6 Object Recognition in Humans and Machines

Christian Wallraven and Heinrich H. Bülthoff

1 Introduction

The question of how humans learn, represent and recognize objects has been one of the core questions in cognitive research. With the advent of the field of computer vision – most notably through the seminal work of David Marr – it seemed that the solution lay in a three-dimensional (3D) reconstruction of the environment (Marr 1982, see also one of the first computer vision systems built by Roberts et al. 1965). The success of this approach, however, was limited both in terms of explaining experimental results emerging from cognitive research as well as in enabling computer systems to recognize objects with a performance similar to humans.

More specifically, psychophysical experiments in the early 1990s showed that human recognition could be better explained in terms of a view-based account, in which object representations consist of snapshot-like views (Bülthoff and Edelman 1992) instead of a full, 3D reconstruction of the object. The most important result of these experiments is that recognition performance is critically dependent on the amount of view-change between learned and tested object view. This stands in stark contrast to the predictions from frameworks using 3D representations such as the often-cited Recognition-By-Components theory (Biederman 1987) which is based on a 3D alphabet of basic geometrical shapes (so-called geons) and predicts a largely view-invariant recognition performance. To date, psychophysical and neurophysiological experiments have provided further evidence for the plausibility of the view-based approach (see, e.g., Tarr and Bülthoff 1998; Wallis and Bülthoff 2001 for two recent reviews).

Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, Tübingen, Germany

In a recent paper, an attempt has been made to reconcile these two approaches to object processing (Foster and Gilson 2002): a careful study of view-dependency of novel objects that were created by combining structural properties (number of parts) with metric properties (thickness, size of parts) has found that both view-dependent and view-independent processing seem to be combined in object recognition. Thus, instead of taking the extreme perspective of either view-based or view-invariant processing one might envisage a visual processing framework in which features are selected according to the current task, where the optimality, efficiency and thus the dependency on viewing parameters of the features depend on the amount of visual experience with this particular task.

Robust extraction of structural, view-invariant features from images, however, has proved to be difficult for computer vision. Therefore, parallel to view-based approaches to object recognition in human psychophysics, view-based computer vision systems began to be developed. These sometimes surprisingly simple recognition systems were based on two-dimensional representations such as histograms of pixel values (Swain and Ballard 1991), local feature detectors (Schmid and Mohr 1997) or on pixel representations of images (Kirby and Sirovich 1990). The good performance of these recognition systems using complex images taken under natural viewing conditions can be seen as another indicator for the feasibility of a view-based approach to recognition.

To date, most theories of object recognition as well as most computer vision systems have mainly focused on the *static* domain of object recognition. Visual input on the retina, however, consists of dynamic changes due to object- and self-motion, non-rigid deformations of objects, articulated object motion as well as scene changes such as variations in lighting, occluding and re- and disappearing objects, and at any given point in time several of these changes can be interacting. The central question for this chapter will thus be: To what extent do object recognition processes rely on *dynamic* information per se? Several psychophysical experiments, which will be discussed below, suggest an important role for dynamic information, in both learning and recognition of objects. Based on these findings, an extension of the current object recognition framework is needed in order to arrive at truly spatio-temporal object representations.

In this chapter, we therefore want to explore the idea of learning and representing objects in a spatio-temporal context by developing a computational object recognition framework motivated by psychophysical results. Specifically, we are interested in developing a recognition framework, which can learn and recognize objects from natural visual input in a continuous perception-action cycle. In the following, we will first briefly summarize the psychophysical experiments that guided the development of the recognition framework. Subsequently, we will present details of the framework together with results from several computational recognition experiments. Finally, we will summarize experiments conducted with a humanoid robot in which the framework was applied to multi-modal recognition of objects using proprioceptive and visual input. These experiments represent a first step towards a closely coupled perception-action system based on and motivated by psychophysical research.

2 Psychophysical Experiments

2.1 Temporal Continuity for Object Learning

To illustrate the importance of temporal context, consider a view-based object recognition system faced with the task of learning object representations. Input to this system consists of a series of views that the system acquires. The problem for this recognition system is how to link the many views of an object to create a consistent and coherent object entity; especially since these views can be very different from each other. One solution to this problem is the observation that in real life we seldom see only isolated snapshots of objects. Usually novel objects are explored either actively through manipulation by our hands or by walking around them. This results in a sequence of images that gradually change from the initial view of the object to a very different one within a short period of time – temporal contiguity. This general observation about natural visual input in a continuous perception-action context motivates the following question: Does the human visual system use temporal contiguity to build a mental representation of the object in order to associate views together? This temporal association hypothesis was investigated in two studies (Wallis and Bülthoff 2001; Wallis 2002), which we briefly review below.

