Focus of Attention on Relevant Multimodal Events

A Developmentally Inspired Architecture for Active Vision

by

Miranda Grahl

Dissertation

submitted to the

Faculty of Technology at Bielefeld University

in partial fulfillment of the requirements for the degree of

Doktor der Ingenieurwissenschaften

(Dr.-Ing.)

December, 2011

A dissertation submitted to the Faculty of Technology at Bielefeld University for the degree of Doktor-Ingenieur (Dr.-Ing.) on December 13, 2011.

Reviewed by:

Prof. Dr.-Ing. F. Kummert	Bielefeld University, Bielefeld, Germany;
Dr.-Ing. F. Joublin	Honda Research Institute Europe GmbH, Offenbach/Main, Germany;

Accepted on April 26, 2012, on behalf of the Faculty of Technology at Bielefeld University, Germany, by the following dissertation committee:

Prof. Dr. P. Cimiano	(chairman)
Prof. Dr.-Ing. F. Kummert	(advisor)
Dr.-Ing. F. Joublin	(co-advisor)
Dr.-Ing. Hendrik Koesling	

Miranda Grahl, ≫Focus of Attention on Relevant Multimodal Events≪

Printed on permanent paper °° ISO 9706.

Danksagung

An erster Stelle möchte ich mich bei Prof. Dr.-Ing. Franz Kummert, meinem Betreuer an der Universität Bielefeld, sowie Dr.-Ing. Frank Joublin vom Honda Research Institute Europe (HRI-EU) dafür bedanken, dass sie mir die Möglichkeit gegeben haben, am Institut für Kognition und Robotik der Universität Bielefeld meine Promotion durchzuführen. Dem HRI-EU danke ich in diesem Zusammenhang für die finanzielle Unterstützung in Form eines Stipendiums. Mir hat es große Freude bereitet, in einem solch interessanten Forschungsumfeld arbeiten zu dürfen.

Meinen Betreuern, Franz und Frank, danke ich zudem für deren kontinuierliche Unterstützung. Beide hatten immer ein offenes Ohr für meine Angelegenheiten und verstanden es, meine Arbeit in Form von Ideen und Hinweisen mit voranzutreiben. Unsere lebhaften Diskussionen werde ich sicherlich in Erinnerung behalten.

Des Weiteren möchte ich mich bei meinen Kollegen aus der Arbeitsgruppe Angewandte Informatik bedanken. Sie sind dafür verantwortlich, dass ich in einer angenehmen Arbeitsatmosphäre forschen durfte. Insbesondere danke ich meinem Büronachbarn Dipl.-Inform. Marc Kammer für Diskussionen, die auch einmal über den Rand des Computerbildschirms hinweg gingen.

Danken möchte ich außerdem den Korrekturlesern meiner Doktorarbeit und entstandenen Publikationen: Dr. rer. nat. Sina Kühnl, Dr.-Ing. Claudius Gläser, Dipl.-Inform. Andrea Finke, Dipl.-Inform. Alexander Lenhardt und M.A. Griffiths Sascha. Ihr Feedback hat wesentlich zur Verbesserung dieser Arbeit beigetragen.

Persönlich möchte ich mich bei meinen Eltern und meinen Großeltern bedanken, die mir immer zur Seite standen und mich bestmöglich unterstützt haben. Außerdem danke ich meinen Freunden für das Daumen drücken und die Abwechslung von der Arbeit, die Ihr mir beschert habt. Dir ganz besonders Ulrike. Und natürlich meinem Freund Claudius, der es auch in schwierigen Zeiten verstand, mich immer wieder aufs Neue zu motivieren. Tschakka. Auf gehts!

Dezember, 2011 M. Grahl

Abstract

Future robots should autonomously operate in their environments. The autonomous exploration and understanding of scenes constitutes one of the key challenges that have to be solved on the way towards this goal. Thereby, existing research follows different strategies. One of the most promising approaches – which will be pursued in this thesis, too – is to take inspiration from infant development. Similar to robots, infants are confronted with a wealth of information that initially can not be interpreted by them. Over the course of development, however, infants acquire capabilities that allow them to explore and understand previously unknown scenes.

In this respect, the ability to focus on multimodal events is of particular importance in this thesis. The present work aims at developing a computational framework which allows a robot to control its gaze on important aspects of its surrounding. Here, we define important aspects as those scene elements which exhibit characteristic features in multiple sensory modalities. The development of the framework is based on principles known from infant development. One of these principles is the transition from an initially reactive gazing behavior, which is solely based on bottom-up visual filtering processes, towards an expectation-driven gazing, which makes use of already acquired object knowledge.

