ICIAR - Int'l Conference on Image Analysis and Recognition, Porto, Portugal, Sep.29 - Oct.1 2004

Appearance-based object detection in space-variant images: a multi-model approach*

V. J. Traver[1], A. Bernardino[2], P. Moreno[2], and J. Santos-Victor[2]

[1] Dep. de Llenguatges i Sistemes Informàtics, Universitat Jaume I, Castelló (Spain)
[2] Instituto Superior Técnico, Instituto de Sistemas e Robótica, Lisboa (Portugal)
vtraver@uji.es, {alex,plinio,jasv}@isr.ist.utl.pt

Abstract. Recently, log-polar images have been successfully used in active-vision tasks such as vergence control or target tracking. However, while the role of foveal data has been exploited and is well known, that of periphery seems underestimated and not well understood. Nevertheless, peripheral information becomes crucial in detecting non-foveated objects or events. In this paper, a multiple-model approach (MMA) for top-down, model-based attention processes is proposed. The advantages offered by this proposal for space-variant image representations are discussed. A simple but representative frontal-face detection task is given as an example of application of the MMA. The combination of appearance-based features and a linear regression-based classifier proved very effective. Results show the ability of the system to detect faces at very low resolutions, which has implications in fields such as visual surveillance.

1 Introduction

The combination of space-variant images and active-vision systems represent a biologically plausible approach to reduce the complexity of visual tasks. Attentional mechanisms [5] allow potential-interest objects be detected in periphery, so that the fovea can be directed to the selected object for its fine-detail inspection. As a remarkable space-variant image model, log-polar images have a central fovea with a very high resolution, which decreases with the eccentricity. The size of these images is given by $R \times S$ (the number of rings R and sectors S in which the original cartesian space is sampled). The particular log-polar model used in this work leaves a central circle unmapped (the *blind spot*). In an example of log-polar transformation (Fig. 1), it is worth noticing the important data reduction achieved, by comparing the sizes of images in Fig. 1(b) and Fig. 1(c).

Most of past work on visual attention has focused on salience computation [6] in static images, and scarce work has been done in active-vision and foveal systems. While the benefits of the high-acuity fovea and the implicit focus of attention of log-polar images have been exploited in some active-vision problems in the past [1], the role of coarse-resolution peripheral information has

* Research partly funded by grants E-2003-03 from *Fundació Caixa-Castelló Bancaixa*, European IST 2001 37540 (CAVIAR project), and CTIDIB/2002/333, from *Conselleria de Educació, Cultura i Ciència, Generalitat Valenciana*.

(a) (b) (c) (d)

Fig. 1. Log-polar mapping: (a) grid layout example (10×16); (b) original cartesian image (256×256); (c) cortical image (32×64); (d) retinal image (256×256) reconstructed from (c) by the inverse mapping.

been mostly neglected. However, for active-vision systems to get a real benefit of space-variant vision, such as a significant data reduction, *both* foveal and peripheral information needs to be considered appropriately.

In this paper, we focus on the problem of model-based detection of objects across the field of view (top-down attention [9]). Since the appearance of imaged objects changes significantly with their position, we introduce a multi-model approach, and illustrate it with the important problem of human-face detection. While computers and robots interacting with people require this kind of social ability, past research only considered uniformly-sampled images. Interesting results have been achieved [11], but the problem remains very difficult due to the complexity and wide variability in human faces and environmental conditions. In addition, further challenges arise when faces are to be detected within space-variant images. Therefore, the aim of the paper is *not* to propose a face detector outperforming existing systems, but to suggest a framework for object detection in space-variant images.

It is worth stressing that the problem of face detection *within* log-polar images should not be confused with the use of log-polar mapping as a *tool* for face detection/recognition. The former is the problem considered here, and, to the best of our knowledge, it has only been studied before in [7]. A few more works exist on the latter.

Section 2 describes our proposal: the multiple-model approach. For face representation/detection, a PCA-based technique has been studied (Sect. 3), and some classifiers have been tried (Sect. 4). Results are later presented (Sect. 5), before the final discussion (Sect. 6).

