8.07f/W(Jul162004) +model NEUCOM : 10665 Prod.Type:FTP pp.1–16(col.fig.:f1,3–8) ED:SreejaGA PAGN:mamatha SCAN:

ARTICLE IN PRESS

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Neurocomputing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Visual object concept discovery: Observations in congenitally blind children, and a computational approach

Jake V. Bouvrie, Pawan Sinha*

Department of Brain and Cognitive Sciences, MIT, USA

Received 4 February 2005; received in revised form 21 September 2005; accepted 30 January 2006

Abstract

Over the course of the first few months of life, our brains accomplish a remarkable feat. They are able to interpret complex visual images so that instead of being just disconnected collections of colors and textures, they become meaningful sets of distinct objects. Exactly how this is accomplished is poorly understood. We approach this problem from both experimental and computational perspectives. On the experimental side, we have launched a new humanitarian and scientific initiative in India, called 'Project Prakash'. This project involves a systematic study of the development of object-perception skills in children following recovery from congenital blindness. Here, we provide an overview of Project Prakash and also describe a specific study related to the development of face-perception skills following sight recovery. Based in part on the results of these experiments, we then develop a computational framework for addressing the problem of object concept discovery. Our model seeks to find repeated instances of a pattern in multiple training images. The source of complexity lies in the non-normalized nature of the inputs: the pattern is unconstrained in terms of where it can appear in the images, the background is complex and constitutes the overwhelming majority of the image, and the pattern can change significantly from one training instance to another. For the purpose of demonstration, we focus on human faces as the pattern of interest, and describe the sequence of steps through which the model is able to extract a face concept from non-normalized example images. Additionally, we test the model's robustness to degradations in the inputs. This is important to assess the model's congruence with developmental processes in human infancy, or following treatment for extended congenital blindness, when visual acuity is significantly compromised.

Keywords: Object concept learning; Unsupervised learning; Holistic object recognition; Recovery from blindness; Face detection

1. Introduction

Through experience, the brain becomes adept at parsing complex visual scenes into distinct objects. An infant, for example, learns that certain constellations of visual cues correspond to her father. Later, when the relation has solidified, the child is capable of recognizing "father" quickly, and from a variety of viewpoints or illumination conditions. But how does a child, with no prior information, identify an object of interest across a collection of possibly ambiguous or cluttered images? Understanding how the human visual system learns to perceive objects in the environment is one of the fundamental challenges in neuroscience, and it is this question that motivates our work. We do not wish to take a dogmatic stand on whether the brain of an infant is a 'blank slate' or loaded with innate knowledge. What we do wish to investigate, however, is the computational feasibility of acquiring object concepts without presupposing the existence of innate representations.

The research presented in this paper comprises two complementary components: in the first half, we describe an experimental field study from "Project Prakash." A new humanitarian and scientific initiative launched in India, Project Prakash involves a systematic study of the development of object-perception skills in children following recovery from congenital blindness. In the latter half of the paper, we introduce a computational model for visual object concept discovery. Through this computational

*Corresponding author. Tel.: +1 617 253 1434.
E-mail address: psinha@mit.edu (P. Sinha).

0925-2312/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.neucom.2006.01.035

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

investigation, we seek to illuminate from a different angle the process by which humans build object models. Constructing a machine analog of this process, however, is a difficult computational challenge given that natural images are often complex and contain many attributes irrelevant to the concept that needs to be extracted. Several issues concerning this process remain open: how much visual experience is needed for the development of this ability? What are the intermediate stages in the evolution of object representations? How does the quality of early visual input influence the development of object concepts? These questions are of central importance for both the children enrolled in Project Prakash and our computational system alike. In pursuing the latter, it is our goal to determine the nature of learning processes which individuals recovering from blindness might plausibly use to discover visual object concepts.

2. Project Prakash

From an experimental standpoint, there are two dominant approaches for studying object learning: (1) experimentation with infants and (2) experiments with adults using novel objects. These approaches have yielded valuable results, but their usefulness is limited by some significant shortcomings. For instance, infant experiments are operationally difficult and the development of object-perception processes is confounded with the development of other brain subsystems such as those responsible for attention deployment and eye-movement control. Experiments with adults, on the other hand, are necessarily contaminated by the subjects' prior visual experience, even though the objects used as stimuli may be novel.

We have identified a unique population of children in India that allows us to adopt a very different approach. According to the WHO, India is home to the world's largest population of blind children. While the incidence of congenital blindness in developed nations such as the USA and UK is less than 0.3/1000 children, the incidence in India is 0.81/1000. These rates translate to an estimated 25,000 children being born blind each year in India. Many of these children have treatable conditions, such as congenital cataracts or corneal opacities. However, poverty, ignorance and lack of simple diagnostic tools in rural areas deprive these children of the chance of early treatment. Recently, in response to government initiatives for controlling blindness, a few hospitals have launched outreach programs to identify children in need of treatment and perform corrective surgeries at low cost. These initiatives are beginning to create a remarkable population of children across a wide age-range who are just setting out on the enterprise of learning how to see. We have launched Project Prakash with the goal of helping children receive treatment and then following the development of visual skills in these unique children to gain insights into fundamental questions regarding object concept learning and brain plasticity.

Such a population is not available in developed countries such as the United States. Given the extensive network of neonatal clinics and pediatric care in these countries, congenital cataracts are invariably treated surgically within a few weeks after their discovery. Consequently, in the developed world, it is rare to find an untreated case of blindness in a child of more than a few months of age. In India, on the other hand, many children with congenital cataracts spend several years, or even their entire lives, without sight. The societal support and quality of life for blind children in India is extremely poor, leading to a life expectancy that is 15 years shorter than that of a sighted child. There is clearly a humanitarian need to help such children get treatment, and a key goal of Project Prakash is to help address this need. Furthermore, in tackling this need, the Project is presented with a unique scientific opportunity.

The scientific goal of Project Prakash is to study the development of low-level visual function (such as acuity, contrast sensitivity and motion perception), as well as object perception following recovery from congenital blindness. We are investigating the time-course of different object-perception skills as assessed behaviorally, the concurrent changes in cortical organization, and also the development of neural markers associated with object perception. Of special interest to us is face perception, including face localization, identification and expression classification. Few object domains can rival the ecological relevance of faces. Much of the human social infrastructure is critically dependent on face-perception skills. We are studying both the deficiencies and proficiencies of children after onset of sight. The former allows us determine the visual skills that are susceptible to early visual deprivation while longitudinal studies of the latter yield insights about how face perception develops and what the underlying processes might be. These studies complement work on the development of visual abilities after short durations of congenital blindness [26]. This body of work has examined the consequences of a few months of early childhood blindness on visual abilities several years later. Our work looks at the development of visual skills following more extended durations of congenital blindness.

We call this project 'Prakash', after the Sanskrit word for 'light', symbolizing the infusion of light in the lives of children following treatment for congenital blindness and also the illumination of several fundamental questions in neuroscience regarding brain plasticity and learning.

The potential impact of this work extends beyond pure science, into the domain of pediatric eye-care. Significant advances have been made in surgical procedures to treat many cases of childhood blindness, such as those due to congenital cataracts or corneal opacities. However, merely treating the eyes is not sufficient for ensuring restoration of normal visual function. An equally important requirement for sight recovery is that a child's brain be able to correctly process the visual information, after having been deprived of it for several years. Based on past animal studies of the

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

consequences of visual deprivation on subsequent function [2,17,19,24,39], we can expect that the treated children will exhibit visual deficits relative to normally developing children. However, we know very little about what the nature of these deficits will be, and Project Prakash is a step towards acquiring this information. Determining which skills the children are impaired at is crucial for creating effective rehabilitation schemes that would allow the children to be integrated into mainstream society and lead a normal active life. It is important to emphasize that although the patient population for this study is drawn from India, the results are relevant to child health in general. Furthermore, the spotlight this project is bringing to bear upon the problem of treatable childhood blindness is likely to strengthen outreach programs not just in India but globally.

Within the broad context of Project Prakash's motivations and goals, we have conducted several specific studies of object perception. Here, we report an investigation of face-classification skills following recovery from blindness.

3. A specific study from Project Prakash: face classification following long-term visual deprivation

Past work has suggested that early visual deprivation profoundly impairs object and face recognition [9,12,31,36]. Even relatively short periods of deprivation, ranging from the first 2 to 6 months of life, have been shown to have significant detrimental consequences on face-recognition abilities [23]. However, we currently lack experimental data that address the more basic issue of the influence of early visual deprivation on face versus non-face discrimination (hereinafter also referred to as 'face classification'), i.e., can face classification skills be learned later in life? Results from infant studies of face perception are not too helpful in formulating a hypothesis in this context. While it is generally accepted that visual experience during the first 2–3 months of life is sufficient for the babies to exhibit a reliable preference for face-like patterns [11,20,25,28,29], it is not known whether similar learning processes continue to be available later in life. It is possible that long-term visual deprivation might permanently impair an individual's face-classification skills.