Study 1 – Stimuli: Twelve faces from 3D-laser-scanned female heads were used as stimuli. The faces were separated into three training sets of four faces each. Using a technique by Blanz and Vetter (1999), 3D morph sequences between all possible combinations of face pairs within each set were created. A sequence consisted of a left profile view (–90°) of an original face A, a –45° view of morph A→B (the average of face A and B), a frontal view (0°) of face B, a +45° view of morph A→B, and finally a right profile (+90°) of face A (Fig. 1). A backward

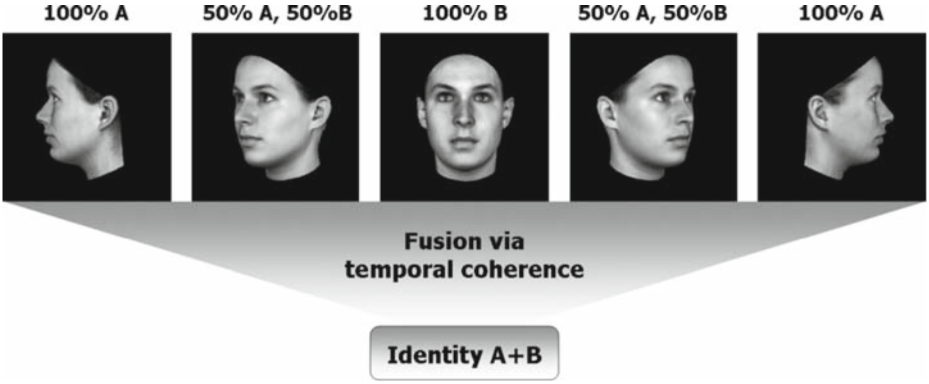

Fig. 1. Illustration of the morph experiment. A morph sequence of two individuals (A and B) is shown to participants who fuse the sequence into one coherent identity due to the temporal continuity present in the visual input

sequence showed the same images in reversed order. The training sequence consisted of a forward sequence and a backward sequence, followed by the forward sequence again and the backward sequence again.

Study 1 – Experimental design: Participants were divided into two groups. In the first group, each participant was trained using sequential presentation of the sequences. In the second group, training used simultaneous presentation of all morph images shown together on the computer screen for the same total time. After training, the participants performed a simple image matching task in which they had to decide whether two subsequently shown images were different views of the same face or not. Half of the trials presented matches, whereas in the other trials half of the test face pairs belonged to the same training set (within set, WS) and the other half to different training sets (between set, BS). If views of objects are associated based on temporal contiguity, then training with sequential presentation should cause the images grouped in one training sequence to be fused together as views of a single object. After such training, participants in the testing phase would be expected to confuse faces that were linked together in a training sequence (WS) more often than between-faces that were not (BS). Training with simultaneous presentation was included to rule out the possibility that the morphs alone were sufficient for the training effect, in which case an effect should appear after both training procedures.

Study 1 – Results: The results of the experiment confirmed that participants were more likely to confuse those faces that had been associated temporally in a sequence (WS). Thus, participants learned to fuse arbitrary views of different faces into one coherent identity without any explicit training. In addition, the results from the second group indicated that the presence of morphs among the training images alone was not sufficient to cause the association of two different faces with each other.

Study 2 – Stimuli and design: In a follow-up study (see Wallis 2002), the results were replicated using two different sets of stimuli for sequential presentation. The sequences here consisted of images of *different* faces instead of morphs thus further increasing the visual difference between frames. In the second experiment, training sequences were created by scrambling the poses in the sequence such that at most two consecutive images showed a consistent and smooth rotation (of 45°). The remaining experimental parameters in the two experiments closely followed the design of the first study for the morph group. This experiment tested whether temporal association based on temporal contiguity could still be detected even when the spatial similarity between consecutive images was low.

Study 2 – Results: The main result from the first experiment was that confusion scores in the WS condition were significantly higher than those in the BS condition indicating that temporal association, indeed, is possible even with more dissimilar sequences. However, the relative effects of temporal association on the two test conditions were *reduced* compared to that of the morphed sequences in the first study. This is an important finding as it indicates that association is influenced by *spatial similarity as well as temporal contiguity*. In the second experiment, there were no significant main effects for a scrambled presentation of images, which destroyed the consistent rotation interpretation but otherwise should have left the pure temporal contiguity intact. However, over the course of three blocks, a significant trend towards a slow dissociation between the two test conditions could be detected. The author interpreted this as a sign that temporal association can take place under such conditions – albeit at a much slower rate.

2.2 General Discussion

Summarizing the two studies, one can conclude that the learning of object representations is strongly influenced by the temporal properties of the visual input. One successful strategy of how the brain might solve the task of building consistent object representations – even under considerable changes in viewing condition – seems to be to assign consecutive images to the same object. This process is not only influenced by temporal parameters but also to a significant degree by the similarity properties of the input. Arbitrary images seem to be much harder to learn, suggesting a crucial influence of the spatial similarity of visual input. These findings therefore are consistent with the extended concept of *spatio-temporal continuity* resulting in integration of images that are below a certain similarity threshold and that are presented within a certain time window.

