synthetic image generated by a 3D rendering system. Distortion curves and vignetting profiles are used to build deformation maps and compositing masks speeding up the image generation process. Depth of fields effects are realized efficiently by spreading the value of each pixel over its corresponding *circle of confusion* whose diameter depends on lens focal length, aperture and point distance, the latter available with no additional computational cost from the ray tracing system employed. The simulation of the depth of field effects is important in the task considered as the related blurring, being depth dependent, could be exploited for ego location by the observing system. The implementation of digital mosaicking and color quantization effects enables the simulation of low quality digital sensors of the kind expected on cheap palmtop cameras. All the effects can be turned on/off for the images generated by the graphical simulator, providing data of different quality to test the sensitivity of the vision algorithms developed.

3. The sensing process

Visual hyperacuity, the perception of a difference in the relative spatial localization of visual stimuli with a precision exceeding the characteristic size of the sensors, is a task dependent ability improvable by training. It has been demonstrated in Snippe and Koenderink, 1992, that population coding of spatial signals, sampled by units with overlapping fields, supports hyperacuity performance. Based on the insight of Poggio et al., 1992, and the ideas of Schiele and Crowley, 2000, an artificial visual system has been realized sampling the incoming optical array through a set of *receptive field* histograms describing the local distribution of several image features. Large receptive fields are more likely to overlap across different snapshots of the world, possibly supporting increased performance of the ego location system via a mechanism similar to the one underlying hyperacuity. However, maximizing the size of the receptive fields by letting them cover the whole image may be less convenient than sampling multiple, smaller areas as in the latter case the image processing module could be trained by blinding specific sensors subsets to cope with occlusions. Each image signature would then become more discriminative including spatial information into the overall statistical description provided by the histograms. Furthermore, image normalization via histogram equalization to cope with variation of light intensity and color would make single luminance (or color channel) histogram meaningless. Three different image features have been considered: hue weighted by saturation and the energies of the first and second derivatives of image luminance convolved with a regularizing Gaussian kernel. Two different sensor responses have been computed: the integral of the sensed quantity over the receptive field, and a vector-like output, represented by an 8 bin histogram. The detailed response of the differential operators is given by the integrated response along different directions, describing the local edge structure following Freeman and Adelson, 1991. Two sensor layouts have then been compared: a grid structure with no overlapping and a random sequence of overlapping sensors whose positions are derived from a Sobol pseudo random sequence as presented by Press et al., 1992, whose points maximally avoid each other. Retinas with different sensor overlapping have been generated by varying d

$$s = \frac{1}{2}\left(\frac{1}{i}\right)^{d} \tag{1}$$

where i is the position of the sensor in the sequence and s its normalized size (see Figure 2). The value $d = 0.5$ corresponds to the half side of a square sensor covering $1/i$-th of the imaged area, while $d = 0.25(0.75)$ correspond to a slower (faster) decay of size within the sequence, resulting in increased (decreased) overlapping of the receptive fields. Progressive reduction of sensor fields is well matched to the characteristic of the Sobol sequence, providing

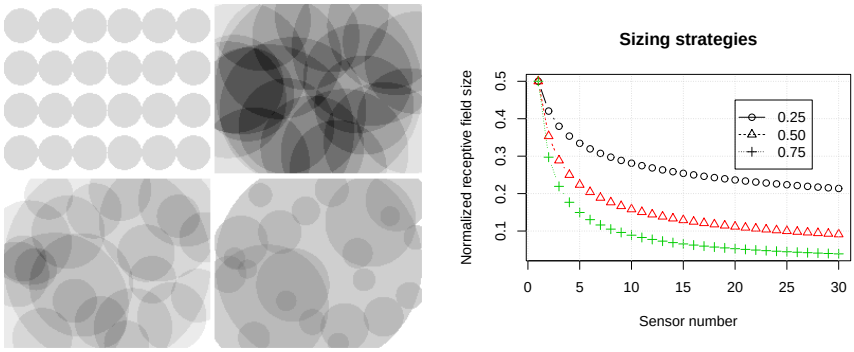

Figure 2. Regular grid and progressive resolution sensors with a Sobol layout (left) with different size decay curves (right).

an effective coarse to fine sampling strategy upon which more sophisticated sensor polling strategies could be based. Training data have been generated by the developed graphical simulator: 10000 low resolution, distorted, vignetted, digitally sampled images (using a Bayer sensor mosaic) with depth of field at a typical lens fstop (5.6), fixed focusing distance ($2m$) as well as autofocusing. Camera positions and target directions have been chosen following a 4-dimensional Sobol sequence progressively sampling the two colored rectangles in Figure 1. The unit of the spatial values reported in the plots is $1m$.

4. The learning process

Regularization Networks, introduced by Poggio and Girosi, 1990, are a tool for multivariate function approximation and include as a special case Radial Basis Functions networks. A scalar function can be approximated from a sparse set of points $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\ldots,N}$, by an expansion in radial functions:

$$\boldsymbol{F}(\boldsymbol{x}) = \sum_{\alpha=1}^{K} \boldsymbol{c}_\alpha g(\| \boldsymbol{x} - \boldsymbol{t}_\alpha \|_{\boldsymbol{W}}) \tag{2}$$

with $\{\boldsymbol{t}_\alpha\}$ a proper subset of the available data, the expansion centers, $\boldsymbol{W}$ defining the metrics, $\boldsymbol{x}$ a vector obtained by concatenation of the available histograms and $\boldsymbol{y}$ the position and gaze of the observer. The computation of $\{\boldsymbol{c}_\alpha\}$ corresponds to a least-square estimation problem:

$$\boldsymbol{c} = \boldsymbol{G}^{+}\boldsymbol{y} \tag{3}$$

where $\boldsymbol{G}_{ij} = g(\| \boldsymbol{x}_i - \boldsymbol{x}_j \|)$ and $\boldsymbol{G}^{+}$ is the Moore-Penrose pseudo inverse $\boldsymbol{G}^{+} = (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T$ whose dimension is fixed by the number of centers K. Several choices are available for g, the Gaussian and the multiquadric being