The proposed framework comprises multiple computational methods which may be used by an autonomous robot to exhibit a similar gaze development. Firstly, a model for unsupervised visual object learning is presented. The method autonomously gathers objects knowledge during the exploration of a scene. Thereby, scene exploration is initially reactive. Later on, however, already acquired knowledge can be used to bias gaze selection in an expectation-driven manner. Additionally, the framework comprises a model for multimodal association learning. The method combines learned visual knowledge with information from other sensory modalities – particularly auditory object characteristics. Once a multimodal association is achieved, visual as well as auditory object knowledge can be used to guide robot's attention. The methods are evaluated in simulations using real video sequences. An integration in a robot remains for future work.

Contents

1. Introduction

Infants are initially equipped with a minimal set of innately given cognitive abilities. Over the course of development, however, they rapidly gain experience from interactions with the environment. Thereby, infants exhibit enormous learning capabilities that by far exceed those of current artificial systems. Infancy research hence constitutes an important source of inspiration that may ultimately lead to a way to overcome the restrictions of current artificial systems.

Studies in the field of developmental psychology already served as motivation for the development of different robotic systems (Lungarella et al., 2003). In the present work, particular emphasis will be given to the development of visual competencies. Like newborns autonomous robots initially act in an unknown environment. They are confronted with a richness of information stemming from different sensory modalities. Equipping the robot with mechanisms that allow an efficient exploration of this information, an unsupervised learning from the input stream, and finally the use of the gathered knowledge are key problems that have to be solved on the way towards building robots that possess child-like abilities. This also includes questions like what kind of knowledge has to be predefined (i.e. innately given to the system) or which other sensory modalities may have an influence on visual development. Infant development can serve as a road map in this respect.

An example for a learning situation, in which infants are typically engaged, is the one where objects are joggled and additionally augmented with sound by a tutor. A similar scenario can occur for robot learning. For example a humanoid robot such as the iCub (Metta et al.) may attend to an object, e.g. a red ball presented by a human. But how should the robot behave, if an object is presented with object-specific auditory characteristics like the sound of a rolling ball? Furthermore, how should it behave if a speaking person is interacting with the robot while holding the red ball in its hand?

In humans, the sound of a rolling ball arouses attention. This includes both auditory attention and visual attention, since we are interested in why the ball makes noise and in which direction the ball rolls. Similarly, during a communication we expect the gaze towards our face and are confused if our opponent is looking somewhere else. We are able to control the gaze towards audiovisual events, because we have learned to suppress incoherent audiovisual distractors that do

not stem from the same object. Therefore, it is desirable to equip a robot with a gazing behavior towards coherent multimodal aspects. Existing artificial systems lack an adequate gazing behavior because they miss a learning of audiovisual information. In contrast to this, such a learning exists in early infancy. This means that infants are able to process and recognize complex objects and learn to extract relevant audiovisual information.

A proper robot behavior in such situations presumes a detection and a memorization of relations between visual and auditory inputs. Thereby, the learning of audiovisual object representations strongly relies on an appropriate synchronization of the inputs from both modalities. In the example, such audiovisual object knowledge may underlie the response of the robot to shift its gaze away from the ball. More precisely, the heard speech can trigger the robot to attend to the person's mouth.

1.1. From Human to Artificial Vision

From an engineering point of view, it is impossible to specify all situations during the design phase, that the robot will be confronted with during operation time. Therefore, it is important that a system continuously extends its knowledge on demand in order to appropriately behave in novel situations. To bootstrap such a developmental process, some minimal innate behavior structure and scene knowledge may be required. Furthermore, an unsupervised acquisition of object knowledge is needed. The learning of relevant scene information implicates that the robot attends to it. For this reason, a gaze control strategy as part of an active vision system constitutes an integral part of an appropriate system architecture.

In humans, eye movements are the result of two complementary processes in the brain: Firstly, a bottom-up process resulting from early retinal filtering which produces "saliency driven eye movements" and secondly, a top-down process based on learning, memorization, and cognition that results in "task driven eye movements". An exploration of a scene is characterized by *saccades* and *fixations*. Thereby, *saccades* are rapid eye movements that shift an interesting location in the center of the visual field, whereas a *fixation* enables stable visual information processing such that aspects from the environment can be learned. The complexity of image processing algorithms constitutes a severe problem with respect to the management of computational resources and memory. Therefore, in machine vision the term *active vision* assumes a *foveation* of the processed visual information. This means peripheral observations are less weighted and the main processing focuses on a small *region of interest*. This is in contrast to *passive vision*.