In order to investigate face/non-face classification skills following extended visual deprivation, we studied two children, SB and KK, who had both recovered sight after several years of congenital blindness. SB is a 10-year-old boy who was born with dense bilateral cataracts. Prior to treatment, he showed no awareness of people's presence via visual cues and could orient to them only on the basis of auditory cues. The cataracts severely compromised his pre-operative pattern vision. He was unable to discern fingers held up against a bright background beyond a distance of 6 in. By comparison, subjects with normal acuity can perform this task at 60 ft and even an individual with 20/400 acuity, who would be classified as legally blind according to WHO guidelines, would be able to do this task at approximately 36 in. It is an indicator of the poor state of awareness in rural India regarding childhood blindness, that when SB was brought in to a hospital, it was not to treat his eyes, but rather a leg injury he had suffered after tripping on an obstacle. After having been blind for 10 years, SB underwent cataract surgery in both eyes (the two procedures were conducted a month apart). The opacified lenses were replaced with synthetic intra-ocular lenses (IOLs). Post-operative acuity in SB's eyes was determined to be 20/120, significantly below normal, but a great improvement over his original condition. SB's left eye currently exhibits significant strabismus.

KK is an 11-year-old girl, also born with dense bilateral cataracts. Visual deprivation appears to have been severe right from birth since the pupils were seen to be white (due to opacified lenses) even while she was an infant and KK did not exhibit any visually guided responses. Furthermore, the nystagmus that KK currently exhibits also suggests severe visual deprivation during infancy [14,43]. In tracing KK's family history, we found that her father had also been born with congenital cataracts. Thus, KK's blindness at birth was considered 'destined' (a blind father being expected to have a blind daughter) and no effort was made by her family to seek medical attention. It was only when KK was 7 years old that she happened to be examined by an ophthalmologist visiting her village as part of an outreach program. She was treated shortly thereafter and the opaque lens in her right eye was replaced with an IOL. Current visual acuity in this eye is approximately 20/120. Her left eye is still untreated and provides no useful vision.

With their guardians' permission, we conducted simple experiments to study SB and KK's face/non-face classification performance. The experiments were conducted 6 weeks post-(first) surgery for SB and 4 years post-surgery for KK. Fig. 1 shows SB and KK's eyes at the time of the study. SB's strabismus (squint) and KK's dense cataract in the left eye are evident in the images.

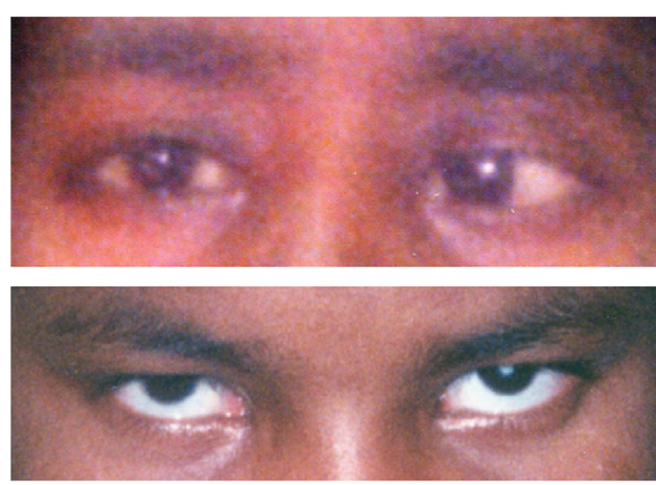

Fig. 1. Views of SB's (top) and KK's eyes at the time our studies were conducted. Both have recovered functional vision in their right eyes. However, SB has significant strabismus in his left eye while KK continues to have a dense cataract in her left eye.

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

The first set of studies involved discriminating between face and non-face patterns and locating faces in complex scenes. We also assessed the performance of two age- and gender-matched controls with normal vision. Our stimulus set for the 'face/non-face discrimination' task comprised monochrome face images of both genders under different lighting conditions and non-face patterns. The non-face distracters included patterns selected from natural images that had similar power spectra as the face patterns and also false alarms from a well-known computational face-detection program developed at the Carnegie Mellon University by Rowley et al. [30]. Sample face and non-face images used in our experiments are shown in Fig. 2a and b, respectively. All of the face images were frontal and showed the face from the middle of the forehead to just below the mouth. Face and non-face patterns were randomly interleaved and, in a 'yes–no' paradigm, the subject was asked to classify them as such. Presentations were self-timed and the images stayed up until the subject had responded verbally. No feedback was provided during the experimental session. The patterns subtended 10° of visual angle, horizontally and vertically.

Fig. 2. The kinds of stimuli we used in our experiments (rows are labeled a–f top to bottom): (a) images of upright faces, (b) non-face distracters, (c) scenes with front-facing people, (d) blurred upright faces, (e) inverted faces, and (f) isolated face parts.

For the 'face-localization' task, we used natural scenes, containing one, two or three people (a few sample stimuli are shown in Fig. 2c). Face sizes ranged from 2° to 4° of visual angle. The subjects' task was to indicate the locations of all faces in a scene by touching the display screen with the index-finger. The response was recorded as a 'hit' if the first touch was within a face boundary. Incorrect locations were recorded as 'false-alarms'. Both the number and correctness of responses to each scene were recorded.

As the top row of Fig. 3 shows, SB and KK exhibited a high hit-rate and a low false-alarm rate on the face/non-face discrimination task, achieving performance similar to that of the age-matched controls. On the face localization task as well, the two groups were comparable. However, SB and KK's face localization performance on gray-scale scenes was completely compromised; they reported seeing no faces at all anywhere in the images (these data are not shown graphically in Fig. 3, since they can be summarized adequately in words: the controls exhibited a 100% detection rate while SB and KK were at 0% in gray-scale images), pointing to the great significance of color information. These data suggest that the ability to discriminate between faces and non-faces and also to localize faces in complex scenes can develop despite prolonged visual deprivation. Furthermore, the fact that SB exhibited this performance within 6 weeks of treatment suggests that the development of face-classification abilities does not require exceedingly long periods of visual experience after sight onset.

These results bring up the important issue of the nature of information used by SB and KK for accomplishing face-classification tasks. Past work [23] suggests that individuals with a history of deprivation are impaired at processing faces configurally and instead analyze them in terms of isolated features such as the eyes, nose and mouth, at least when asked to individuate faces. A similar issue exists for face detection. We attempted to determine whether SB and KK's face classification abilities were based on the use of such a piecemeal strategy wherein the presence of a face was indicated by the presence of specific parts. To this end, we performed an additional set of experiments that specifically investigated the use of holistic versus featural information. These experiments used images that were transformed to differentially affect featural versus configural analysis.

We created three stimulus sets, each containing 20 items. The first comprised low-pass filtered face and non-face patterns. The low resolution of these images obliterated featural details while preserving the overall facial configuration. The second comprised vertically inverted faces.