The findings of these experiments as well as further psychophysical (most notably Stone 1999) and physiological studies (e.g., Miyashita 1988) provide strong evidence for an integral role of temporal characteristics of visual input in object representations and for their active use in learning and recognizing objects. However, the question remains how exactly spatial and temporal information can be integrated in object representations. In the next chapter, we propose a computational implementation, which provides such an integration as part of the recognition and learning procedure.

3 Computational Recognition System

3.1 The Keyframe Framework

The abstract framework shown in Figure 2 consists of several key elements. First, and most importantly, the system processes incoming images in a sequential manner in order to extract so-called keyframes, which represent an extension of the view-concept followed in the view-based approach. Each input frame of an image sequence is first processed in order to extract local features (so-called interest points), which are then tracked across subsequent frames. Eventually, the changes in visual input will be too large and will lead to a loss of tracked features. The core idea behind the framework is that keyframes are precisely defined by that point at which tracking breaks down. If this happens, a new keyframe is inserted into the object representation and the process repeats.

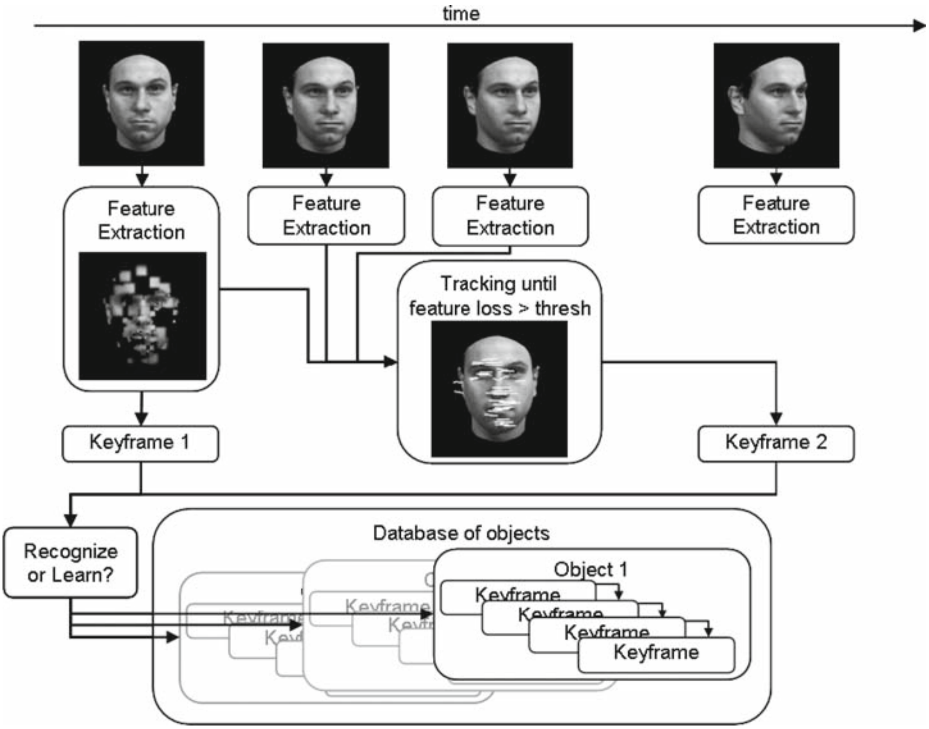

FIG. 2. Abstract description of the keyframe framework. Local feature tracking is used to extract visual events ("keyframes") from an image sequence, resulting in a view-based, connected object representation

Keyframes are thus two-dimensional views (snapshots) of the scene, which are defined by the temporal continuity (in close connection to the psychophysical experiments described in the previous section) of the visual input and form a connected graph of views (see Fig. 2).

Note that in this abstract form the keyframe approach resembles the concept of "aspect graphs" (Koenderink and van Doorn 1979), in which objects are defined by their aspects, i.e., by visual events, where a sudden change in the observed shape of the object occurs. Even though the rigorous mathematical formulations were highly appealing to the computer vision community due to their geometric interpretations, computational realizations of the aspect graph framework for arbitrary objects proved to be difficult. One of the core ideas, however, namely the representation of objects by visual events remains a powerful concept, which our proposed framework retains. Whereas the focus of aspect graphs mainly lies in representations of 3D objects by well-defined views, we want to go one step further with the keyframe concept by representing all kinds of dynamic visual input with the help of two-dimensional views.

Furthermore, learning and recognition are not separated in our framework with new keyframes constantly being compared against the learned library. This means that the system continuously learns new data that can be used to augment existing object representations or form new ones. This is a crucial pre-requisite for any cognitive system, as it is embedded in a dynamic sensory environment and thus constantly receives new input that has to be evaluated and categorized in order to create appropriate (re-)actions.

