

Reconstructing mental object representations: A machine vision approach to human visual recognition

EROL OSMAN¹, ADRIAN R. PEARCE², MARTIN JÜTTNER¹
and INGO RENTSCHLER¹

¹*Institute of Medical Psychology, University of Munich, Germany*

²*School of Computing, Curtin University, Perth, Australia*

Received 23 August 1999; revised 15 March 2000; accepted 29 March 2000

Abstract—This paper introduces a new approach to assess visual representations underlying the recognition of objects. Human performance is modeled by CLARET, a machine learning and matching system, based on inductive logic programming and graph matching principles. The model is applied to data of a learning experiment addressing the role of prior experience in the ontogenesis of mental object representations. Prior experience was varied in terms of sensory modality, i.e. visual versus haptic versus visuohaptic. The analysis revealed distinct differences between the representational formats used by subjects with haptic versus those with no prior object experience. These differences suggest that prior haptic exploration stimulates the evolution of object representations which are characterized by an increased differentiation between attribute values and a pronounced structural encoding.

Keywords: Object recognition; representation; recognition-by-parts; graph matching.

1. INTRODUCTION

The question concerning the quality of internal representations underlying human object recognition is still unresolved. Prominent computational theories have postulated, on the one hand, that objects are mentally represented by three-dimensional (3D), object centered, part-based descriptions (Marr and Nishihara, 1978; Biederman, 1987). On the other hand, more recent studies have argued for representations of 3D objects in terms of multiple, viewer centered, two-dimensional (2D) views, among which the visual system interpolates if necessary (Tarr and Pinker, 1989; Poggio and Edelman, 1990; Ullman and Basri, 1991; Bühlhoff and Edelman, 1992).

In a comparative study using signal-detection analysis Liu *et al.* (1994) have found that the internal representations humans use for object recognition may be characterized by falling between the extremes of a 2D and 3D format. However,

their analysis did aim to delimit human performance rather than making details of these representations explicit. Here we present a novel method to reconstruct properties of visual representations underlying object recognition, by modeling human performance by a recognition system adopted from computer vision. In contrast to previous approaches (see e.g. Poggio and Edelman, 1990; Bühlhoff and Edelman, 1992) the objective here is not just to fit recognition performance in terms of the global error rate but to reproduce the specific pattern of confusions, i.e. classification matrices, between objects in the recognition task at hand.

The machine vision system we are using is CLARET (consolidated learning algorithm using relational evidence theory — Pearce and Caelli, 1999). CLARET is based on inductive logic programming, attribute generalization and graph matching principles employing a recognition-by-parts paradigm. It provides a structural object representation in terms of components and relational rules. Within the field of machine vision CLARET has been successfully applied to classical object-recognition problems such as the interpretation of handwritten characters (Pearce and Caelli, 1999). In the current context, CLARET will serve us as a computational model to analyze the structure of mental object representations concerning the relative importance of different attributes, the degree of differentiation between attribute values and the relational depth of rules.

Using this technique we have investigated in a supervised learning paradigm, how various forms of prior knowledge and the presence of depth information influence learning speed, recognition performance and the ability of spatial generalization. The objects to be learned were composed of spheres, which allowed for an easy and unambiguous segmentation into spatially bounded components. Based on the behavioural data we compare the structure of the underlying mental object representations, as reconstructed by CLARET, and discuss their differences in relation to the various learning conditions.

2. OBJECT LEARNING EXPERIMENT

The learning set consisted of three objects, constructed and displayed on a SGI O2 computer using the Open Inventor software package. Each object was composed of four spheres, with three of them forming an isosceles triangle and the fourth being placed perpendicularly above the center of one of the base spheres (Fig. 1). Object 2 and object 3 were mirror symmetric to each other. 2D views were generated as perspective projections of the objects onto the screen plane of the computer display. For the training views, the viewing sphere was sampled in 60 deg steps; views redundant due to object symmetry were eliminated. This resulted in 22 views in total (6 views for object 1, 8 views for object 2 and object 3). In a similar way, 2D test views were generated by sampling the viewing sphere in 30 deg steps, leading to 83 views (21 for object 1, 31 for both object 2 and 3). Because of the sampling interval, 19 of the test views were old views (views used during training, which were not considered during evaluation of the results), whereas the rest of them were