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

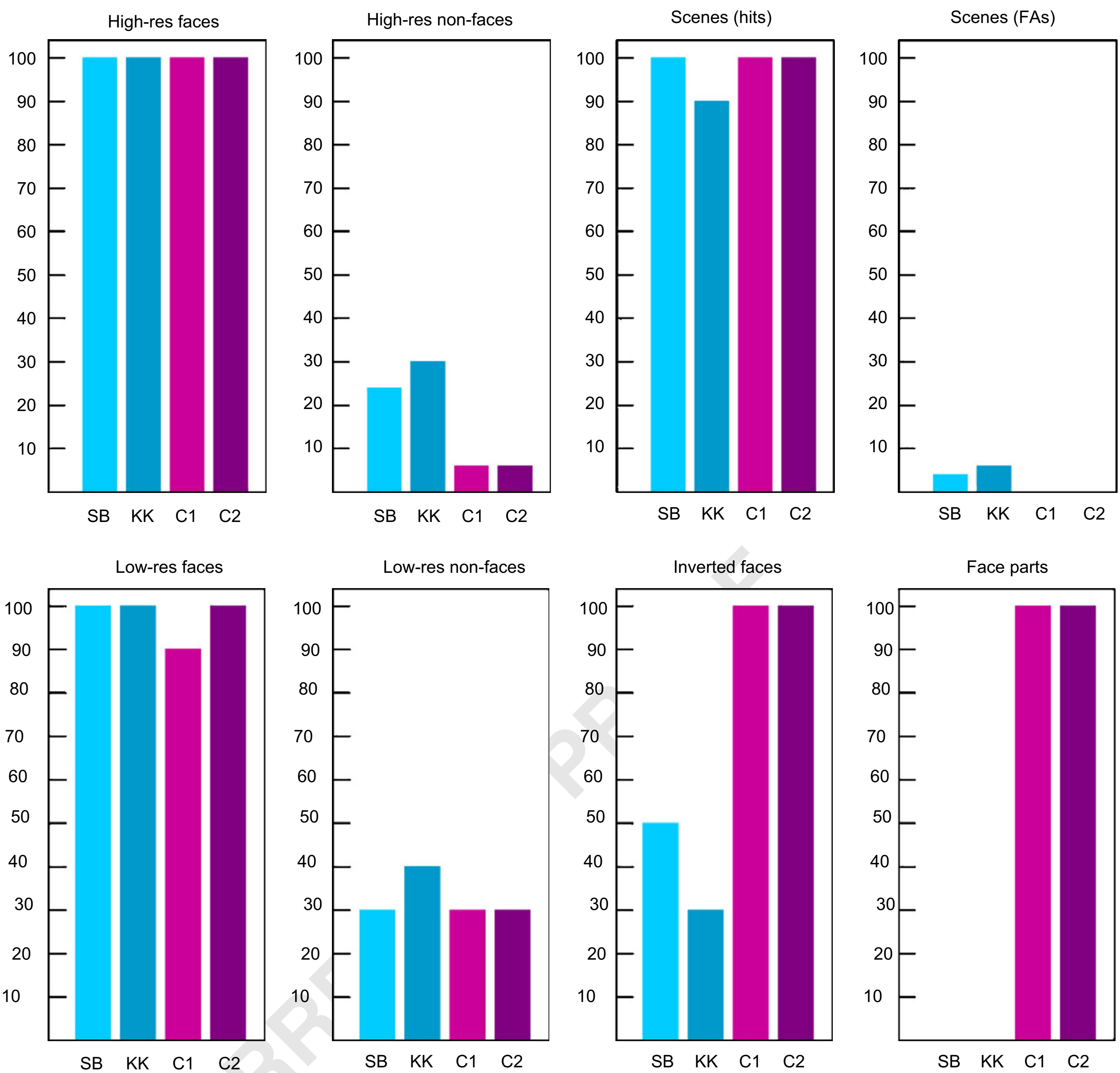

Fig. 3. Results from SB, KK and two age-matched controls on various face-perception tasks. All values are percentages. For 'High-res faces', the bars represent percent correct performance (what proportion of faces are classified as faces). For 'High-res non-faces', the bars represent false-alarm rate (what proportion of non-faces are classified as faces). For 'scenes (hits)', the bars represent the proportion of faces correctly localized. For 'scenes (FAs)', the bars represent the false-alarms as a proportion of the total number of faces in the set. The meanings of the bars for 'Low-res faces' and 'Low-res non-faces' are identical to those for 'High-res faces' and 'High-res non-faces', respectively. For 'Inverted faces', the bars represent percent correct performance (what proportion of the inverted faces are classified as faces). For 'Face-parts', the bars represent percent correct naming performance (what proportion of the face-parts are correctly recognized).

Vertical inversion is believed to compromise configural processing while leaving featural analysis largely unaffected [6]. There were no separate distractor patterns for the inverted face set. Since all of the stimuli were presented in random order, non-faces from the low-resolution and high-resolution sets were also interspersed with the vertically inverted faces. The third set comprised images of individual features (eye, nose and mouth). These feature images were enlarged so that low-level acuity issues would not confound the recognition results. Sample stimuli from each of these sets are shown in Fig. 2d–f. The first two sets were used in a face/non-face discrimination task, while for the third, the subjects' task was to verbally name what the image depicted. The children were not pre-informed that they

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

would be seeing face features. A feature-based strategy would predict that performance would be poor with the first set (low-resolution images devoid of featural details), and comparable to controls for the second and third sets.

The results are summarized in the lower row of Fig. 3. We found that SB and KK performed as well as the age-matched controls on the low-resolution face classification task. However, their performance was significantly poorer with inverted faces and isolated features. Notice that the controls do not exhibit impaired performance with inverted faces. This lack of an 'inversion effect' is not surprising, since the task here is not identification, but simply face/non-face classification. SB and KK's poor performance on the isolated feature set is unlikely to be due to extraneous factors such as an inability to understand the instructions or a lack of labels. The naming task included other non-face objects as well, and SB and KK performed well on naming them. Thus, they demonstrated that they understood what the task entailed. Furthermore, SB and KK did possess the labels 'eyes', 'nose' and 'mouth', and could point to these features on their own faces and could also appropriately label the blobs on full-faces. However, when the features were presented alone, the children could not recognize what they were. Thus, SB and KK showed an ability to label the parts of a face based on the overall configuration of the face pattern, rather than via their individual details. This pattern of results strongly suggests the use of holistic information by SB and KK. Details of individual face parts appear to be neither necessary nor sufficient for classifying a pattern as a face.

Taken together, our experimental results suggest that children can rapidly develop face classification abilities even after prolonged visual deprivation, lasting for several years after birth. Furthermore, the face concept used for classification appears to encode holistic information rather than piecemeal featural details. This particular encoding strategy may well be a consequence of the relatively poor acuity the children possess following treatment for prolonged blindness. Acuity limitations reduce access to fine featural details and may, thereby, induce the use of holistic face information available in low-resolution images. In drawing inferences from these data, two caveats deserve note. First, our data show only accuracy, not reaction times. We cannot, therefore, say that the equality of performance between the experimental and control groups in terms of accuracy is a definitive indicator of the normalcy of the former. It is possible that reaction-time data might reveal systematic differences between the two groups. Second, our results derive from only two experimental subjects. This brings up obvious issues of generalizability. However, in our more recent work, we have found a similar pattern of results with other individuals who have recovered sight following extended congenital blindness. Two of the most interesting cases, to be described in detail in a separate paper, comprise a woman who gained sight at the age of 10 years after surgery for corneal opacity, and another who was treated for bilateral cataracts at the age of 12 years. We worked with these individuals more than 15 years after their surgeries, and found that their face perception skills were well developed, and that the cues they appeared to rely on were similar to those that SB and KK used. This gives us confidence that the results from SB and KK are not idiosyncratic.

These findings are also interesting in that they may guide the development of computational models of human face-detection skills. Most current models implicitly assume that faces are encoded in terms of their parts [16,22,35]. Face concept learning in these models proceeds by first acquiring facial parts which are then, optionally, combined into a larger representation. This emphasis on the use of face-parts as pre-requisites for face classification is not reflected in our experimental results. A model that proceeded by developing a holistic face representation without need for featural details, which may be added later as higher acuity information becomes available, would be more congruent with these experimental data.

One way of reconciling our results with past reports of piece-meal processing is by assuming that visual deprivation does not compromise the encoding of overall facial configuration per se, but rather, the ability to discern differences between variants of the basic configuration. This has the consequence of increasing reliance on featural differences for distinguishing one face from another, a characteristic of piecemeal processing. It is worth pointing out that the distinction between configural and piece-meal processing that has been considered in the literature thus far has focused on the task of face individuation. Our task here is different in that it requires classifying patterns as members of the 'face' category. Thus, there is no obvious conflict between the results we have reported here and those presented in earlier papers such as [23]. A more accurate characterization of our results is that SB and KK tend to use 'first-order' configural information for the face-detection task. We do not know whether they possess sensitivity for 'second-order' configural relationships, but based on earlier studies [23], we would expect that they probably do not.

In considering whether these results have any bearing on the development of face perception skills in normal infants, it needs to be remembered that children like SB and KK differ in many ways from neonates. Unlike the newborn, SB and KK have had extensive experience of the environment through sensory modalities other than vision. This experience has likely led to the creation of internal representations that may well interact with the acquisition of visual face concepts. Furthermore, the deprivation may have led to structural changes in neural organization. For instance, projections from other senses may have claimed sections of the cortex that, in normal brains, would be devoted to visual processing [32,39]. Thus, a priori, we cannot assume that the developmental courses of face perception in a 10-year old recovering from blindness will have much similarity to that in the newborn. However, some interesting parallels deserve further scrutiny. Primary

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

among these is the quality of initial visual input. Both these populations typically commence their visual experience with poor acuity. The compromised images that result may constrain the possible concept learning and encoding strategies in similar ways. Thus, there exists the possibility that normal infants, and children treated for blindness at an advanced age, may develop similar schemes as a consequence of the similarity in their visual experience. However, the validity of this conjecture needs to be tested via further experimentation.

As we shift our focus towards the modeling of object learning, five aspects of the aforementioned experimental results stand out as useful constraints on the design of the computational system:

1. In the natural setting, the faces can appear at any of a large number of positions. Typically, there is not an explicit pointer provided by a 'teacher' as to the exact location where the face was, in any given image. However, SB and KK were able to acquire the face concept despite these complexities. While we do not know precisely which cues they used to acquire the face concept, we can make a general inference that object learning can proceed with weakly labeled inputs.
2. Color appears to be an important orienting cue during the early development of visual recognition skills. SB and KK performed very well on the task of face location with colored images, but were at chance with gray-scale ones.
3. Detailed featural information appears to be less important than overall object configuration. SB and KK were able to classify faces even when the images were degraded, leading to the individual features being reduced to indistinct blobs.
4. The internal representation is limited in terms of the in-plane rotation it can tolerate. Vertical inversion of the face images greatly compromised SB's and KK's face classification performance.
5. Object discovery can be accomplished with limited training instances. SB's excellent face classification performance, just a few weeks after sight acquisition, while not a definitive proof of this conjecture, is suggestive of its validity.

In the following sections, we describe a simple computational model that is guided by these observations. The specific details, such as the choice of algorithms used, cannot lay claim to being biologically accurate (we simply do not have enough information about the biological processes to make that claim), but the overall constraints on their input–output mappings respect the findings we have reported above.