This embedding of object learning and recognition in a temporal context is reminiscent of the "active vision" paradigm that was developed in the 1980s in computer vision (for example, Aloimonos et al. 1987). Most of the research in active vision was focused on how to control the optics and mechanical structure of vision sensors to simplify the processing for computer vision. Here, we go one step further by endowing object representations themselves with a temporal component through tracking of features and the graph-like keyframe representation.

3.2 Properties of the Framework

As indicated in the introduction, learning and recognition of objects certainly seems possible using only the static dimension – one of the key questions then of course becomes: What – apart from psychophysical motivations – is the advantage of using the temporal dimension in the framework?

Keyframes: In the most extreme case of a view-based framework, learning would involve storing all input images. This strategy is certainly not feasible for any reasonable amount of learning data due to storage constraints. In addition, it also represents a severe problem for recognition as the time it takes to index into the representation becomes prohibitively large. The question thus is: which views to select for learning? Here the keyframe concept provides an intuitive answer to that question: select the views in which an important visual event occurs. In order for this strategy to be successful, one needs to make the assumption that the visual input is, on average, slowly changing, which, given the psychophysical evidence presented above, certainly seems to be valid. Furthermore, the keyframes are organized in a directed graph structure, which allows for pre-activation of frames during recognition of image sequences. If two connected keyframes in a row could be recognized, chances are good that the next incoming keyframe will be the next node in the graph. This strategy thus dramatically reduces the search time during recognition of known sequences or sequence-parts.

Visual features: We chose to include local features in the framework in order to focus on locally informative visual aspects of each frame (see Fig. 2). These local features consist of simple image fragments (much in the spirit of Ullman et al. 2002) extracted around interest points that are detected in the image at several scales. Whereas of course the exact nature of these features is open to further experimentation (for example, Krieger et al. 2000; Lowe 2004 for other approaches), already these relatively simple image fragments are effective in compressing image data. In addition, the most important contribution of the tracking that is used to determine the keyframes is that it allows access to feature trajectories. In our framework, the trajectories follow features from one keyframe to the next. The larger the visual difference between keyframes, the more discriminative these feature trajectories are – this is because the chances of false matches are reduced, the longer a feature can reliably be tracked (see also Tomasi and Kanade 1991). More importantly, the trajectories describe the transformation of each feature from one keyframe to another and thus can be used to generate priors for matching feature sets. Consider, for example, a sequence of a rotating object for which the feature trajectories between keyframes will have a shape that is specified by the direction of the (3D) rotation. For recognition, a matching prior can now be derived directly from the trajectories by constraining feature matches to that very direction. Whereas this strategy obviously works only for some simpler cases of object motion, it nevertheless will provide a much more robust feature matching. In addition, we want to stress that our focus on visual features and their transformations between visual events is a much broader concept not restricted to object motion alone. Going beyond a simple matching prior, this information can also be used to explicitly model generic object or category transformations (Graf 2002), which expands the keyframe framework into a general learning concept for any dynamic visual data.

3.3 Computational Experiments

In the following, we will briefly present results from computational experiments, in which we tested the performance of the keyframe implementation on a highly controlled database (for details of the implementation as well as additional experiments also including real-world video sequences, see Wallraven and Bülthoff 2001).

Stimuli: The database consisted of 60 sequences of faces taken from the MPI face-database (Troje and Bülthoff 1996). This database contains highly realistic 3D laser-scans of faces and allows full control of all aspects of rendering (pose, lighting, shadows, scene, etc.) for benchmarking recognition algorithms. Each sequence showed a face turning from –90° (left) profile view to +90° (right) profile view consisting of 61 frames at pose intervals of 3 degrees. All faces were rendered from a viewing distance of 1.3 m on a black background using a frontal point-light source. Our test sets consisted of images from the same sequences in addition to novel images containing pose variations of +/–15° (upwards and downwards) as well as two different illumination directions.

Keyframes: Using the local feature tracking algorithm described above, the system found 7 keyframes for each of the 60 sequences (Fig. 3a shows some example keyframes and their average poses). Furthermore, the angular distance between subsequent keyframes is smallest for the frontal poses (between keyframes 3 and 5). This is due to the fact that a rotation around the frontal view causes larger variations in features (such as ears disappearing and appearing) leading to an earlier termination of tracking. Note also that even