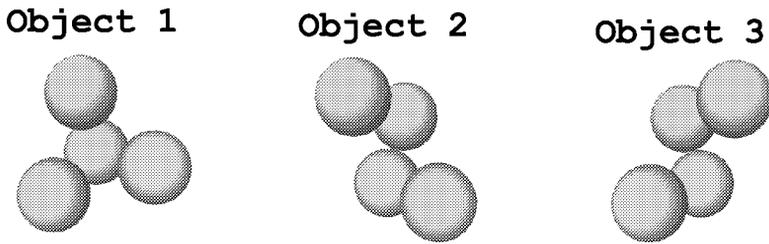


Figure 1. Learning objects used in the experiment. Each object consisted of four spheres, with three of them forming an isosceles triangle and the fourth being placed perpendicularly above the center of one of the base spheres. Note that object 2 and object 3 are mirror symmetric to each other.

Table 1.

Specification of the five learning groups. The groups differed concerning the possibility for prior exploration of the objects and concerning viewing condition

| Group | Prior exploration | Viewing condition |
|------------|-----------------------|-------------------|
| (1) mono | none | monoscopic |
| (2) stereo | none | stereoscopic |
| (3) visual | visual only, by mouse | monoscopic |
| (4) vishap | visual/haptic | monoscopic |
| (5) haptic | haptic only | monoscopic |

new for the subjects. At the viewing distance of 1 m the images appeared under a visual angle of approximately 1.5 deg.

The experiment was divided into three stages, an optional exploration of the objects, a supervised-learning stage and generalization test. The supervised-learning procedure consisted of a number of subsequent learning units (see Rentschler *et al.*, 1994). In each unit the learning set was presented in random order, with an exposure duration of 250 ms for each view, followed by the presentation of the object label for 1 s. The learning unit ended with a recognition test to assess the learning status of the subject. Here, each view had to be assigned to an object by the observer. The sequence of learning units continued until the subject had reached the learning criterion of 90% correct responses in the recognition test. After having completed learning the observers entered the generalization test, where they had to assign all views of the test set which were presented once and in random order.

25 subjects with normal or corrected-to-normal vision, who were divided into 5 groups, participated. The five groups differed in terms of the prior exploration phase and in terms of the depth information provided during the learning stage (Table 1), whereas the generalization test was identical for all groups. The degree of depth information was controlled by the presence of disparity information (via shutter glasses), i.e. by comparing stereoscopic versus non-stereoscopic (mono) viewing conditions. Prior knowledge was varied in terms of sensory modality: Visual versus purely haptic versus visuohaptic. Haptic and visuohaptic exploration was done

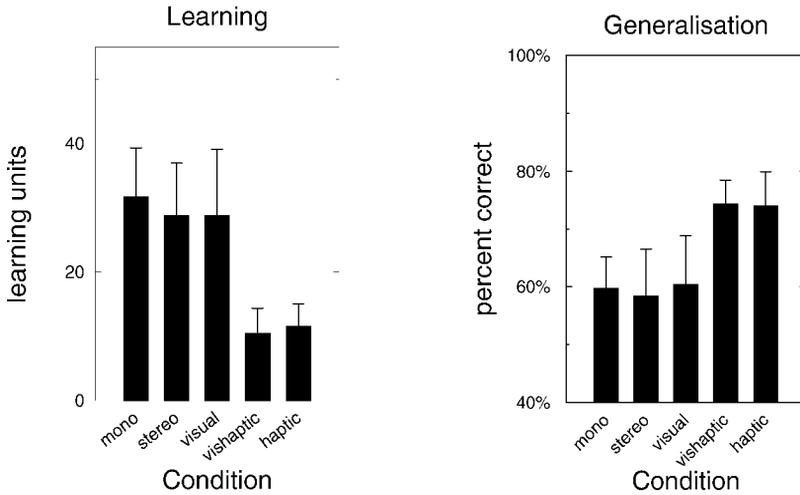


Figure 2. Left: Learning time required to reach the 90%-correct criterion during training. Note the substantially reduced number of learning units in case of the two haptic learning groups. Right: Generalization performance of the five learning groups, each consisting of five observers.

using physical object models constructed of styrofoam balls. In the purely haptic case the subjects were blindfolded. In the visual case the observers could actively rotate the objects on the screen via the computer mouse. The exploratory phase lasted about 5 min, and was immediately followed by the visual learning phase.