4. Computational modeling of object discovery

In this section, we propose a computational model for the process by which humans might build a set of general object concepts (e.g. "car" or "dog"), from a collection of visual stimuli. Typical computational schemes for concept learning require that the learning system be provided with a training set of images showing the target object(s) isolated and normalized in location and scale [5,15,16,27]. But, while such a "pre-processed" training set simplifies the problem, it also renders the approaches unrealistic and circular from a developmental standpoint, because in order to normalize an image, one needs already to possess the object concept. Recently, several unsupervised and mixed supervised/unsupervised attempts have been made that partially avoid the inherent inconsistencies between supervised machine learning and behavioral learning in humans. Both Weber et al. [40] and Agarwal and Roth [1] built parts-based representations for several categories of real-world objects. The latter automatically identify visual objects in a scene and extract descriptive features in order to build a classifier from a set of labeled images. Weber et al. take an entirely unsupervised approach, and build generative representations for object parts, the parameters of which are learned via the expectation-maximization algorithm. Fei-Fei et al. [8] also take a Bayesian approach to unsupervised learning of visual object categories, whereby "generic" knowledge from previously discovered models is brought to bear on novel categories. By setting prior class probability distributions on the basis of previous experience, the authors are able to model additional categories using only a handful of examples. While the system we present does not attempt to model more than one category simultaneously, previous experience does play an important role in guiding later object detection processes.

Our model for visual object concept learning is motivated largely by the experimental findings presented above, in addition to studies of object perception in infancy [33]. In particular, the model is designed to work with non-normalized training data: given a set of images, each of which contains an object instance embedded in a complex scene, the system attempts to automatically discover the dominant object concept. We employ an unsupervised learning strategy at the outset to formulate hypotheses over the set of possible concepts. At this stage, the processing is, of necessity, bottom-up. The only means of complexity reduction are low-level image saliencies and a priori regularity within an object class. As visual experience accumulates, however, the object concept undergoes concurrent refinement, allowing the model to utilize a top-down strategy in an effort to reduce search complexity in subsequent images. Such a mixture of bottom-up and top-down strategies represents a plausible computational analog to the gradual use of prototypes in object recognition as observed experimentally in humans.

The optimal point at which an artificial system ought to begin applying prototype knowledge is typically problem-specific, and there has been little prior research to establish either empirically proven heuristics, or any sort of precedent based on biological processes occurring in

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

humans. It is therefore an additional goal of this research to consider cases where object concepts are to be learned under conditions simulating visual impairment, and to ultimately discern how and where image resolution directs the application of prototypes to unseen images so as to increase the probability of concept discovery. We begin, in the following section, with a description of the computational system, from feature extraction to object identification. We then present some experimental results given both full-resolution and low-resolution (low-pass filtered) image sets in Section 6, and lastly in Section 7 offer a discussion of the results and their significance within the context of the experimental findings presented in the first half of this paper.

5. System design

We first present a bird's eye-view of the overall system, before describing the individual modules in greater detail. The first step is to create an initial training dataset. Next, each image is processed so as to identify the most salient sub-regions. A large number of such regions are extracted from the training images, and encoded via a low-dimensional representation. These patches are then clustered using a hierarchical algorithm, thereby grouping extracted regions by theme in the expectation that one of them will provide an initial implicit representation of the concept. If necessary, further clustering or merging iterations are performed in conjunction with additional unseen datasets to successively refine the concept representation. It should be noted that the success of the system does not depend on the specific choices of algorithms we have made, and we have chosen where possible the most ''vanilla'' techniques that can accomplish the desired processing.

The overall structure of our approach is similar in spirit to those of Weber et al. [40] and Agarwal and Roth [1]. This is perhaps inevitable since the task of unsupervised learning must of necessity have a clustering operation at its core. However, there are some important differences that deserve note. First, instead of using the Foerstner interest operator [10], we have attempted to design the front end of our model to be consistent with biologically motivated schemes for saliency estimation. Second, our representation of image structure uses attributes that are plausibly computable by neural hardware. Third, we propose a novel scheme for the iterative refinement of object concepts, which allows the model to keep improving its representation with increasing experience. We now provide details for each of the steps in our model. With an eye towards making our work accessible to an interdisciplinary audience, we have minimized the use of equations and technical derivations. We have also attempted to draw all of the specific techniques we have used in this work from the standard collection of tools in computational statistics. Details regarding these techniques are readily available in several textbooks (see for instance [42]).

5.1. Image processing for training dataset generation

We generated a dataset by embedding concept examples (faces) of fixed size within background scenes (interiors) at random locations (Fig. 4). Individual faces and backgrounds were not repeated in the training set to prevent unnatural cluster biases from arising. Face images were derived from the USF database [3] and backgrounds were

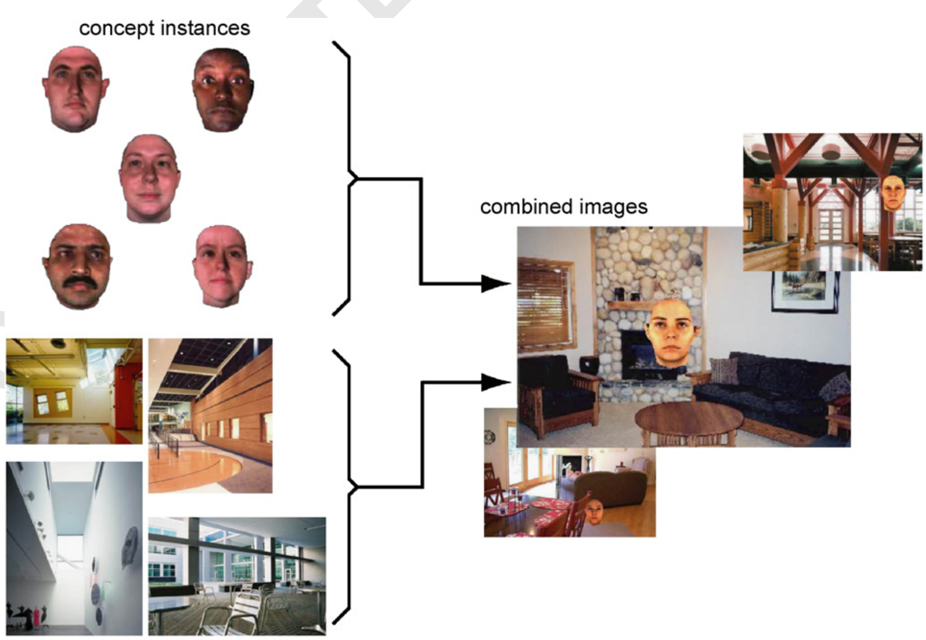

Fig. 4. The set of training images is generated by embedding examples of the concept to be learned (cropped faces) at random locations within background scenes (interiors). Neither backgrounds nor faces are reused in the generation process.

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

obtained from several different commercially available image collections. All images were in color and there were no constraints on the complexity of the backgrounds. While the face concept examples were fixed in size, we did not attempt to normalize the relative scale of the background images, and a substantial variation in the relative size of face features versus background features was maintained. Furthermore, faces occupied less than 5% of the total area in any given training image. Thus, the overwhelming majority of visual attributes in the training set were irrelevant to the desired face concept. Finally, the training set included background-only images (different from the backgrounds upon which the faces had been superimposed). We will return to the role of these images below. The size of the faces was chosen so as to mimic the information available to a limited acuity system (approximating 20/100, similar to SB and KK) when a real face is about 5 ft from the eye. In a typical household where the size of a room is about 10 ft square, the average distance of interaction between two people is approximately 5 ft. Having separate sources for the faces and backgrounds facilitated our computational tests by allowing us to generate arbitrarily many training/test images. The natural (non-composited) images of humans we had access to had a limited range of background complexity. By including complex scenes as backgrounds, and placing no restrictions on where the faces could appear, we were better able to test our model by ensuring that faces would fall within the vicinity of a wide range of different background objects and textures. Without playing down the significance of working with a non-composited image set, we believe that the database generation strategy we have employed serves as a good initial way to test our object discovery approach.