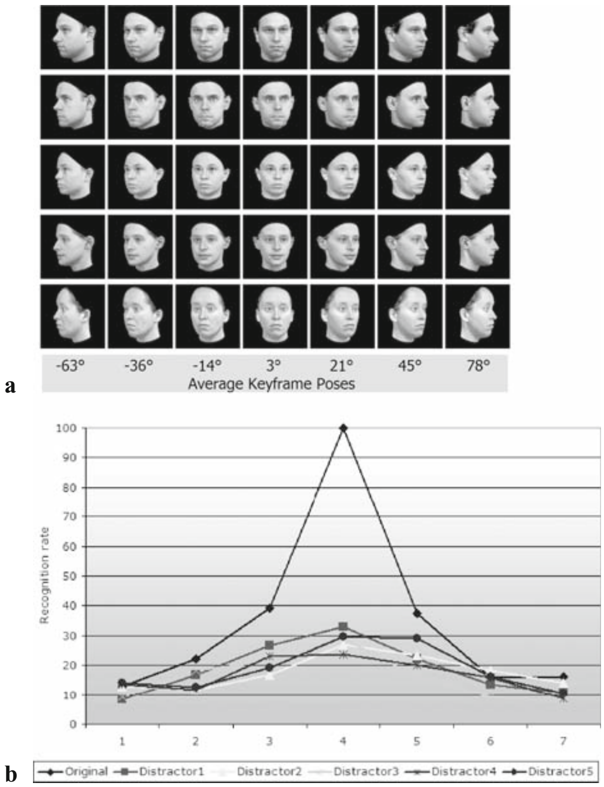

Fig. 3. **a** Examples of keyframes extracted from image sequences of rotating faces. The bottom figures list the average pose in degrees found across the whole database. **b** Matching scores for 6 "novel" faces. Note that the target face has a much higher matching score than the five other distractors

though the framework itself has no prior knowledge about the object class or the object motion, similar views are selected as keyframes. This is a demonstration that our framework is able to generate consistent representations provided the input also exhibits consistent characteristics. Finally, the representation of each image sequence consists of a number of keyframes containing local features, resulting in a significant, size reduction. This is an essential property for any view-based system working with dynamic data since otherwise huge amounts of data would have to be stored. To calculate the size reduction of the representation, we compared the size of the final sequence models to the raw pixel data and determined a reduction of 99.1% (7 keyframes compared to 61 original frames corresponds to a reduction of ~90%; each keyframe contains ~200 local features, each of which consists of 5 × 5 pixels. Given the original image size of 256 × 256 pixels, this results in a reduction of ~92% per keyframe).

Recognition results: Our first recognition experiment concerned a validation of whether the resulting keyframe representation could support recognition of intermediate views of the original sequences. We therefore tested all keyframe representations with the remaining $(61 - 7) * 30 = 1620$ frames not included in the keyframe representation, which resulted in a total recognition rate of 100%. To illustrate the robust matching, Figure 3b shows the matching scores for a set of keyframes with one target image and 5 distractor images, all of which show the same view. First of all, one can see that the target has a much higher matching score than the distractors. Interestingly, the highest match score for the distractors is almost exclusively achieved for the correct pose. In addition, all curves show a consistent view-based behaviour with a fall-off around the best matching keyframe. Recognition rates in the second experiment testing novel views under pose and illumination variation were 98.6% and 89.4%, respectively. Although pose variation is tolerated well by the system, changes in illumination clearly show the limits of the simple matching scheme. Taking the amount of compression into account, however, we think that these results demonstrate the feasibility and robustness of our approach (see also Wallraven and Bülthoff 2001).

4 Multi-Modal Keyframes

So far, the keyframe framework has been treated as a framework for recognition of objects in the visual modality. The general idea of spatio-temporal object representations, however, can of course be extended to other modalities as well. In the following, we will introduce such a multi-modal object representation combining visual with proprioceptive information, which was successfully implemented on a robot-setup and subsequently tested in object learning and recognition scenarios.

Recent research in neuroscience has led to a paradigm shift from cleanly separable processing streams for each modality towards a more integrative picture consisting of multi-modal object representations. Such cross-modal integration of data from different modalities was also shown, for example, to play an important role for haptic and visual modalities during object recognition. In a recent psychophysical experiment (see Newell et al. 2001), participants had to learn views of four simple, 3D objects made of stacked LEGO™ bricks either through the haptic modality (when they were blind-folded) or through the visual modality (without being able to touch the objects). Testing was then done using an old-new recognition paradigm with four different conditions: two within-modality conditions, in which participants were trained and tested in either the haptic or the visual domain and two between-modality conditions, in which information from the learned modality had to be transferred to other modalities in order to solve the recognition task. For each condition, in addition, either the same viewpoint or a viewpoint rotated 180° around the vertical axis was presented in order to test the viewpoint-dependence of object recognition.

The recognition results for the four conditions showed first of all that cross-modal recognition occurred at levels well above chance. Not surprisingly, recognition of rotated objects in the within-modality condition was severely affected by rotation in both modalities. This shows that not only visual recognition is highly view-dependent but also that haptic recognition performance is directly affected by different viewing parameters. One could thus extend the concept of view-based representations of objects also to the haptic modality. Another interesting finding of this study is that recognition performance in the haptic-to-visual condition increased with rotation. The authors assumed that this was an example of a true cross-modal transfer effect – the reason for such a transfer lies in the fact that during learning the haptic information extracted by participants was mainly derived from the back of the object. When presented with a rotated object in the visual modality, this haptic information was now visible, which enabled easier recognition. The results from this experiment thus support the view that haptic recognition is also mediated by view-based processes – although the exact dependence on viewing angle remains to be investigated. In addition, the authors shed light on how information from the haptic modality can be used to enable easier recognition in the visual modality. Taken together with the spatio-temporal framework outlined above, this cross-modal transfer might be an important reason for the excellent visual performance of human object recognition – after all, it is known that infants learn extensively by grasping and touching objects, which thus could provide a "database" of object representations for visual recognition.