Object recognition in an online learning scenario comes along with a wide range of challenges. For example, most vision-based architectures suffer from training with hand annotated data resulting in an inflexibility regarding the recognition of new objects not covered by the training set. This means that those architectures are built for known operation scenarios that are defined during the design time. So far, little work has been done in the computational modeling of an object recognition process, that automatically extracts a structure out of auditory and visual cues in order to gain object knowledge, i.e. to build up a system that incrementally learns an object representation. In order to recognize objects, an active exploration system needs to remember learned scene information. For example, Itti et al. (1998) proposed a computational model for a data driven exploration of the scene by means of a gaze control towards visually salient regions. The model allows to specify salient object locations but misses an adequate object-driven search. This is because a process that acquires visual object knowledge is missing, such that a robot cannot attend to similar objects at a later time.

Findings from Cognitive Psychology point to the fact that the ability to track and to recognize objects is strongly influenced by additional modalities like haptics or audition. For example, Newell (2004) proposed that an object identity can be maintained by a multisensory representation linking vision with haptic cues. Xiao et al. (2007) studied the effect of task irrelevant sound on the oculomotor system. The analysis with different pitch deviants showed that the ability to follow a moving object increases with an increase in pitch. Lehmann and Murray (2005) investigated the influence of past audiovisual object representations on an unimodal object recognition task. Memory performance was improved if an object has been perceived in both modalities before. Molholm et al. (2004) also reported that an audiovisual representation leads to a faster and more accurate object detection. She hypothesized that auditory input modulates visual brain areas by which object recognition gets biased.

1.2. Problem Statement

Current artificial systems for active vision do not tackle an autonomous learning of a gaze control that focuses on multimodal aspects of the environment. They are either based on a reactive gaze control or presume defined models for visual and auditory input in order to attend to objects. However, it is well researched that infants do not rely on predefined models (Spelke, 1981). They rather flexibly learn audiovisual associations: For example associations that comprise a temporal synchrony between faces and speech (Flom and Bahrick, 2007), or between moving toys and objects' rhythmic sound characteristics (Flom and Bahrick, 2010).

Evidence in favor of this view is provided by the work of Richardson and Kirkham (2004). In an experimental study, saccade movements of six month old infants were analyzed with respect to audiovisual information. In a learning phase, infants were introduced to look on a screen. On this screen, a bouncing toy was presented with a rhythmic sound. The toy appeared in a rectangle either on the left or right side of the screen. The sound was always produced in the center of the screen. Subsequently, in a testing phase the gaze behavior towards the learned toy location was analyzed in the presence of an associated sound. The toy location was presented with an empty rectangle whose position was either not modified or rotated. In both conditions, infants showed a longer gaze fixation to the location that was learned with a sound. This gazing behavior was interpreted by Richardson as a kind of visual prediction mechanism to object locations in the presence of an object specific sound.

The finding of this experiment can serve as an inspiration for designing a gaze control strategy for robots. As a key aspect, it suggests a modulation of a visual filtering process by auditory modalities. This entails the question on how to equip a system with a minimal innate perceptual knowledge and a mechanism that learns to structure the perceptual information by itself. This also means to overcome the problem of audiovisual correlation in the location cue and hence implies the need for a direct measurement of causality between visual and auditory concepts.

1.3. Relation to Infant Development

In the following, the visual competencies of an approximately eleven month old infant are illustrated in an object learning scenario. This example describes details on infants' object learning and gaze control capabilities to cope with audiovisual scene information that is eligible for an artificial system. The image sequence is extracted from a video corpus that is part of an infant study which took place at the Bielefeld University *. In this study, parents were invited to demonstrate different objects and their functions to their infants. One kind of objects were cups of different colors and the task was to explain how to stack them.