5.2. Image processing for saliency-based image sampling

Following work by Itti et al. [18], and motivated by the receptive field structures of cells in the initial sections of the mammalian visual pathway, salient regions within the training images were identified utilizing center-surround operations (Fig. 5). Color and intensity information was accumulated across the levels of a dyadic Gaussian image pyramid as pixel-wise differences between fine (center) and coarse (surround) scales. The image pyramid was constructed by recursively downsampling (by a factor of 2) and low-pass filtering, starting at the highest-resolution image. The resultant feature maps, encoding intensity and double-opponent (red-green, blue-yellow) color responses, were collapsed into a single normalized "saliency map." The map in effect guides the selection of locations, functioning as a biologically motivated model for early visual attention. An image's saliency map was then thresholded so as to

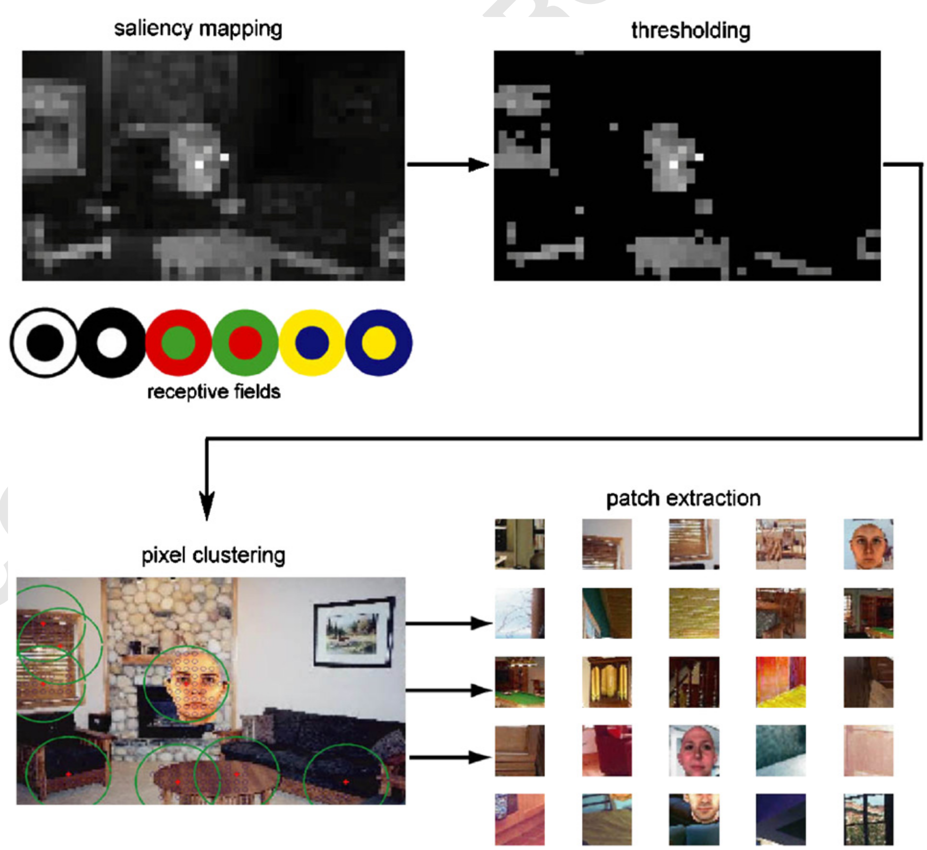

Fig. 5. Each training image is transformed (from left to right, panel 1) and thresholded (panel 2) to produce a "map" of salient locations, the most prominent of which are identified and chosen according to spatial clustering of the map's image pixels (panel 3). The locations of the top few clusters determine the location in the original (un-processed) image from which we extract patches of fixed size (panel 4).

identify only the most salient of locations and facilitate later patch extraction. In this system, the threshold was chosen dynamically at each image such that the top 10–30% most salient pixels were retained.

Following thresholding, salient pixels were clustered in space with a simple hard-membership algorithm, in this case K-means. We have chosen $K = 8$ (the choice is arbitrary, and the results do not change significantly for other values of K), and additionally seed the K-means algorithm with the results of a preliminary divisive clustering pass in order to stabilize and improve results. It is possible that a more elaborate clustering scheme such as IsoData could be used to tailor the number of extracted patches to each individual stimulus, as a way to reduce the number of non-concept patches extracted at each image. However, in the interest of simplicity and to reduce the amount of (possibly data-dependent) parameter tuning necessary, we have chosen to use a generic K-means implementation with a static number of centers. Square patches centered at each of the (K) converged means were then extracted and saved into a master database of potential concept examples. Patch size was set so as to mimic information gathering within a 10° window, assuming an acuity of 20/100 (which approximates the acuity of SB and KK). The heuristic underlying this choice is that acuity falls to 30% at 5° on either side of the fovea, and drops off even more substantially beyond this. This makes it difficult to acquire image information in a single fixation if the patch size is any larger than 10°. We could, in principle, have chosen a much smaller patch size, but given that the model at the outset has no information about the size of the object, it is expected to learn, it is less presumptive to adopt a size limited only by acuity considerations. These are admittedly heuristic choices, and a more principled method of selecting a patch size (or possibly several sizes) would be an interesting avenue for further study. For the present discussion, we shall commit to the heuristic described here. Overall then, the purpose of this stage in the system is to automatically identify and extract the top K most salient regions in the image.

5.3. Feature extraction and dimensionality reduction

Given a dataset of image fragments gathered as described above, feature extraction and a reduction in dimensionality is necessary not only for storage efficiency, but also to obtain some degree of robustness against image variations. While recent machine learning work has suggested that dimensionality enhancements might also be useful for facilitating classification [7,37], we adopt the more conventional approach of employing dimensionality reduction as a way to gain limited tolerance to variations in image appearance. The use of higher-dimensional representations is a topic we set aside for future investigations.

We therefore computed principal components by projecting each patch in the dataset onto the space spanned by the eigenvectors corresponding to the N largest eigenvalues of the feature-wise covariance matrix of the dataset. The principal components thus describe as much of the variance in the dataset as possible. Selecting only the "dominant" components provides a reduced representation that captures only large variations in the dataset. Additional features summarizing frequency information were calculated by taking the 2D Discrete Cosine Transform (DCT) of the 2D Fourier magnitude [34]. The magnitude of the discrete Fourier transform is shift-invariant and reveals periodic textures in an image while the DCT combines related frequencies into a single value and conveniently focuses energy into the top-left corner of the resulting image. After transforming the original patch, a triangle from the top-left corners of each (RGB) color pane was normalized and combined with the principal components information to form the final feature vector. We chose to reduce $40 \times 40 \times 3$ sized color image patches to 188 features: 80 of them principal components and the remainder DCT/DFT triangles. For trials involving low-resolution datasets (1/8th and 1/16th of full resolution), the number of distinct pixels falls below 188 and the feature extraction step is eliminated completely. In this case, the raw patch pixel values by themselves are used for further computations since the dimensionality is already low, and minor high-frequency variations in the image are eliminated a priori during reduction to the lower resolutions.

For resolutions higher than 1/8th of full resolution, the "patches" are now represented by features from the corresponding low-dimensional representation described above, but we will continue with our original terminology and refer to these feature vectors as simply "patches", even though they are not, strictly speaking, raw image pixels.

5.4. Patch labeling

At this stage of the system, we have on our hands two separate collections of patches: those that were extracted from images containing an embedded concept instance, and those that were drawn from background scenes devoid of concept instances. We will call the former collection "concept patches" and the latter "non-concept patches", but it is important to bear in mind that the "concept patches" may or may not actually include the desired concept—the distinction is that the concept patches *might* include the concept while the "non-concept" patches will, by definition, not have the concept. The backgrounds used to generate non-concept patches are different from those used to generate concept-containing images, but represent similar themes (here, more interiors). Regions from these background-only images are selected, extracted, and processed in the same manner as those coming from images which do contain an embedded concept example.

Finally, we assign a binary label to the patches as a means by which to remember the original source. We chose the following convention: if a patch's label reads $+1$, then it is a "concept patch" and, by definition, came from a

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

concept-containing image. Conversely, if the label is −1 then this tells us that the patch did not come from an image that contained the concept, and thus will not contain the desired concept. This weak labeling only provides information about whether or not a patch is *possibly* a concept example, and cannot be used to identify which patches contain the concept. In this sense, the labels can be interpreted as additional sensory cues which might be used to assist in the development of an object concept. For instance, in an infant's world, the presence of auditory speech cues could serve as an implicit label that differentiates between face and non-face images. In such situations, intermodal cues provide a reliable hint that the concept is present in the visual field. They do not, however, identify which concept among others is important, where the concept is in space, nor whether the concept is occluded or in some way transformed from any previous notion of the concept the infant may or may not already possess. We also do not require that the infant (or our system) learn a direct mapping between labels or additional cues and the visual concept. The auditory cues we wish to compare to the system's labels need only reveal that a face is present and we do not assume that the cue is, for instance, a veridical label such as "father".

The precise role of labels in the system will be described below.

5.5. Patch merging and hierarchical clustering

In order to ultimately separate concept-containing patches from the others, we wish to determine which ones ought to be grouped together using some notion of distance between patches or groups of patches. Given our two collections of patches, we combine all "concept patches" with "non-concept" patches into a single dataset, while retaining the labels as previously assigned. The background-only patches and potential concept patches are then clustered jointly using a hierarchical agglomerative algorithm designed to form successively larger groups of similar patches by agglomerating groups or individual patches together according to distance. This technique is completely unsupervised, and label information is not used. The inter-cluster distance metric we have chosen computes the average intra-component variance for each cluster pair, tentatively combined. The pair of clusters yielding the minimum combined variance, taken over all possible pairs, is thus merged into one and the iteration repeats until the desired number of clusters has been reached. Typically, the final cluster count is selected so that merging of large, substantially different clusters does not occur, and can be chosen heuristically based on the number of data points to be clustered. This overall approach gives results similar to that of Ward's method [21,38], which also evaluates cluster variance, and empirically outperforms other popular linkage methods (single, complete, group-average) for this particular clustering problem. For the experiments presented below, we used 130 composited training images and 40 background-only images. From each of these 170 images, we extracted 8 patches giving a dataset of 1360 patches in total. We found that 100 final clusters was an appropriate balance between cluster size and patch homogeneity; hierarchical clustering was thus applied for 1260 iterations until only 100 clusters remained as desired.