4.1 Multi-Modal Keyframes – the View-Transition Map

Taking these psychophysical experiments as inspiration, we now want to describe how visual and proprioceptive input can be combined to create and test a multi-modal keyframe representation.[1]

Let us consider a person who is examining an object by holding it in their hand and turning it around – the sensory information that is available in this situation consists of not only dynamic visual data but also dynamic haptic information. More specifically, we will focus on the proprioceptive information as a subset of the haptic information, which consists of the 3D configuration of the hand (such as the exact configuration of the fingers holding the object) as well as that of the wrist. How could this information be of use for learning and recognition?

First of all, proprioceptive information about the 3D configuration of the hand could actually be used in a similar manner as in the psychophysical experiment described in the previous section. Since it is three-dimensional, it can for example generate a 3D viewing space in which keyframes (derived from the visual infor-

[1] The multi-modal representation, as well as the experiments were developed in collaboration with Sajit Rao, Lorenzo Natale, and Giulio Sandini at the Dipartimento di Informatica, Sistemistica e Telematica at the University of Genoa.

mation of the image sequence) can be anchored at proprioceptive coordinates. This would link the visual appearance from the keyframe with the hand position and configuration and thus provide a proprioceptively anchored visual space. Returning to Figure 2, we see that one of the inherent disadvantages of the keyframe framework is that the real-world topology of the keyframe graph is undefined – only the outgoing and incoming links for each keyframe are known. Although this provides enough information to resolve recognition tasks (see previous section), being able to convert the viewer-centered keyframe graph into an object-centred keyframe graph would provide additional constraints for matching visual appearances since such a representation would be more closely integrated into a perception-action loop.

One of the problems with the idea of proprioceptive space, however, is that absolute coordinates in such a space make little sense from the perspective of recognition. Although it might be the case that objects suggest a canonical grasp (in much the same manner as they might suggest an affordance in the Gibsonian sense), usually it is possible to pick up and hold an object in a number of ways – all of which will change the absolute proprioceptive coordinates to which keyframes will be attached. Our solution is to interpret the proprioceptive space in a similar manner as the keyframe graph: as a representation based on *changes* in its underlying modality. Thus, rather than using an absolute frame of reference, each generated keyframe could be attached to a relative change in proprioceptive coordinates. One way to implement such a multi-modal representation is as a lookup table, in which each entry can be accessed via its relative change in proprioceptive space – this change can, for example, be simply the difference between the proprioceptive state vectors of the hand (including wrist angles, finger positions, etc.). This novel representation – which we call a view transition map – would for *n* visual keyframes consist of *n(n – 1)* entries for all possible proprioceptive transitions between keyframes.

How could one use this view-transition map to recognize objects? First of all, a keyframe representation of an object is learned in an active exploration stage using a pre-learned motor program, which for example grasps an object and turns it around. Each new keyframe is entered into the transition map at a position specified by the relative change in proprioceptive state from the previous keyframe. In addition, the transition map is enlarged by adding transitions from this keyframe to all previous ones. In a second step, a test object is picked up and keyframes are extracted again while the same motor program is executed. In order to recognize this object using the transition map, the first keyframe that was generated from the test sequence is matched against all of the keyframes of the training sequence using visual similarity (in our implementation, similarity consisted of simple Euclidean distance – using local feature matching would further increase the robustness of the system). Once this match has been established, the transition map can be used to quickly find neighboring keyframes by looking for the most similar proprioceptive transition from the keyframe that matches the current change in the proprioceptive state. With this strategy one could expect to recognize objects in a much more efficient manner as indexing proprioceptive transitions allows for direct matches in an object-centered reference frame.

4.2 Computational Recognition Experiment

The proposed view transition map representation was tested in a computational experiment in which we explored its use for object recognition.