As shown in Fig. 2 (Left), learning time, as measured by the number of learning units necessary to reach the criterion, was significantly affected by sensory modality during the exploration phase. Despite its short duration of approximately 5 min, the haptic exploration led to a drastic reduction of learning duration by about 60% relative to conditions with either none or visual-only exploration. Similarly, the haptic exploration also distinctly improved the ability for generalization to novel views. The mean level increased from about 60% correct responses for the non-haptic groups to a mean of about 75% for the haptic groups (Fig. 2, Right). A closer inspection of the data yielded that this improvement was mainly due to the better recognition rate for objects 2 and 3, i.e. the two mirror-symmetric objects. This advantage could be observed only for the haptic groups, whereas a purely visual exploration had no significant effect. Furthermore, stereo viewing yielded no benefit compared to monocular viewing in terms of learning duration and generalization.

3. CLARET SIMULATIONS

A CLARET object recognition system (see Pearce and Caelli, 1999) in general involves the following processing stages: (1) decomposition into parts, (2) generation of relations, (3) induction of rules and (4) matching of parts.

Decomposition into parts: Due to the intentionally simple structure of the objects, we assumed for the experiment an unambiguous segmentation into component

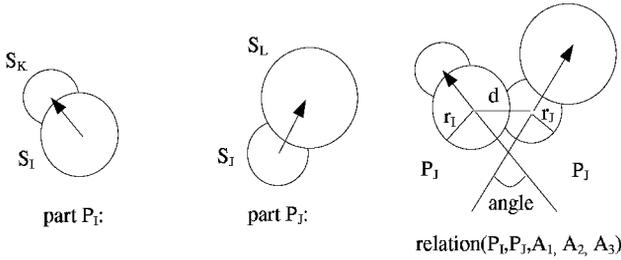


Figure 3. Parts, P_I , are comprised of projected two-dimensional (2D) views of object spheres, S_I , and their relationship (edges) to adjacent spheres, S_K (Left). Pair-wise relations between parts P_I and P_J , defined by, $relation(p_I, p_J, A_1, A_2, \dots, A_n)$, where attributes are comprised of distance between parts (normalised by size of sphere s_I), $A_1 = d/r_I$; projected size ratio (size of S_I to size of S_J), $A_2 = r_I/R_J$ and angle between parts (edge $S_I S_K$ to edge $S_J S_L$), $A_3 = angle$ (Right.)

spheres. The sphere-based objects are subsequently decomposed into parts suitable for the recognition-by-parts approach. Each part, p_I , comprises a two-dimensional (2D) view of a sphere and its relationship to an adjacent sphere (Fig. 3, Left.)

Generation of relations: For each object view, sets of relations are generated over pairs of parts, defined by, $relation(p_I, p_J, A_1, A_2, \dots, A_n)$ which represents the relationship between parts p_I and p_J in terms of numerical attributes A_1, A_2, \dots, A_n . In our implementation attributes comprised of: distance between parts (normalised by size of sphere s_I); projected size ratio (size of S_I to size of S_J); angle between parts (edge $S_I S_K$ to edge $S_J S_L$) (Fig. 3, Right.)

Induction of rules: Input to the matching procedure involves all the relations from all existing (learned) objects, together with relations from the current (to be recognised) object. A taxonomy of rules must now be generated which discriminates between objects based on both attributes and parts. In addition, the rules must be specific enough to discriminate between existing objects but general enough to accommodate new input objects. The technique is based on inductive-logic programming principles and involves a general-to-specific search over first-order relational rules.