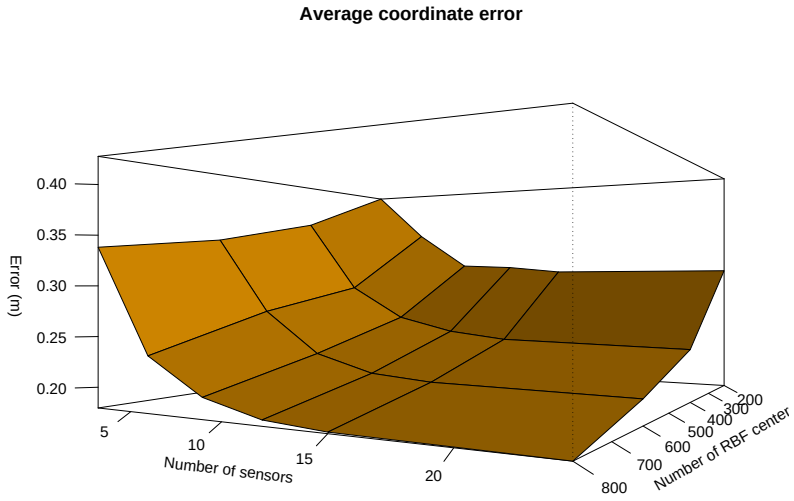

Figure 3. Average error over the 6 coordinates of the observer for a Gaussian RBF network with varying number of centers and of sensors. Good performance can be obtained with as few as 12 sensors and 400 centers (examples) even when using averaged sensor responses.

among the most popular, with the former being chosen for the present experiments due to the similarity of its localized support to the properties of receptive fields. The choice of $\boldsymbol{W}$ is important as it defines the *overlapping* of the network units, represented by the Gaussian functions centered at the expansion centers $\boldsymbol{t}_\alpha$. Good results have been obtained by choosing $W_{ij} = \delta_{ij}/d_M$, where d_M represents the maximum cluster radius obtained when setting the expansion centers by clustering. Hyperacuity-like advantages are expected to derive from the joint action of large receptive fields of the histograms, providing data changing slowly with the position of the observer and sustained response of the network units due to small values of $\boldsymbol{W}$. The resulting networks exhibit good performance even with integral sensor responses (Figure 3). No advantages have been found for fixed focus sensing over autofocusing. The experimental results show a convincing advantage of overlapping Sobol layouts with respect to a non overlapping grid structure (Figure 4), favoring large receptive fields.

5. Conclusions

Artificial retinas built from largely overlapping sensors responsive to different local visual features have been proved to be an economical yet rich enough representation for solving the ego location problem in a known environment.

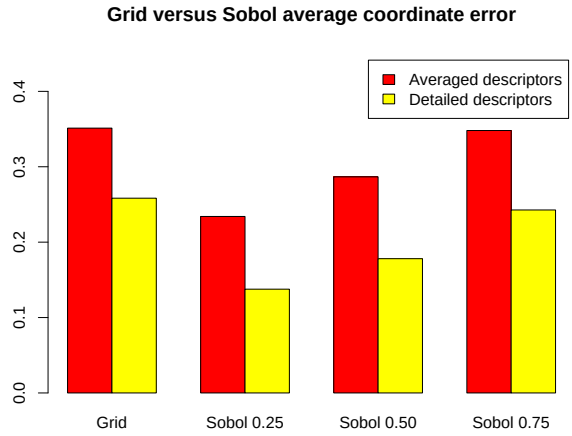

Figure 4 Average error over the 6 coordinates of the observer for Gaussian RBF networks with 400 centers, 24 sensors and different inputs. Note the significant advantage of the proposed Sobol sampling.

The impact of the number of sensors, of their layout and complexity of response has been investigated in an attempt to simplify as much as possible the architecture of the image processing module. The use of Sobol sequences for a progressively refined sampling of images can be of wider interest, being applicable to the development of more efficient search strategies in image retrieval applications based on the query-by-example paradigm such as those mentioned in Brunelli and Mich, 2000. Future work will extended the results obtained so far using different learning techniques (e.g. Support Vector Machines), different feature sets targeted at gray level images and occlusion management by training several modules for specific occlusion patterns.

References

Bertamini, F., Brunelli, R., Lanz, O., Roat, A., Santuari, A., Tobia, F., and Xu, Q. (2003). Olympus: an ambient intelligence architecture on the verge of reality. In *Proc. of ICIAP 2003*, pages 232–237, Mantova, Italy.

Brunelli, R. and Mich, O. (2000). Image Retrieval by Examples. *IEEE Transactions on Multimedia*, 2(3):164–171.

Freeman, W. T. and Adelson, E. H. (1991). The design and use of steerable filter. *IEEE Transactions on PAMI*, 13(9):891–906.

Poggio, T., Fahle, M., and Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, 247:1018–1021.

Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, 2nd edition.

Schiele, B. and Crowley, J.L. (2000). Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50.

Snippe, J. and Koenderink, J.J. (1992). Discrimination thresholds for channel-coded systems. *Biological Cybernetics*, 66:543–551.