Fig. 1.1 shows a tutoring situation in which a father aims to teach his child "Rasmus" the usage of cups. The presentation of the objects is complemented by speech as well as sounds produced by the objects themselves. The accompanied acoustic segments are highlighted with the color red below the figures. Additionally, the father shakes the cup in order to capture the infant's attention. The infant consistently fixates the object (Fig. 1.1(a)-Fig. 1.1(b)) with exception of

*Motionese Corpus, Bielefeld University 2006

(a) green cup and 'look' (b) green cup and 'cup clutter' (c) yellow cup and 'then' (d) yellow cup and 'hello rasmus'

Figure 1.1.: A cup stacking task is presented to a child. The father explains the functions of different cups. The images (a)-(d) illustrate the interaction sequence.

the combination yellow cup and 'then' (Fig. 1.1(c)). After a repetition of this exposure with 'hello Rasmus' (Fig. 1.1(d)), the infant seems to be interested in again, shifts its gaze on the yellow cup, and tracks it.

The infant moves its gaze to the position, where a coherent object representation in both modalities occurs. At the same time, it inhibits gazing towards the mouth of the father. This implicates that the infant already built an object representation for the cup during a learning phase. This representation comprises a visual concept, an auditory concept derived from the coherent speech, and a location in which the object has been previously observed. The inhibition of the "mouth-speech-coherence" indicates that the infant learns to enhance the attention for the cup in the presence of speech. Therefore, it is reasonable to research the visual information processing system of infants in order to infer possible principles that can be likewise applied in artificial systems.

Onset of Memory Based Eye Movements

An automatic recognition and an active gaze control towards objects in the presence of environmental sounds requires the understanding why infants can detect and discriminate temporal coincidences of audiovisual events (Hollich et al., 2004). Further, it requires knowledge about the learning mechanism of audiovisual events that are not associated with a specific location (Morrongiello et al., 1998). This implicates the question how infants configure a visual filtering process in the presence of auditory input in order to increase the sensitivity to an object that is repeatedly presented during a learning phase. In other words, infants need to suppress reflexive eye movements towards auditory events to follow a voluntary gaze control towards visual aspects. The literature (Spelke, 1994, 2000) shows that infants start to explore their environment with a minimal scene

knowledge and begin to learn visual object representations. Afterward, learned object representations are combined with other sensory modalities.

In the first months of life, infants' gazing on objects as well as the tracking of them is reactive (Von Hofsten and Rosander, 1996). Later on, learned competencies substitute these reactive mechanisms. The age of three months defines the onset of such memory based eye movements (Von Hofsten and Rosander, 1997). At birth an infant is equipped with an immature visual system. The ability to process visual information is restricted to the periphery of the visual field and leads to an incomplete perception of an object. In detail, detection of object characteristics is limited to conspicuous lines and edges (Fogel, 2000). The oculomotor skills are not fully developed. Hence, infants fail to sustain attention to an object shortly after birth. This means infants innately direct attention only on simple visual stimuli (e.g. color and intensity) and are not able to negotiate stable visual information processing on attended information (Frank et al., 2009). With three months infants start to develop an active coordination strategy towards objects and develop visual expectations about their environment that are successively improved during their first year of life. At that time, infants are able to track moving objects.

The influence of attended locations on gaze control is characterized by two aspects that are important for sustaining attention on objects and consequently for a stable visual information processing. Firstly, a saccade to a visual stimuli occurring in the peripheral visual field is facilitated, if the stimuli has been already cued during attending another object. The second aspect relates to the ability of suppressing previously attended locations in order to focus on new salient targets (*inhibition of return*). Richards (2004) showed that three to six month old infants continuously benefit from the *facilitation effect* but the performance differs with respect to the ability to inhibit previously attended locations. The result showed that five month old infants start to coordinate visual attention. Amso and Johnson (2008) found that two to six month old infants missed the suppression of a distractor stimuli in the periphery during targeting a visual stimuli. Moreover, in this study nine months old infants showed the maintenance of the distractor inhibition over different time delays. This means infants are able to memorize the suppression of a distractor cue and hence show delayed saccades. Furthermore, this memorization ability implicates a discrimination between a learned target and a distractor.

Johnson (1990) related this development to the cortical maturation of the visual system. He suggested that the primarily orienting gaze behavior in one month old infants is characterized by a saccadic following of moving objects, whereas the ability to predict an object location is missing. Two month old infants show the

ability to track *smooth pursuit*. This means the location of a moving target is estimated more correctly, but the gaze fixation to the new location is characterized by a time delay. A more precise estimation of a moving target takes *anticipatory looking* into account, i.e. predicting the position of a moving target. This ability is achieved during three to six months and is mediated by the stronger recruitment of the frontal eye field which is a region involved in memory based saccade and intended smooth pursuit, but also with the development of area MT and MST involved in the processing of optical flow and motion prediction.