2 Multiple spatial models

In practice, top-down attention consists of using some *a priori* model of a target being searched. While just a single model is usually considered, when it comes to space-variant sensing, it makes sense to have multiple models. In the case of log-polar images, due to its lack of translation invariance and its varying resolution, the appearance of an object is non-linearly distorted in different parts of the visual field (Fig. 2). Therefore, we propose to have M models of the target (faces here). Each model $i \in \{1, \ldots, M\}$ is defined by a set of image positions $\mathcal{L}_i$, and a

set of features $\mathcal{F}_i$ describing the target as viewed at $\mathcal{L}_i$. Thus, instead of having a single model representing the target and distort it *on-line* while searching for instances of it, the target is mapped *off-line* to a set of locations and a multiple-model representation of that target is built. We argue that, for space-variant images, a multiple-model approach (MMA) have significant advantages over a single-model approach (SMA):

- It is ***intuitive***. Because resolution is different at different sensor locations, it makes sense to have different models at different spatial positions.
- It offers a ***natural*** solution to a number of issues. Particular conditions such as the central blind spot, varying resolution between and within models, targets partially visible, etc., represent no problem at all, because targets imaged at different locations are never compared one to each other.
- Target detection can be as ***efficient*** in MMA as in SMA. Model acquisition usually requires a learning stage. Under MMA, each model requires its own learning process, but this is only at *off-line* time, while the *on-line* detection stage can proceed as fast as, or faster than, in SMA.
- It exploits ***data reduction***. Peripheral models can benefit from the fact of targets being imaged with fewer log-polar pixels. This is in clear contrast with the approach in [7], where a SMA is adopted. In SMA, for the feature set of all models to be directly comparable, they *all* must have the same length, which, in turn, requires image data oversampling. As a result, most feature sets are bigger than strictly necessary, and contain redundant information.

Examples of targets at different model locations are shown in Fig. 2. Notice that faces further away from the center take up less log-polar pixels. The MMA makes full sense in active vision scenarios. For example, because only a discrete number M of models can be considered, the detection of faces at positions different to $\mathcal{L}_i$ requires the collaboration of purposeful movements of a robotic head. With such motions, the same set of models $\{\mathcal{L}_i, \mathcal{F}_i\}$ can be reused, but each time observing different parts of the scene, where potential faces might be. Thus, a larger number of *virtual models* are possible under an active vision setup.

It is worth stressing that the application to face detection is just an example, and the multi-model framework is perfectly suited to detecting general visual classes of objects. Similarly, while PCA is used here for illustrating an actual application of MMA, a variety of techniques are possible under the MMA. Finally, even though we are particularly interested in log-polar images, the idea of MMA is easily applicable to other space-variant models and, more generally, whenever image distortions may happen (omnidirectional images, fish-eye lenses, etc.).

3 Principal Component Analysis (PCA) under MMA

PCA is a well-known technique allowing a high-dimensional space (e.g., images) be represented in a low-dimensional one (the eigenspace)[3]. The *eigenface* technique (PCA applied to faces) has been used since the early 90s [10], mostly for

[3] A discussion on PCA *vs.* ICA, which is beyond the purpose of this paper, can be found, e.g., in [2].

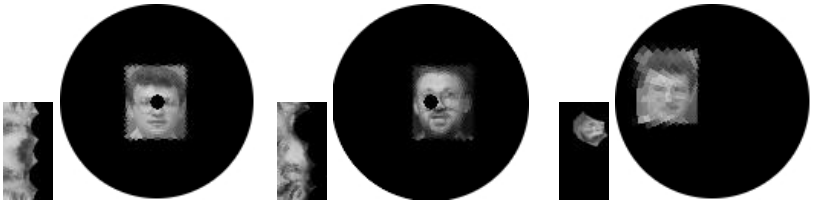

Fig. 2. Multiple models: example of human faces as viewed at different spatial locations

face recognition and, much less, for face detection. One of the things that should be decided when using PCA is how many eigen vectors k to use: the smaller k, the more compact the representation, but the worse the samples in the training set (TS) can be approximated from the eigen vectors. Several heuristics have been suggested to choose k. Consider the eigen values $\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$, with n the size of the TS, and let $V(k) = \sum_{i=1}^{k} \lambda_i$. Here, we choose the smallest k such that $\frac{V(k)}{V(n)} > \theta_\lambda,\ \theta_\lambda \in [0, 1]$. Thus, the higher θ_λ, the more variability in the TS is accounted for [8].