Experimental setup: Figure 4a shows the robot setup from the Dipartimento di Informatica, Sistemistica e Telematica at the University of Genoa that was

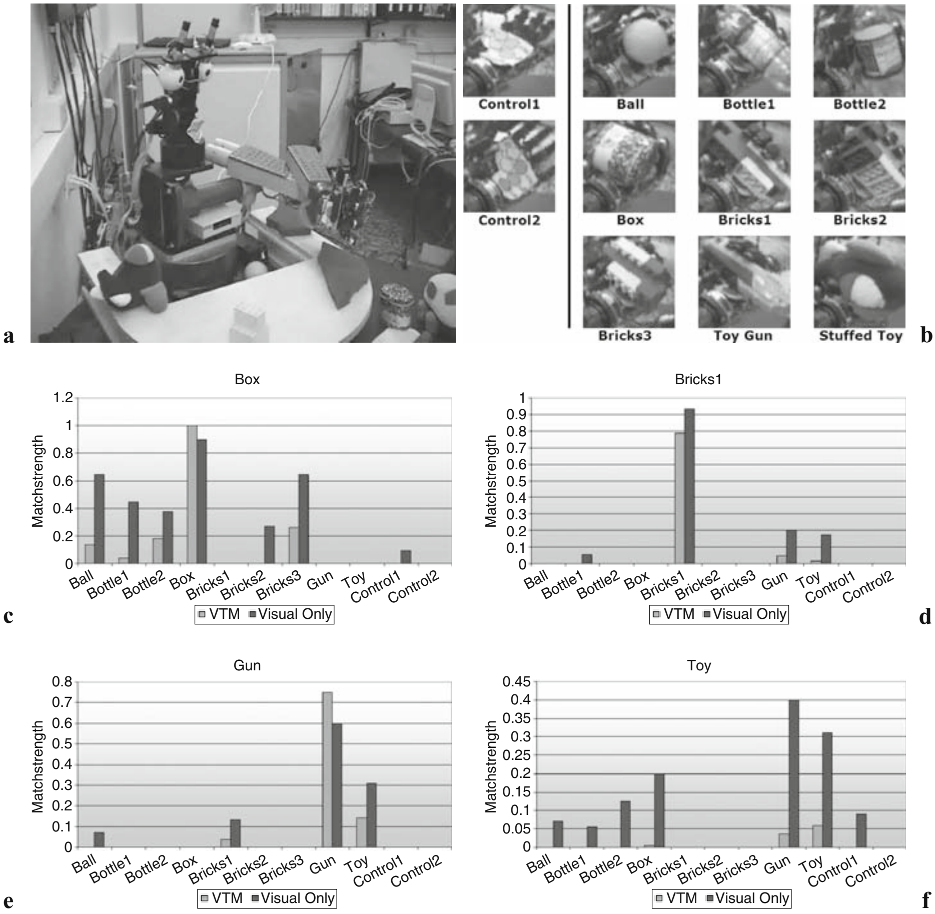

Fig. 4. **a** The robot setup (Metta et al. 2000) that was used in the multi-modal keyframe experiments **b** the objects used in the learning and recognition experiments. **c–f** Example results from the object recognition experiment showing the increase in discriminability when using multi-modal representations. The bright bars show matching using the view transition map, the dark bars show visual matching only

used in this experiment (Metta et al. 2000). The important components of the setup for this experiment consist of an actively foveating stereo camera head (using space-variant image sensors mimicking the human retinal structure) and an anthropomorphic robotic arm with a fully controllable hand. The camera head was pre-programmed to fixate on the location of the hand in order to track the hand during all movements. In addition, a trajectory for the hand movement was defined, which consisted of the hand rotating first around the axis defined by the arm ("turning the hand") and then around a second axis resulting in an up-and-down movement of the hand. This exploratory motion sequence ensured an adequate visual coverage of any grasped object. The test objects for the experiments consisted of 9 household and toy objects and are depicted in Figure 4b – note that some of the objects are rather similar in terms of their visual appearance.

In order for the robot to learn an object, it was placed into the robot's hand and the exploratory motion sequence was initiated. The visual input from the foveated cameras was then used to track local features in real-time using the keyframe framework as described in the previous section. Each time the system found a keyframe, the proprioceptive transition leading from the last to the current keyframe was used as an index into a matrix where each entry stored the visual information of the frame (in this case simply consisting of the whole frame rather than its local feature representation). In addition, each incoming keyframe was matched against all existing keyframes using the view-transition map matching procedure outlined above. If a match of suitable strength was found, the keyframe was discarded, otherwise the keyframe was inserted into the representation. A total of 9 objects were learned in this manner; in addition two control conditions were recorded, which simply showed a sequence of empty hand moving.

Recognition results: To test recognition performance, six of the objects were given again to the robot and the same movements were executed. Each new keyframe was then compared against all learned transition maps using the matching procedure described above and the amount of matches in each transition map was added up to a final matching score. If the sequence would be identical, all keyframes would be found in the map and therefore the matching score would be 1. To provide a baseline, visual-only matching was also run in addition to the multi-modal matching procedure. Figure 4c–f shows histograms of the matching scores for the two matching procedures for four test-objects. For the "box" object, both procedures correctly predict the right category; the multi-modal matching, however, has a much higher discriminability compared to the visual-only matching. The same is true for the "bricks1" and "gun" object. Finally, the "toy" object is an example of an object, which is not correctly recognized by visual-only but is recognized by the multi-modal matching.