Starting from an initial rule, defined by an attribute range which is satisfied by the relations of all existing objects, rule refinement proceeds in two ways: (i) By repartitioning the attribute space by clustering, thus producing more specific rules, and (ii) by incorporating relational information in terms of rules which include other rules in their body. The initial rule is of the form,

$$r_0(P_1, P_2) \leftarrow relation(P_1, P_2, A), 0 \leq A \leq 1.0,$$

where r_0 is the rule number, P_1 and P_2 are part variables and A is a continuous attribute variable which must fall into the range $0.0 \leq A \leq 1.0$ (only the distance attribute is shown for clarity, normalised between 0 and 1.0). Such a rule, can be satisfied by any of the relations at this stage, and will not discriminate between any of the objects.

In step one (i), new rules are generated by partitioning attribute values forming two new ranges. Rule r_0 is replaced with two new rules r_1 and r_2 ,

$$\begin{aligned} r_1(P_1, P_2) &\leftarrow \text{relation}(P_1, P_2, A), 0 \leq A \leq 0.37, \\ r_2(P_1, P_2) &\leftarrow \text{relation}(P_1, P_2, A), 0.49 \leq A \leq 1.0. \end{aligned}$$

For partitioning, a technique is used that minimises the attribute range of each new rule by maximising the partition between rules, it approximates K-means clustering (Jain and Dubes, 1988).

In step two (ii), relational information between attribute values of different parts is captured, achieved by generating new rules which include existing rules as terms in their body,

$$r_3(P_2, P_3) \leftarrow \text{relation}(P_2, P_3, A), 0.49 \leq A \leq 0.86, r_1(P_1, P_2).$$

Notice that rule r_3 includes rule r_1 on the right hand side. This rule is satisfied by relations with parts P_2, P_3 and with parts P_1, P_2 that satisfy rule r_2 . In rule r_3 , the range of distances $0 \leq A \leq 0.86$ between parts P_2 and P_3 is conditional upon the distance between P_1 and P_2 being within the range $0 \leq A \leq 0.37$. This is equivalent to traversing paths through spheres of an object as follows: ‘*sphere p_1 is between 0 and 0.37 in distance away from sphere p_2 and p_2 is between 0.49 and 0.86 away from sphere p_3* ’.

Since the process is recursive, attribute repartitioning and relational term extension of rules results in a hierarchy of rules characterised by two properties: Its range of attribute specificity and depth of relational term extension. A partition measure is calculated for all rules (see Pearce and Caelli, 1999, for details), such that repartitioning occurs in attributes which are either more or less relational, adaptively. The question can now be asked: *At what stage are these rules specific enough to differentiate between the different relational structures present, to correctly interpret the current object?*

Matching of parts: In order to match a new object to the set of existing ones, the hierarchy of rules is used to set up correspondences, or matchings, between the parts of the current (new) object and those of the existing objects. During rule induction, the correspondence of changes from many-to-many toward one-to-one matching. As a result, the correspondence of spheres from the current objects is solved with respect to spheres in one of the existing objects. This is achieved, at each stage during the rule specialization process, by inspecting which relations satisfy rules in terms of which parts are bound to rule variables. In step one all relations satisfy rule r_0 and bindings between variables P_i and parts from current object q_1, q_2, \dots, q_n and existing object p_1, p_2, \dots, p_n are represented as,

$$r_0(P_1, P_2), \{P_1 \setminus (q_1, q_2, q_3, p_1, p_2, p_3, p_4), P_2 \setminus (q_1, q_2, q_3, p_1, p_2, p_3, p_4)\}.$$

As all relations satisfy this rule, notice that possibilities for correspondence of parts from the current to existing is many-to-many. In step two, rule r_0 is partitioned into

two new rules, r_1 and r_2 . As these rules are more specific, less parts bind to their variables,

$$r_1(P_1, P_2), \{P_1 \setminus (q_1, p_1, p_2, p_3), P_2 \setminus (q_2, q_3, p_2, p_3, p_4)\},$$

$$r_2(P_1, P_2), \{P_1 \setminus (q_2, q_3, p_1, p_2, p_3, p_4), P_2 \setminus (q_1, q_2, q_3, p_1, p_2, p_3, p_4)\}.$$

At this stage, a partial matching exists such that part q_1 in the current object either matches to part p_1, p_2 or p_3 for the existing object. As re-partitioning occurs and rules become more specific the correspondence of parts, and hence spheres in the matched objects, solve changing from many-to-many possibilities toward one-to-one matchings. As compatibility of new rules is carried out pair-wise, each new rule needs to be checked with respect to all existing rules. Search for the best match amounts to searching for the most compatible sets of rules which solves the matching to one of the existing objects.