Under the MMA, for each model i, a rectangle of the size of the face subimage is placed over a log-polar grid at an eccentricity ρ_i and an angularity θ_i, so that the set of log-polar image positions $\mathcal{L}_i = \{l_j = (u_j, v_j),\ j = 1, \ldots, m_i\}$ underlying this rectangular region is found. Then, for each model i, a feature vector $\mathbf{I}_s^i$ is computed from the gray-level values of each face image in the TS, I_s, $s \in \{1, \ldots, n\}$, being mapped to a log-polar image A, and a function f being applied: $\mathbf{I}_s^i = \{f(A(l_j)),\ l_j \in \mathcal{L}_i\}$. The function f weights the gray-level values taking into account the size of receptive fields and to attenuate background information present in the (non-segmented) face database. Finally, each feature set $\mathcal{F}_i$ encapsulates the first k_i eigenfaces computed from all $\mathbf{I}_s^i$.

If we consider $V(k)$ for models at different eccentricities ρ, we find that more peripheral models require less eigen vectors to account for the same variability θ_λ (Fig. 3). This makes sense because a set of faces observed at coarser resolution look like more similar between them (there is less variability in the set). Therefore, peripheral models are computationally efficient, not only because the reduced number of log-polar pixels they occupy (i.e., $m_i < m_j$ for $\rho_i > \rho_j$), but also because fewer eigenfaces are needed (i.e., $|\mathcal{F}_i| < |\mathcal{F}_j|$ for $\rho_i > \rho_j$).

Notice that the proposed PCA-based approach implies a learning mechanism to discover a face model for each location. The suitability of a learning-based scheme to represent and detect faces in space-variant images is further confirmed by recent evidence showing the superiority of learning over mathematical models heuristically defined [3].

To measure the distance of an input image to face space, two distances (or a combination of them) have typically been used: the distance *from* feature space and the distance *in* feature space. In theory, we could use this *faceness* for face/non-face classification, but there is one important problem: the need

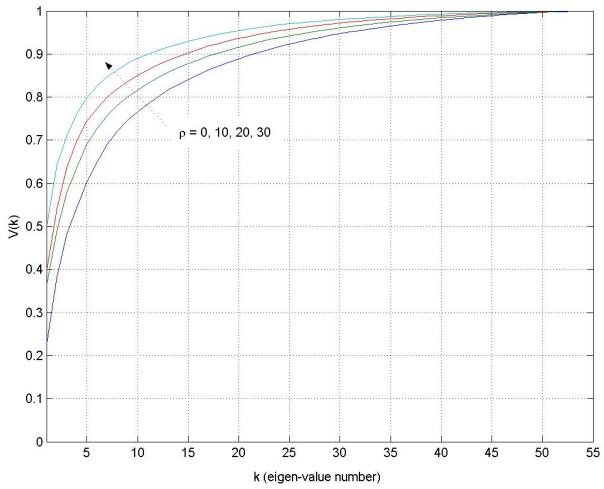

Fig. 3. $V(k)$ for different eccentricities ρ

to set distance thresholds, a generally tricky and undesirable task. Next section proposes an effective alternative strategy overcoming these difficulties.

4 Exemplar-based face/non-face classification

Because of the drawbacks of PCA-based distances, and our negative experience using them, we looked into different ways to approach the problem. One interesting idea was having two sets, each representing the face and non-face classes, so that the system is better at discriminating between them. The problems now are that examples of both classes are needed, and that the non-face class has a huge variability and can be difficult to represent compactly. Even with these disadvantages (which, otherwise, are common in many techniques [11]), we believe this approach is preferable to having to choose thresholds.

As in any classification task, two things have to be defined: the features representing the samples and a classifier. As for the features, we have tried several possibilities, but in this paper we report on the one using the projections to face space as the feature vectors. Regarding the classifier, several alternatives have also been explored, but here we use the *Linear Regression of an Indicator Matrix* (LRIM) [4], a technique we found simple yet effective.