For a given collection of patches extracted from concept-containing images, only an eighth or fewer may actually contain a concept example: most patches come from interesting but irrelevant background edges and color shifts. Hierarchical clustering, dutifully performing the task of identifying common patterns, will usually uncover both the concept theme as well as other peripheral themes, leaving one with a number of concept possibilities (Fig. 6). How then, is the system to know whether the concept we are interested in is "chair" or "face" if both are found in large numbers across the set of training images? It is here that labeling comes in as the crucial means by which the

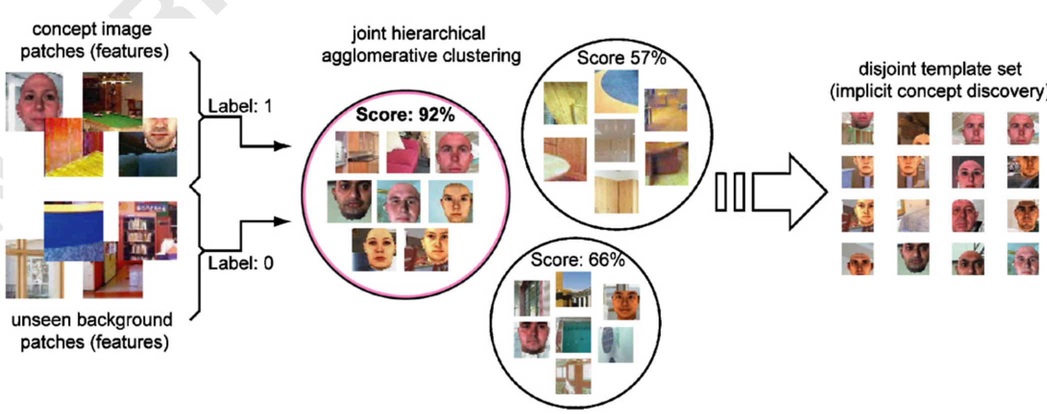

Fig. 6. Patches are labeled according to whether they were extracted from an image with a concept example (label = 1), or from a background-only image (label = 0). The labeled patches are then merged together and jointly clustered. Each resulting cluster is scored by computing the fraction of patches that were originally taken from concept-containing images. The cluster scoring highest gives rise to a template set representation of the concept.

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

concept cluster can be distinguished among other contending themes.

5.6. Scoring and selection

After applying the hierarchical clustering algorithm to the dataset of combined "concept" and "non-concept" patches, we are left with approximately N clusters of varying size (where N was chosen to be 100 given our dataset size and choice of K above). In order to determine which of these clusters best represents the desired concept, each of the N clusters is scored according to the fraction of its members (patches) previously labeled as originating from concept-containing images.

If enough negatively labeled (background only) patches are clustered alongside "concept" patches, non-concept themes arising due to background commonalities are generally sufficiently diverse in origin (extracted from concept-containing versus4 background-only images) so that no theme present in only the backgrounds used to form concept-containing images can be misinterpreted as the desired concept. If a spurious concept (e.g. windows, wood textures, or swathes of sky) is present in the scenes used to form the original concept-containing dataset, then the unseen backgrounds will hopefully also contain those themes and a mixture will form during the joint hierarchical clustering. The concept is then left as the only theme which could possibly have arisen from the set of concept-containing images alone.

The scores assigned to each of the clusters, in effect, denote the probability that a given cluster represents the concept. With such scores, we can then rank-order the clusters and identify one or more of them as candidate representations of the desired object concept.

5.7. Concept discovery

In the final stage of the system, we simply select the highest scoring cluster as determined in the previous step, and discard patches labeled as having come from background-only images, leaving behind a set of patches rich in the concept. If the ratio of concept to non-concept patches in this final set is high, the collection can be interpreted as a preliminary disjunctive "template" for the concept. This template can then be immediately applied to further unseen data. Additional (test) images might be processed using the template set in place of another hierarchical clustering stage, yielding a search significantly reduced in complexity: if a test patch of unknown identity is sufficiently "close" to the template, then we simply declare it to be representative of the concept. We need not collect another dataset of experience or determine a second hierarchy of patch clusters. Several reasonable choices for determining proximity to the template set exist, including distance to the mean of the set, or a distance equal to the minimum of the distances between the test patch and each member of the template set. A variance-scaled distance ought to work well in theory; however, the number of examples comprising a template set is typically not great enough to compute accurate covariance estimates for a Mahalanobis-based distance criterion—even with harsh constraints imposed. Empirically, both the minimum distance and mean-template metrics have been found to work comparably well.

5.8. Template evolution

Each concept representation can be combined with further unseen data to iteratively generate improved concept representations via a feedback mechanism. Given a distance metric and an initial template patch collection, a small set of unseen images is processed in order to determine how well the preliminary concept prototype is able to guide searches for the concept. Even if the initial template set does not appear to be mostly homogeneous in the concept, a subset of patches can still be selected from test images using the template guidance method described above. Although the resulting collection will include many spurious patches due to erroneous members in the template set, it will typically have a much improved ratio of concept to non-concept patches compared to the test set. The original template set, augmented with a collection of patches selected from the set of test images, can be clustered a second time yielding a second, more homogeneous template collection. If the original template set contained N members, then we select from the results of the second clustering pass the N closest members to the highest scoring cluster mean to arrive at our second version of the template set. The feedback process then repeats, refining the template set until the concept is suitably represented. The presence of spurious members in the starting template set is an inevitable consequence of the weakly labeled training set. The system at the outset does not know what attributes of a given image lead to its inclusion in the 'positive' set. Any attribute that is strongly enough correlated with the label is, as far as the model is concerned, a valid member of the template set. It is only with an accumulation of instances, as with the iterative feedback mechanism described above, that the spurious members get weeded out.

6. Model performance

6.1. Results with high-resolution datasets

The dataset of full resolution images is composed of 130 color scene/concept composites, using backgrounds of average size $175 \times 200 \times 3$ pixels and concept examples of constant size $25 \times 25 \times 3$ pixels. Performance at full resolution, using patches of fixed size $40 \times 40 \times 3$ pixels, shows that object concepts are discovered quickly and accurately after a single hierarchical clustering iteration (Fig. 7). The implicit template set can be used to effectively guide the search for concept patches in an unseen test set,

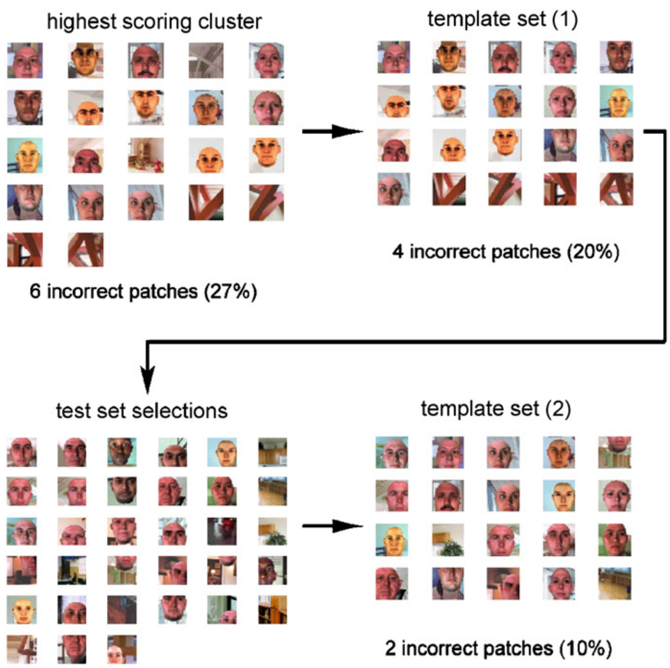

Fig. 7. Given a full resolution dataset, the model provides accurate concept discovery. (From left to right) The highest scoring cluster (from left to right, panel 1) resulting from the agglomerative clustering process gives rise to the first template set representation of the concept (panel 2). The template is used to guide the search for concept patches in an unseen set of test images (panel 3). Finally, patches extracted from the test set of images are used to produce a second, refined, template set (panel 4).