Summary: The results of these initial computational experiments are very encouraging. Through a straightforward extension of the keyframe approach to include proprioceptive information, we have shown how multi-modal object representations can be learned as well as how such representations can help to increase the discriminability of object recognition. Since our representation is in part three-dimensional (i.e., coupled to proprioceptive coordinates), some of the robustness actually comes from 3D information in a viewer-/manipulator-centred coordinate system. It would be interesting to see how such a representation might capture the results in the chapter by Gschwind et al. (this volume) on exploration of 3D shapes.

Among several extensions that can be envisioned, adding more sophisticated local feature matching, better classification schemes as well as different cue combination approaches should further improve the performance of the framework. Another interesting property of the transition map is that it enables execution of specific motor actions based on visual input. Consider, for example, a situation in which an object has to be manipulated in order to insert it into a slot. The inverse of the transition map would allow such a task to be solved by executing motor commands that trace out a valid motor path to the desired view based on the current view. In a similar manner, the transition map could also be used for efficient imitation learning based on visual input and for executing mental rotations. The key to all of these applications is that the transition map provides a strong coupling between proprioceptive data (action) and visual data (perception) and in this manner facilitates representation of a perception-action loop in an effective and efficient way.

5 Conclusion

In this chapter, we proposed an abstract framework for learning and recognition of objects that is inspired by recent psychophysical results which have shown that object representations in the human brain are inherently spatio-temporal. In addition, we have also presented results from a computational implementation of this keyframe framework, which demonstrate that such a system can reliably recognize objects under a variety of conditions. Finally, experiments with multimodal keyframes have shown that by integrating non-visual cues, object learning and recognition becomes more efficient and effective. We believe that this framework can represent a significant step in designing and implementing a truly cognitive system, which is embedded in a constantly changing environment and thus has to constantly analyze and learn in order to plan its (re-)actions.

References

Aloimonos JY, Weiss I, Bandopadhay A (1987) Active vision. Int J Comput Vis 1:333–356

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

Blanz V, Vetter T (1999) A morphable model for the synthesis of 3d faces. Proc ACM SIGGRAPH 1999:187–194

Bülthoff HH, Edelman S (1992) Psychophysical support for a 2-d view interpolation theory of object recognition. Proc Natl Acad Sci U S A 89:60–64

Foster DH, Gilson SJ (2002) Recognizing novel three-dimensional objects by summing signals from parts and views. Proc R Soc Lond B 269:1939–1947

Graf M (2002) Form, space and object. Geometrical transformations in object recognition and categorization. Wissenschaftlicher Verlag Berlin, Berlin

Kirby M, Sirovich L (1990) Applications of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Trans Pattern Anal Mach Intell 12:103–108

Koenderink J, van Doorn A (1979) The internal representation of solid shape with respect to vision. Biol Cybern 32:211–216

Krieger G, Rentschler I, Hauske G, Schill K, Zetzsche C (2000) Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. Spat Vis 13:201–214

Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110

Marr D (1982) Vision. Freeman Publishers, San Francisco

Metta G, Panerai F, Sandini G (2000) Babybot: a biologically inspired developing robotic agent. Proc. Sixth International Conference on the Simulation of Adaptive Behaviors, 1–10

Miyashita Y (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature 335:817–820

Newell FN, Ernst MO, Tjan BS, Bülthoff HH (2001) Viewpoint dependence in visual and haptic object recognition. Psychol Sci 12:37–42

Roberts L (1965) Machine perception of three-dimensional solids. In: Tippett J, Clapp L (Eds) Optical and electro-optical information processing. MIT Press, Cambridge MA, pp 159–197

Schmid C, Mohr R (1997) Local greyvalue invariants for image retrieval. IEEE Trans Pattern Anal Mach Intell 19:530–535

Stone JV (1999) Object recognition: view-specificity and motion-specificity. Vision Res 39:4032–4044

Swain M, Ballard D (1991) Color indexing. Int J Comput Vis 7:11–32

Tarr M, Bülthoff HH (1998) Object recognition in man, monkey, and machine. MIT Press, Cambridge MA

Tomasi C, Kanade T (1991) Detection and tracking of point features. Carnegie-Mellon Tech Report CMU-CS-91–132

Troje NF, Bülthoff HH (1996) Face recognition under varying pose: the role of texture and shape. Vision Res 36:1761–1771

Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5:682–687

Wallis GM (2002) The role of object motion in forging long-term representations of objects. Vis Cogn 9:233–247

Wallis GM, Bülthoff HH (2001) Effects of temporal association on recognition memory. Proc Natl Acad Sci U S A 98:4800–4804

Wallraven C, Bülthoff HH (2001) Automatic acquisition of exemplar-based representations for recognition from image sequences. In Proc. CVPR'01 – Workshop on Models versus Exemplars