In LRIM, the p features of N training instances belonging to one of K classes (here, $K = 2$) are given in the *model matrix*, $\mathbf{X}_{(N\times(p+1))}$, with one leading column of 1's. The *indicator response matrix*, $\mathbf{Y}_{(N\times K)}$, has N rows of K indicators $y_k,\ k = 1,\ldots,K$, with $y_c = 1, y_j = 0, j \neq c$ to indicate class c. Thus, $\mathbf{Y}$ is a matrix of 0's and 1's, with a single 1 per row, representing the class of feature vector in the corresponding row in $\mathbf{X}$. With these two matrices as input, a *coefficient matrix* is estimated as $\hat{\mathbf{B}}_{((p+1)\times K)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

Fig. 4. Examples of images

PCA	LRIM clusters		Test set	
Training set	Faces	Non-faces	Faces	Non-faces
130	64	64	120	190

Table 1. Number of images used

To classify a new observation (a p-feature vector) $\mathbf{x}$, the fitted output $\hat{\mathbf{f}}$ is computed as $\hat{\mathbf{f}} = ([1\ \mathbf{x}] \cdot \hat{\mathbf{B}})^T$, and its largest component, identifies the predicted class c to which $\mathbf{x}$ belongs: $c = \arg\max_k \hat{\mathbf{f}}_k(\mathbf{x})$. We note that, for classification efficiency, it is valuable the fact that the size of $\hat{\mathbf{B}}$ depends on the number of features, but *not* on the number of samples in $\mathbf{X}$. And because the feature vector in our case is small (projections in eigenspace), the choice of LRIM as a classifier turns out to be particularly convenient when used in conjunction with PCA.

5 Experimental results

In the experiments, we used the AT&T face database (`www.uk.research.att.com/facedatabase.html`), consisting of 400 images (40 individuals, each at 10 different approximately frontal views). We collected 190 non-face images for the LRIM stage and for testing purposes. Some of the non-faces images are upside-down faces (which are not in the *face* class, as meant here). This helps the system classify test images that could otherwise be misclassified. Examples of face and non-face images are shown in Fig. 4.

The distribution of images in different stages of the algorithm is summarized in Table 1. The face cluster for LRIM was a subset of the TS used in PCA. The test set included faces of subjects not included in the TS, as well as different views of faces of subjects in the TS. Regarding the log-polar images, the considered size was as small as $R \times S = 32 \times 64$.

With all the elements comprising this experimental setup, there are many and interesting issues that deserve exploration. The influence on the performance of some factors is plotted in Fig. 5, where, besides recognition error, recognition rates for faces and non-faces are shown, so that the number of false positives and false negatives can readily be perceived. On the other hand, more important than the actual rates, it is the performance *trend* when varying these factors.

In the first experiment, we used the central model ($\rho = 0$), and the number of eigenfaces was decreased by changing θ_λ, which accounts for the variability in the training set. For these θ_λ, the number of eigen vectors were 29, 10, 4 and 2, respectively. The less eigen vectors, the faster the classification can be, but as the results reveal (Fig. 5 (top)), the representativeness of a face diminishes, and so the recognition rates. It is worth noticing how, with as few as 10 eigen vectors, recognition is reasonable good, which points to the appropriateness of PCA as a technique to represent face patterns economically.

Next, the effect of eccentricity is analyzed. We kept θ_λ fixed to 0.7, while the training and test images were placed at increasing distance from the center, ρ, thus loosing visual acuity with eccentricity. The classification results, shown in Fig. 5 (bottom), are somehow surprising and unexpected: there is no significant decay in performance. The possible explanation of this seemingly counter-intuitive result is that the important, discriminative features in the face pattern can be observed at sufficient detail even at very coarse resolution. This somehow reminds us the amazing human ability of spotting human faces even in adverse conditions, and points to the key role that the periphery in log-polar images may play in attentional tasks.