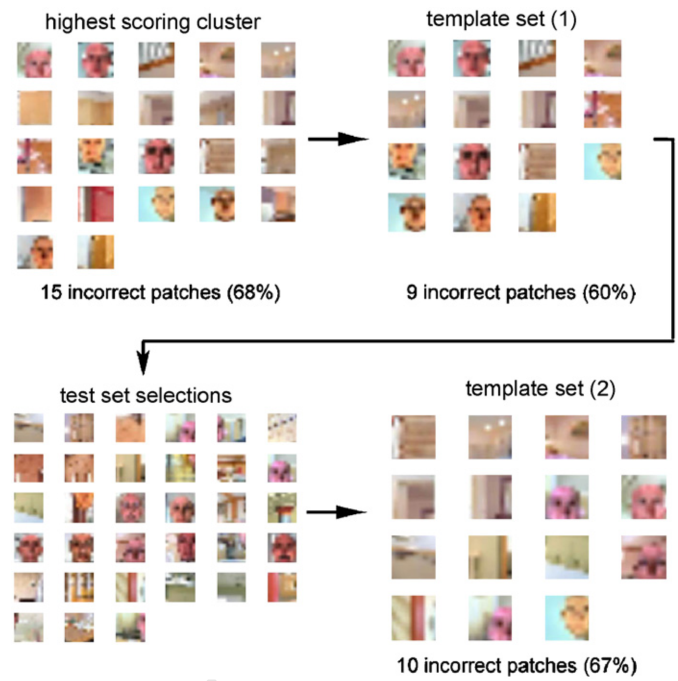

Fig. 8. At one-eighth of full image resolution, the model can still provide a representation of the concept that is useful for future guidance. As in the full resolution case, the highest scoring cluster is used to construct a template set representation of the concept, which is in turn used to guide the search for concept objects in subsequent (unseen) images. Panel 3 shows that guidance is still occurring, with several face patches selected among the hundreds of possible patches. Further refinement at this resolution, however, is unhelpful, as panel 4 reveals.

as shown in the figure. Refinement of the template set using the first collection of processed test patches shows only slight improvement, indicating that template sets generated after clustering are already fairly close to being homogeneous in the concept. Further iterative refinement is largely unnecessary here.

It is possible that if the original dataset had ultra high resolution, the generalization performance of our system might not have been as good as what we have obtained with the current set. In other words, there is likely to be an 'optimal' resolution, intermediate between the very high (which is not conducive to generalization) and the very low (which simply does not provide enough information for reliable discrimination). However, the resolution of the dataset we have used here does not allow us to comprehensively explore this issue.

6.2. *Concept learning with impaired vision*

Impaired vision, in the form of a loss of information due to severe blurring, is simulated as a reduction in image resolution followed by upsampling to the original size. Our choice of resolution reduction as a method for incorporating visual impairment in the inputs is driven by our experimental observations. We find that the primary impairment children in Project Prakash exhibit after treatment is sub-par acuity. We do not yet have any evidence regarding spatial specificities of these impairments (different parts of the visual field being differentially affected). We have, therefore, opted for a uniform degradation in our simulations. Any other choice would have entailed making additional assumptions without the grounding of experimental data. To appropriately mimic the visual consequences of resolution reduction, we use interpolation for image upsampling, rather than a nearest-neighbor-based approach. This reduces pixelation and other artificial effects which might limit effective simulation of vision abnormalities, and additionally prevents artificial saliencies arising from the introduction of false edges.

At 1/8th resolution, with patches of size $5 \times 5 \times 3$ pixels, the discovery of an object concept is significantly more difficult (Fig. 8). The highest scoring cluster becomes infused with many patches bearing only slight similarity to faces, but which nevertheless become close in a reduced feature space. Template set performance reveals, however, that some degree of guidance is still occurring: not all ability to learn the concept has been lost at this resolution. Comparing panels 2 and 4 in Fig. 8, it can also be seen that further applications of the algorithm do not, in this case, improve the concept representation. While the resolution at which iterative refinement proves unhelpful naturally depends on the dataset and the target concept; for this problem, 1/8th resolution marks the point at which optimal performance is likely unachievable via additional itera-

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

tions. At 1/16th resolution, it was found that there is essentially no apparent guidance, as the number of patch pixels becomes reduced to $3 \times 3 \times 3$ in total. At this size, there does not appear to be enough information by which patches might be separated in the pixel space. Further iterations also do not improve the template, indicating that there is perhaps a minimum amount of knowledge one must be able to extract during the initial clustering pass in order to observe further template improvement.

It is worth noting that an infant's (or a Project Prakash participant's) acuity of 20/100 translates to about 2 cycles per degree of visual angle. A face at a distance of 2 ft subtends about 15° of visual angle. Thus, a child with even fairly compromised acuity gets as input a face image that effectively has 60 pixels across its width. As our results attest, our model can accomplish face concept learning very well at this resolution. Even the acuity of a newborn, which has been assessed to be approximately 20/400 is, forms the perspective of the model, adequate to perform well on the face-detection task. This is especially true with the addition of motion cues which we have not considered here. It is only when resolution is brought down dramatically (to levels well below what constitutes legal blindness), that concept learning by our model is compromised.

Overall, the discovery of object concepts is remarkably tolerant to all but the most extreme degradations referred to above. At intermediate levels of resolution, the recovered object concept comprises primarily face patterns, just as is the case at the highest resolution. This suggests that this model is able to mimic the robust ability of Project Prakash participants to learn face concepts despite their impaired acuity.

7. Discussion

We have provided an overview of some of our experimental studies and computational modeling efforts that have the goal of investigating object concept learning. The experimental studies have some obvious limitations, the most significant being the small number of test subjects. Taken just by themselves, the experimental results included here are suggestive rather than definitive. However, we believe that there is reason to expect these findings to generalize. The high consistency between results from SB and KK, and also similar observations in ongoing studies with other subjects in Project Prakash suggest that the pattern of results reported here might hold more generally.

It is worth emphasizing that the current experimental data are not extensive enough to suggest very specific computational strategies for a modeling effort. However, they do provide some broad constraints for a model's design, and simultaneously rule out certain classes of models, such as those dependent critically on the use of fine-grained luminance edge-information. They indicate the potential importance of cues such as color, and the sufficiency of low-resolution image information. The model we have proposed cannot yet be said to faithfully reflect the object learning strategies instantiated in the brain. But, in providing one possible approach, the model allows us to assess the plausibility of the general computational framework it embodies. The particular algorithmic decisions we have made in the model presented here (such as the use of K-means clustering or principal components analysis) are not driven by specific experimental results, but the overall structure of the model is intended to be consistent with empirical data.

Before considering the model's features, let us examine a few of its limitations. The training inputs it can handle at present cannot be entirely unconstrained. While we have placed no restrictions on the location of the target object or the complexity of the background, we have sidestepped the challenges of size, pose and orientation variability by fixing these parameters. We have also not so far devised a completely principled way of selecting patch sizes for the initial analysis of inputs. Furthermore, our representation strategy could be augmented with additional biologically plausible image features/measurements to yield improved clustering performance. However, despite these limitations, we believe that this computational system suggests one plausible model for object concept learning in infants and individuals recovering from blindness. The model has begun serving a useful purpose for demonstrating feasibility of explanations regarding experimental data by mimicking some aspects of human performance. We explore some of these parallels next.

The model's reliable performance at a resolution approximating that of infants and Project Prakash participants, implies that, at least in principle, concept discovery is possible with significantly impaired acuity and some sensory assistance to provide implicit labels, which constitute a weak form of supervision. Future development of the model presents the opportunity to provide concept realization without labels of any sort, thus providing a computational analog to object concept learning in humans given static scenes and zero additional sensory input. Beginnings have already been made in this direction. For instance, recent work by Fei-Fei et al. [30] has explored concept discovery when supervisory input is entirely abolished.

It should be noted that of the cues available to both children during early development [4,13] and the computational system, color information plays a significant role. In infants, color contributes strongly to orientation and segmentation of objects for eventual identification and recognition, and this is especially true in the absence of motion or other peripheral cues (as in static images). Correspondingly, double-opponency color filters are an integral part of the saliency computation we apply to the dataset of images in the computational model.

It is also interesting to consider the congruence between the computational model and the experimental results from human subjects participating in Project Prakash. As described in Section 3, children suffering from varying

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035

degrees of congenital blindness typically exhibit significantly impaired visual acuity. Such acuity deficits are neural in origin and cannot be corrected with refractive aids. A child, therefore, has to acquire object concepts with blurred images of the kind utilized in the preceding section. In agreement with the computational experiments, we found that even with these constraints on input quality, the children participating in Project Prakash can develop face classification abilities using concepts that appear to encode holistic information rather than piecemeal featural details. It is possible that this bias is in fact adaptive since it induces the visual system to use information that is suited for the task of face-recognition. The diffusive degradations obliterate part details to a greater extent than their overall configuration. Analogously, the disjunctive face-concept that the system discovers comprises whole faces, rather than isolated parts. This is possibly because the detailed structure of the parts is more variable than their overall configuration. The holistic configuration, therefore, can support the clustering steps of the model better than the part details. It should be noted, however, that while the system demonstrates the effectiveness of a holistic encoding, it does not show for this particular problem that holistic features arise naturally over parts-based representations. Therefore, with respect to the nature of the internal object concept representation and the development of concepts as a function of visual acuity, the computational model provides one possible explanation for the observations collected from Prakash patients. It also shows that acuity improvements beyond the thresholds of legal blindness can enhance the feasibility of face-concept learning. This has an interesting applied implication. Despite the fact that for the majority of Project Prakash cases, refractive lenses cannot completely correct acuity impairments, the computational simulations predict that even a small increase in visual acuity can yield significant improvements in face classification ability. Thus, in those cases where even partial refractive correction is possible, we might expect the benefits of such aids to be substantial rather than marginal.