To determine the best matching a probability measure is required. We use an underlying Bayesian network model, by hierarchically conditioning over both relations and rules in the tree (Pearl, 1988). The formulation allows for the analytical determination of probabilities for each rule in the rule tree hierarchically according to its parent,

$$p(r_k(P_J, P_K) \mid existing) = p(r_j(P_J, P_K) \mid relation(P_I, P_J, A))$$

$$\times p(relation(P_I, P_J, A) \mid r_i(P_I, P_J)).$$

Relative probability is based on distributions of attributes and their conditional dependencies through hierarchical conditioning from rule $r_i(P_I, P_J)$ to rule $r_j(P_J, P_K)$ to rule $r_k(P_K, P_L)$ etc. The attribute probabilities correspond to the conditional probability of relations satisfying rules, corresponding to the term $p(relation(P_I, P_J, A) \mid r_i(P_I, P_J))$. These are determined from the likelihood ratio of relations from one existing object satisfying rules with respect to all the objects — calculated by dividing the number of relations satisfying rule r_j from the existing object by the number of relations from all other objects. The relation probabilities correspond to the conditional probability of relations from the current object satisfying rules with respect to the existing object, corresponding to the term $p(r_j(P_J, P_K) \mid relation(P_I, P_J, A))$ — calculated by dividing the number of relations satisfying rule r_j from the current object by the number of relations from the existing object.

Such probabilistic reasoning allows a relative probability to be determined for the current object in terms of each existing object. The relative probability (confusion matrix) of each current-existing matching is calculated over the set of rules using,

$$p(existing \mid current) = \frac{\sum_k p(r_k(P_J, P_K) \mid current)}{\sum_k p(r_k(P_J, P_K) \mid existing)}.$$

Results: Since the behavioural results revealed a distinct dichotomy between subjects with haptic versus those with no prior object experience, the data of the subjects were divided into two groups, haptic and no prior exploration, respectively.

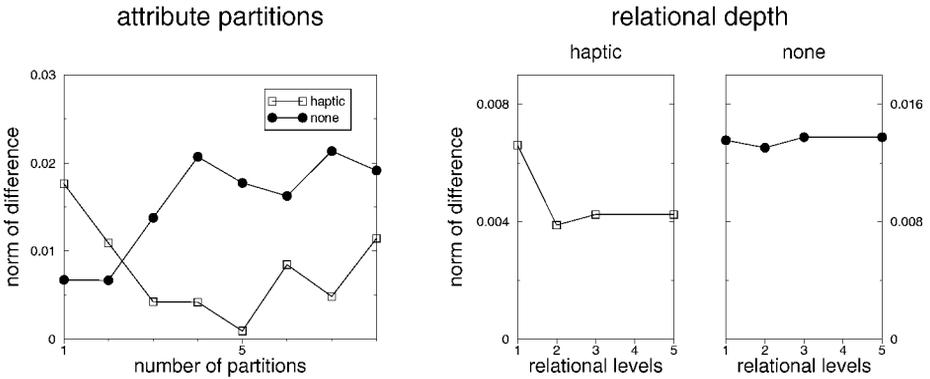


Figure 4. Error between behavioural data (haptic versus non-haptic group) and CLARET-predicted data for classifying the novel views in the generalization test is shown. Error values are in terms of the norm of the difference of the confusion matrices and as a function of the number of attribute partitions (Left) and of relational depth (Right) allowed during the rule induction of CLARET.

For each group, the mean confusion matrices for the learning stage and the generalization test were computed.

In a first attempt, only the correct answers were considered and used as target output in the following simulations. The relative weights that control the degree of partitioning of different attributes during induction of rules were varied as to minimize the error between the confusion matrices of the human subjects and of the CLARET system for identifying the learning views.

In the next step, the generalization data with respect to the test views was used to track down the optimal degrees of attribute partitioning and relational depth in the simulations of the two groups. Attribute partitioning was controlled by limiting the number of partitions within each attribute. Relational depth was controlled by limiting the degree to which rule induction could incorporate other rules as terms in the body of new rules.