The development of active saccade generations requires the learning of visual expectations during the exploration of a scene. This firstly comprises the memorization and retrieval of internal object representations (Henderson, 2003). Hereby, the object representations are learned from observations in the central visual field. Secondly, to develop visual expectations, causal relations between multiple scene events have to be determined- a process commonly known as *contingency learning*. In conjunction with *anticipatory looking* it defines a starting point for the development of visual expectations for an active generation of eye movements. Both learning circuits develop dependently from each other and Johnson et al. (1991, 1994) hypothesized that both circuits inhibit each other during early visual development. Both abilities are emphasized with the age of four months. During this developmental state, infants improve their ability to attend to familiar objects and are able to habituate to them (Fogel, 2000).

Influence of Audition on Gaze

The depicted tutoring situation (Fig. 1.1) well demonstrates the impact of speech on infants' gaze control. The appearance of speech helps the infant to redirect the gaze towards the yellow cup. This indicates that infants use the auditory modality for configuring a visual filtering mechanism such that an active gaze coordination is realized. The following developmental aspects emphasize how infants learn these coupled object representations.

Due to the immaturity of the visual system, the gaze control is initially dominated by acoustic properties of the environment. The auditory dominance gradually decreases with an increase in age (Robinson and Sloutsky, 2010). An active processing of multimodal information starts at an age of five months. For example it has been shown that a visual stimulus is fixated for a longer time, if a novel rhythm is presented at the same time (Bahrick and Lickliter, 2000). This suggests an active discrimination of acoustic properties that temporally coincidence with visual information. This means an innately arbitrary crossmodal processing becomes coordinated as development progresses and infants start to

control their gaze towards audiovisual aspects by suppressing the initial auditory dominance, i.e. they control the gaze independent of the auditory sound location.

The gaze control towards an object may require an active excitation of learned visual object representations in the presence of an acoustic input, i.e. an active selection of learned visual concepts that previously have been associated with acoustic properties. The integration of such top-down knowledge has been observed in one year old infants (Gliga et al., 2010), where an enhancement of brain activities in the visual cortex was measured during the presentation of objects with learned verbal labels.

In summary, the literature shows that infants learn to attend to objects with a minimum of predefined scene knowledge. This minimum becomes evident by reactive saccades (Aslin, 1988). This gazing behavior can serve as an inspiration for equipping an artificial system with an initial mechanism that drives the learning of audiovisual object representations. Furthermore, the learning of an inhibition mechanism enables infants to discriminate between different visual stimuli. For machine vision, this mechanism gives important insights in the design of an appropriate object description that benefits from peripheral scene information. Evidences from experiments that deal with audiovisual information suggest that visual processing is influenced by other modalities. Moreover, infants infer object locations based on a learned object specific sound which suggests a flexible processing between the different modalities.

1.4. Research Contribution

The research goal of this thesis is to develop a gaze control strategy that autonomously learns to focus on audiovisual events in its environment. The approach is inspired by findings from infancy research insofar as it aims at modeling the development from initially bottom-up driven reactive eye movement towards learned multimodal top-down attention. For this approach, different factors that contribute to the learning are investigated. The research contribution lies in the development of a computational model that acquires visual concepts in an unsupervised way and further configures those concepts in the presence of acoustic information.

1.5. Thesis Outline

This thesis is structured as follows. In Chapter 2, related work is reviewed and discussed. Thereby, neurophysiological and functional aspects of the visual sys-

tem are described and related to current computational models of attention. This includes models from biology as well as robotics. In Chapter 3, our approach towards a developmental active vision architecture is presented and motivated by findings from infancy research. Chapter 4 describes our algorithm for the incremental learning of object models during tracking and further analyzes the achieved performance with respect to the models' discrimination and generalization ability. Secondly, a pruning strategy of filters during object learning is proposed. This mechanism consolidates the acquired knowledge and show a way to overcome memory restrictions. Chapter 5 introduces a gradually learned weighting scheme between temporal coincidences of auditory and visual concepts. The learned weighting scheme is analyzed according to visual recognition performance in the presence of different acoustic classes. Chapter 6 finally discusses the proposed attention model and suggestions for future work are given.