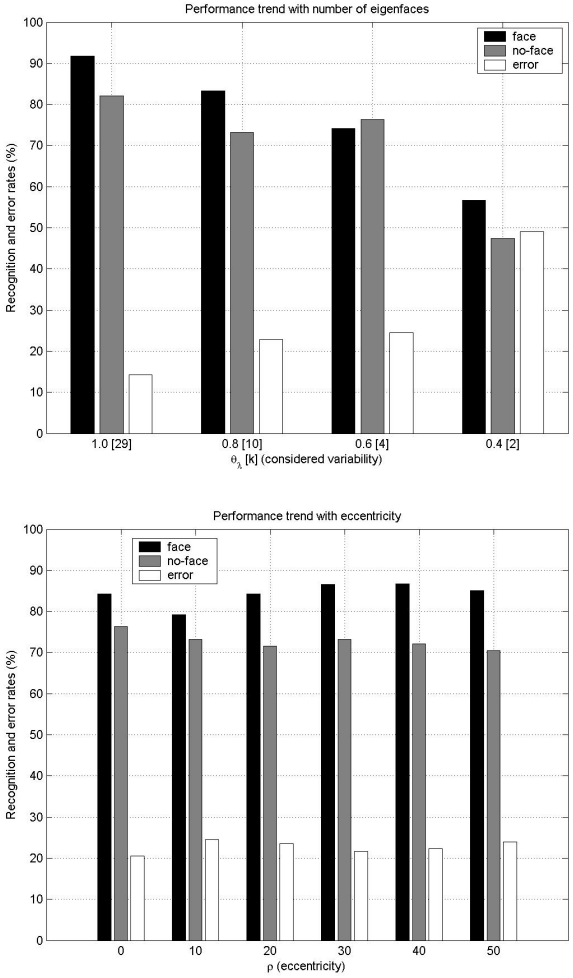

Fig. 5. Performance trends (see text for details)

6 Conclusions

A multiple-model approach (MMA) has been proposed as a general framework for modeling classes of objects in the context of space-variant images and active vision setups. The motivation behind this is the use of log-polar images for top-down visual attention tasks, and the exploration of the role that peripheral information can play. While MMA is a general approach, it has been illustrated with an application to the problem of face detection in log-polar images.

PCA has been used within MMA (i.e., multiple PCAs) as a technique for an efficient representation of a face class (one face class per spatial model), and has been shown to be effective, although quite sensitive to out-of-class variations. To overcome the encountered limitations of existing PCA-based distances, a simple but effective exemplar-based strategy to classify input patterns into face/non-face classes has been provided. Importantly, the use of LRIM in tandem with PCA turned out to be both efficient and effective.

Although traditionally underestimated, peripheral data, even at coarser resolution, plays an important role for fast attentional tasks, as experimental evidence reveals.

Acknowledgments. We want to acknowledge comments from Matthew Turk, clarifications from DoJoon Jung, discussion with Roger Freitas and Lorenz Gerstmayr, comments and Matlab code for PCA from Matthew Dailey, and proofreading from Adolfo Martínez and Raul Montoliu.

References

1. C. Capurro, F. Panerai, and G. Sandini. Dynamic vergence using log-polar images. *Intl. Journal of Computer Vision*, 24(1):79–94, 1997.
2. J. Draper, K. Baek, M. Bartlett, and J. Beveridge. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding (CVIU)*, 91:115–137, 2003.
3. H. M. Gomes and R. Fisher. Learning-based versus model-based log-polar feature extraction operators: a comparative study. In *XVI Brazilian Symp. on Computer Graphics & Image Processing (SIBGRAPI)*, San Carlos, Brazil, Oct. 2003.
4. T. Hastle, R. Tibshirani, and J. Friedman. *The elements of statistical learning; data mining, inference, and prediction*. Springer, 2001.
5. L. Itti. Modelling primate visual attention. In J. Feng, editor, *Computational Neuroscience: A Comprehensive Approach*, pages 635–655. CRC Press, 2003.
6. L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, July-Aug. 2000.
7. F. Jurie. A new log-polar mapping for space variant imaging. Application to face detection and tracking. *Pattern Recognition*, 32:865–875, 1999.
8. H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *Intl. Journal of Computer Vision*, (14):5–24, 1995.
9. Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146:77–123, 2003.
10. M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, Maui, Hawaii, 1991.
11. M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 24(1):34–58, Jan. 2002.