As shown in Fig. 4 (Left), the two groups show a divergent behaviour with respect to attribute partitioning. The haptic group reveals a high degree of differentiation between confusion matrices along the attribute dimensions, reaching its optimum around a value of 5 partitions. In contrast, according to the model, observers with no prior experience tend to differentiate between attribute values much more coarsely, with an optimum partition number around 1–2.

Concerning relational depth (Fig. 4, Right), the fit to the data of the haptic group improves as other rules are included in induction of rules, whereas it remains unaffected with respect to the data of the group with no prior exploration. This warrants the conclusion that subjects with prior haptic experience include relational information to a higher degree in their internal object representations than those without such experience.

4. CONCLUSION

We have shown how CLARET, a relational-learning approach adopted from the field of computer vision, can be used as an analytic tool in cognitive modeling. We applied this technique to an object-learning paradigm which was designed to explore the influence of polysensory prior knowledge on the ontogenesis of mental object representations.

The computer simulations yielded distinct differences concerning the structure of reconstructed representations, depending on whether haptic prior knowledge had been acquired or not. Observers with no prior experience revealed only a low degree of differentiation between attribute values. This suggests that these subjects might have used differences in attribute values only on a coarse scale, i.e. in a categorical rather than in a continuous, or metric sense. The difficulty which this group of observers had to distinguish between the mirror-symmetric objects 2 and 3 would agree with such an explanation. In contrast, the haptic-experienced subjects developed a higher degree of attribute differentiation. This ability, together with the enhanced use of relational information, allowed them to successfully distinguish between all three objects. Thus both experimental and simulation results indicate that even a short period of haptic exploration may stimulate the ontogenesis of object representations which have a different format than representations originating from purely visual learning conditions.

Although the present findings have to be regarded as preliminary and require further consolidation in an extended CLARET implementation they indicate a novel perspective for analyzing human competence in object recognition. It is based on recognition-by-parts techniques (see Bischof, 1999) adopted from the field of machine vision. This implies a scope of modeling which exceeds that of approaches aiming for a qualitative description of global performance measures, such as percentage correct or response latency (e.g. Poggio and Edelman, 1990; Bühlhoff and Edelman, 1992), or from those establishing lower and upper bounds of performance by ideal-observer standards (see Liu *et al.*, 1994). Rather it seeks to explicitly reconstruct properties of internal representations. This is achieved by utilizing the observer-specific profile of confusion errors during learning as some sort of cognitive 'finger print' for parameter estimation, and by using the generalization data for cross validation of the system. Previous work has demonstrated the feasibility of such an account in the domain of 2D pattern classification (Jüttner *et al.*, 1997; Rentschler and Jüttner, 1999). The present results suggest that it may also provide a useful tool to analyze processes subserving 3D object recognition.

REFERENCES

- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding, *Psychol. Rev.* **94**, 115–147.

- Bischof, W. (2000). Learning to recognize objects, *Spatial Vision* **13**, 297–304.
- Bülthoff, H. H. and Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition, *Proc. Natl. Acad. Sci. USA* **89**, 60–64.
- Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs.
- Jüttner, M., Caelli, T. and Rentschler, I. (1997). Evidence-based pattern classification: A structural approach to human perceptual learning and generalization, *J. Math. Psychol.* **41**, 244–259.
- Liu, Z., Knill, D. C. and Kersten, D. (1994). Object classification for human and ideal observers, *Vision Research* **35**, 549–568.
- Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes, *Proc. Roy. Soc. Lond.* **200B**, 269–294.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
- Pearce, A. R. and Caelli, T. (1999). Interactively matching hand-drawings using induction, *Computer Vision and Image Understanding* **73**(3), 391–403.
- Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects, *Nature* **343**, 233–282.
- Rentschler, I., Jüttner, M. and Caelli, T. (1994). Probabilistic analysis of human supervised learning and classification, *Vision Research* **34**, 669–687.
- Rentschler, I. and Jüttner, M. (2000). Dynamics and context dependence of visual category learning, *Spatial Vision* **13**, 231–240.
- Tarr, M. J. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition, *Cognitive Psychology* **21**, 233–282.
- Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 992–1005.