2. Attention: Process, Gaze Control and Object Learning

The recognition of objects in our environment depends on many factors. For example, objects can be perceived by multiple senses, i.e. we can listen to a knocking hand, see the hand or we can touch it. These senses interact with each other and facilitate the recognition of the hand. Many research areas deal with the learning and recognition of such multimodal object representations. In this thesis the focus lies on the development of appropriate models that can explain such interactions between the modalities as well as their influence on the visual attention system. Research on visual attention is not only concerned with learning processes and recognition mechanisms. In particular, it deals with questions like how the human brain synchronizes different sensory modalities and how such processes can be computationally modeled or which role does the similarity between different sensory information plays in their filtering and memorization process.

This Chapter first reviews the different research methods that aim at modeling the interaction of audiovisual information. Then the influence of multimodal sources on the control of visual attention is described from a neurophysiological point of view. This Chapter continuous with a review of biologically motivated attention models, whose focus is the simulation of eye movements. These attention models are applied in robotics to control the gaze of artificial systems. Finally, this Chapter concludes by discussing the role of attention in social robotics, where existing models are reviewed and assessed with respect to their suitability for learning and recognition in tutoring situations.

2.1. Neurophysiological Process and Function

Eye movements are mainly driven by two processes of the visual system: a bottom-up process and a top-down process (Posner, 1980). Whereas the former enables a rapid processing by solely relying on sensory information, the latter modulates the bottom-up process by incorporating already acquired knowledge. The bottom-up process is based on a direct forwarding of visual information to the motor neurons that are linked to eye movements. Those *stimulus-driven* eye

movements are the result of an early retinal filtering process and exclude feedback information, e.g. those resulting from knowledge on the visual or auditory object properties. *Stimulus-driven* eye movements are triggered by a rapid change in the visual scene and lead to a reflexive gazing behavior. These kinds of eye movements are also called *exogenous*, since they do not incorporate learned information into a gaze movement towards an object. In summary, bottom-up visual attention is driven by visual aspects of the scene rather than by the observer's prior knowledge and serves for a rapid scene exploration.

The top-down process additionally incorporates feedback information from higher cognitive functions (Corbetta and Shulman, 2002). This feedback information may be the output of an object recognition process that comprises information on already learned visual aspects of the environment. The resulting eye movements are called *task-driven*, because they enclose the observer's prior object knowledge and thus allow to control the gaze towards known aspects of the environment. This prior knowledge can comprise, for example, the object color or the direction into which an object is moving. It is noteworthy that an object search can not only be biased by visual information of this kind, but also by learned acoustic object properties. For example, we use bird songs as prior knowledge to search for an object that corresponds to a bird.

A brain area that is relevant for a rapid visual scene exploration is the *superior colliculus* (SC). To execute a saccade, motor maps in the SC are accessed to specify a new gaze position. Thereby, the SC plays a key role for the coordination of reflexive saccades as it receives afferent projections from early processing levels, (e.g. direct input from the *retina*), the *occipital lobe* of the visual cortex, as well as the *frontal eye field* (FEF). The SC is directly influenced by inputs from its neighbor structure of the inferior colliculus (IC) that provides information about sound source direction in head centered coordinates. This sub-cortical structure is part of the auditory system and enables us to control the gaze towards acoustic events of the environment (Trepel, 2003). The integration of auditory sources occurs in one of the layer of the SC and specifies gaze fixation points to guide visual attention (Onat et al., 2007). Such interactions of audiovisual information at an early processing level were also studied by Frens and Van Opstal (1995). The authors investigated how audiovisual aspects are perceived and how they trigger a shift in attention. Therefore, saccade characteristics in the presence of such spatio-temporal events were examined. The results demonstrated a reduction of the saccade latency that is evoked by an enhanced synchronization of both modalities. This shows an important link between the synchronization of audiovisual events and eye movements.

Contrary to such *stimulus-driven* eye movements, *task-relevant* saccades incorporate aspects learned by the observer. This means that we need to remember and

retrieve these aspects while we look for something concrete in the environment. Thereby, learned object knowledge can be used for the purpose of different goals for an artificial gaze control: On the one hand, targeted objects can be biased using already acquired object knowledge. On the other hand, our attention system is equipped with a mechanism that allow to ignore stimuli in the periphery in order to stabilize the gaze. This is particularly useful to achieve a stable object fixation and to prevent sudden eye movements. One way to achieve such a resistance may be that the system can distinguish between aspects in the center of the visual field and those that are present in the periphery. This implies that the robot should not only draw attention to a fixated object, but also to those objects in the periphery. More precisely, the robot has to attend surroundings objects, but without bringing them in the center of the visual field. Such gazing behavior is termed *covert* attention and is characterized by a slightly less direct sensing of visual stimuli.