A fairly limited amount of training, a few weeks in duration, appears sufficient to engender face concepts in children. While it is difficult to quantitatively estimate the number of face instances a child would have been exposed to during this period, the result qualitatively suggests that human face learning can proceed with a potentially limited training set. The computational model too requires a small training dataset (100–200 images), placing it within the range of biological plausibility. It would be interesting to compare these results with those from normally developing infants using similar sets of stimuli.

Finally, there are some interesting parallels to consider between visual development in children recovering from congenital blindness, and normally developing infants. Primary among these is the quality of initial visual input. Both these populations typically commence their visual experience with poor acuity. The compromised images that result may constrain the possible concept learning and encoding strategies in similar ways. Thus, there exists the possibility that normal infants, and children treated for blindness at an advanced age, may develop similar schemes as a consequence of the similarity in their visual experience. Our model allows us to develop hypotheses about the nature of internal representations as a function of the nature of visual experience. It can, therefore, serve as a valuable aid not just for modeling what is known about high-level visual development, but also for designing novel studies to be conducted with infants or the children participating in Project Prakash.

Uncited reference

[41].

Acknowledgment

The authors wish to thank Prof. Gedeon Deak for his very helpful comments on earlier drafts of the manuscript.

References

[1] S. Agarwal, D. Roth, Learning a sparse representation for object detection, in: Proceedings of the Seventh European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, vol. 2353, Springer, Berlin, 2002, pp. 113–130.
[2] J.A. Bauer, R. Held, Comparison of visually-guided reaching in normal and deprived infant monkeys, J. Exp. Psychol. Anim. Behav. Process. 1 (1975) 298–308.
[3] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, Proc. SIGGRAPH (1999) 187–194.
[4] M.H. Bornstein, Qualities of color vision in infancy, J. Exp. Child Psychol. 19 (1975) 401–419.
[5] M.C. Burl, T.K. Leung, P. Perona, Recognition of planar object classes, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1996.
[6] R. Diamond, S. Carey, Why faces are and are not special: an effect of expertise, J. Exp. Psychol. Gen. 115 (1986) 107–117.
[7] T. Evgeniou, M. Pontil, C. Papageorgiou, T. Poggio, Image representations for object detection using kernel classifiers, in: Proceedings of Asian Conference on Computer Vision, 2000.
[8] L. Fei-Fei, R. Fergus, P. Perona, A Bayesian approach to unsupervised one-shot learning of object categories, Proc. ICCV (2003).
[9] I. Fine, A.R. Wade, A.A. Brewer, M.G. May, D.F. Goodman, G.M. Boynton, B.A. Wandell, D.I.A. MacLeod, Long term deprivation affects visual perception and cortex, Nat. Neurosci. 6 (2003) 915–916.
[10] W. Foerstner, E. Guelch, A fast operator for detection and precise location of distinct points, corners and centres of circular features, ISPRS Intercommission Workshop, Interlaken, June 1987.
[11] C.C. Goren, M. Sarty, P.Y. Wu, Visual following and pattern discrimination of face-like stimuli by newborn infants, Pediatrics 56 (1975) 544–549.
[12] R.L. Gregory, J.G. Wallace, Recovery from early blindness: a case study, Q. J. Psychol. Monogr. (2) (1963).
[13] M.M. Haith, J.J. Campos, Human infancy, Annu. Rev. Psychol. 28 (1977) 251–293.
[14] C. Harris, Nystagmus and eye-movement disorders, in: D. Taylor (Ed.), Paediatric Ophthalmology, Blackwell Science, Oxford, 1997, pp. 869–896.

[15] B. Heisele, Visual object recognition with supervised learning, IEEE Intell. Syst. Artif. Intell. Second Century May/June (2003) 38–42.
[16] B. Heisele, P. Ho, J. Wu, T. Poggio, Face recognition: component-based versus global approaches, Comput. Vision Image Understanding 91 (2003) 6–21.
[17] D.H. Hubel, T.N. Wiesel, S. LeVay, Plasticity of ocular dominance columns in monkey striate cortex, Philos. Trans. Soc. Lond. B 278 (1977) 377–409.
[18] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1998) 1254–1259.
[19] S.G. Jacobson, I. Mohindra, R. Held, Development of visual acuity in infants with congenital cataracts, Br. J. Ophthalmol. 65 (1981) 727–735.
[20] M.H. Johnson, J. Morton, Biology and Cognitive Development: The Case of Face Recognition, Oxford, UK, 1991.
[21] S. Kamvar, D. Klein, C. Manning, Interpreting and extending classical agglomerative clustering algorithms using a model-based approach, in: Proceedings of 19th International Conference on Machine Learning (ICML), July 2002.
[22] D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[23] R. Le Grand, C.J. Mondloch, D. Maurer, H.P. Brent, Early visual experience and face processing, Nature 410 (2001) 890 (see also, Correction, Nature 412 (2001) 786).
[24] S. LeVay, T.N. Wiesel, D.H. Hubel, The development of ocular dominance columns in normal and visually deprived monkeys, J. Comp. Neurol. 191 (1980) 1–53.
[25] D. Maurer, M. Barrera, Infants' perception of natural and distorted arrangements of a schematic face, Child Dev. 52 (1981) 196–202.
[26] D. Maurer, T.L. Lewis, C.J. Mondloch, Missing sights: consequences for visual cognitive development, Trends Cogn. Sci. 9 (2005) 144–151.
[27] A. Mohan, C. Papageorgiou, T. Poggio, Example-based object detection in images by components, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 349–361.
[28] C.A. Nelson, P. Ludemann, Past, current, and future trends in infant face perception research, Can. J. Psychol. 43 (1989) 183–198.
[29] O. Pascalis, S. de Schonen, J. Morton, C. Deruelle, M. Fabre-Grenet, Mother's face recognition by neonates: a replication and an extension, Infant Behav. Dev. 18 (1995) 79–85.
[30] H.A. Rowley, S. Baluja, T. Kanade, Neural-network based face-detection, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 23–38.
[31] O. Sacks, To see and not see, in: An Anthropologist on Mars Vintage Books, New York, 1995, pp. 108–152.
[32] N. Sadato, A. Pascual-Leone, J. Grafman, V. Ibanez, M.P. Deiber, G. Dold, M. Hallett, Activation of the primary visual cortex by Braille reading in blind subjects, Nature 380 (1996) 526–528.
[33] E. Spelke, G. Gutheil, G. Van de Walle, The development of object perception, in: S. Kosslyn, D. Osherson (Eds.), An Invitation to Cognitive Science, second ed., vol. 2, MIT Press, Cambridge, MA, 1995, pp. 297–330.
[34] M. Szummer, R.W. Picard, Indoor–outdoor image classification, MIT Media Laboratory Perceptual Computing Section Technical Report No. 445, 1998.
[35] S. Ullman, M. Vidal-Naquet, E. Sali, Visual features of intermediate complexity and their use in classification, Nat. Neurosci. 5 (2002) 682–687.
[36] A. Valvo, Sight Restoration after Long-Term Blindness: The Problems and Behavior Patterns of Visual Rehabilitation, American Foundation for the Blind, New York, 1971.
[37] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[38] J. Ward, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244.
[39] M.-C. Wanet-Defalque, C. Veraart, A.G. De Volder, R. Metz, C. Michel, G. Dooms, A. Goffinet, High metabolic activity in the visual cortex of early blind human subjects, Brain Res. 446 (1988) 369–373.
[40] M. Weber, M. Welling, P. Perona, Unsupervised learning of models for recognition, in: Proceeding of the Sixth European Conference on Computer Vision (ECCV), Springer, Berlin, 2002.
[41] T.N. Wiesel, D.H. Hubel, Single-cell responses in striate cortex of kittens deprived of vision in one eye, J. Neurophysiol. 26 (1963) 1003–1017.
[42] I.H. Witten, E. Frank, Data Mining, second ed., Morgan Kaufmann, New York, 2005.
[43] T. Yagasaki, M. Sato, S. Awaya, M. Nakamura, Changes in nystagmus after simultaneous surgery for bilateral congenital cataracts, Jpn. J. Ophthalmol. 37 (3) (1993) 330–338.

Jake Bouvrie is a graduate student in the Department of Brain and Cognitive Sciences at MIT. He received an MEng in electrical engineering and computer science in 2004, and bachelor's degrees in electrical engineering and mathematics in 2003, all from MIT.

Pawan Sinha is an associate professor of computational and visual neuroscience in the Department of Brain and Cognitive Sciences at MIT. He got his undergraduate degree in computer science from the Indian Institute of Technology, New Delhi and his masters and doctoral degrees from the Department of Computer Science at MIT. Using a combination of experimental and computational modeling techniques, research in Pawan's laboratory focuses on understanding how the human brain learns to recognize objects through visual experience and how objects are encoded in memory. Pawan is a recipient of the Alfred P. Sloan Foundation Fellowship in Neuroscience, the John Merck Scholars Award for research on developmental disorders and the Jeptha and Emily Wade Award for creative research.

Please cite this article as: J.V. Bouvrie, P. Sinha, Visual object concept discovery: Observations in congenitally blind children, and a computational approach, Neurocomputing (2007), doi:10.1016/j.neucom.2006.01.035