Covert attention benefits from an inhibition of peripheral stimuli. It plays a key role in top-down attention control. The eye movements are profiting from memorization processes that allow an *endogenous* rather than *exogenous* orienting. These voluntary eye movements can be split into two basic categories: *memory-guided* saccades and *antisaccades*. A *memory-guided* saccade involves the retrieval of cued spatial information of objects during targeting them. This may be important for a machine vision system as objects are repeatedly cued by tutors at different locations and the robot needs to remember the object positions. The incorporation of such spatial information during object targeting is carried out by working memory processes that affect the gaze control (Theeuwes et al., 2005). Engelken (1989) analyzed the presence of auditory cues in a visual-search task and found that onset synchrony of audiovisual components can guide eye movements more efficiently by reducing visual workload. This provides evidence from a neuropsychological perspective that initial visual reflexes can be structured by the auditory modality. It may serve as a strategy to develop a gaze control for an artificial agent. However, it is not desirable that a robot can recognize an object solely based on its position, since a tutor can demonstrate an object at various locations. This means that the learning of audiovisual object representations has to incorporate additional factors that are independent of the object location.

More important for an autonomous learning system may be the mechanism of an *antisaccade*. An *antisaccade* describes the voluntary gaze behavior towards objects that are in an opposite direction to a primed stimuli. Thereby, two mechanisms, that are beneficial for a gazing strategy of a robot, are active. On the one hand, the suppression of reactive saccades and hence an inhibition of peripheral visual stimuli plays a major role for conducting such saccades. On the other hand, the visual information system manages the generation of voluntary

saccades to gaze towards an object. Such interaction of visual processes can help a robot to stabilize its gazing behavior.

For an automatic gaze control towards objects, it is important to orchestrate neuronal functions for eye movements and those involved in the process of object recognition. In a fMRI study cortical activities were measured during the execution of reactive and voluntary gaze movements (Mort et al., 2003). Voluntary gaze orienting was associated with an increased activity in the FEF and additionally the *intraparietal sulcus*. Also Reuter et al. (2010) showed that the FEF as well as the nearby *supplementary eye fields* are strongly involved in the circuit for voluntary gaze control. The mentioned studies show that the FEF plays a dominant role in the generation of voluntary saccades. But the FEF also exhibits correlated response activities with respect to areas that are associated with the recognition process. Monosov et al. (2010) investigated this aspect by analyzing brain activities independent from eye movements. He found a correlation of activities in the FEF and the *inferior temporal cortex* that plays a major rule in recognizing visual stimuli. Similar results were reported by Hopfinger et al. (2000), where cortical areas next to the FEF showed response activities independent of eye movements. This emphasizes the involvement of the FEF in recognition and covert attentional processes even in the absence of gazing.

Whereas the circuits underlying visual gaze control have been extensively studied, the neural correlates of audiovisual processing are less clear. The analysis of neural activities in regard to audiovisual object recognition shows that the promptness as well the accuracy in the presence of a multimodal stimuli is improved. However, there are no specific locations in the brain to which an audiovisual object recognition can be assigned. Rather a large activation of the visual cortex was observed in the presence of both modalities (Giard and Peronnet, 1999). Such modulated activities arising from the processing of bisensory signals that were also observed in audiovisual speech perception. Kaiser et al. (2005) investigated audiovisual speech perception by means of the *McGurk Effekt* (McGurk and MacDonald, 1976). This popular effect shows that perceived syllables and lip movements interact and jointly influence speech perception. In general, humans perceive spoken syllables different in the presence of lip movements if they deviate from the presently spoken ones, e.g. we perceive spoken syllables 'ba-ba' as 'da-da' if the physical lip movements specify the syllables 'ga-ga'. The focus of the study of Kaiser et al. (2005) laid on the effect of artificially induced mismatches between physical lip movements and spoken syllables. The results showed that a manipulation of lip movements (e.g. 'ta' to 'pa') results in a delayed activity in the *occipital cortex*. Based on this delayed response, the authors suggested a top-down modulation of early visual processing by higher motion processing areas to compensate for an incongruent representation of audiovisual aspects. Not

only the processing of acoustic features influences the visual information processing, but also visual stimuli can access brain regions that are associated to speech processing. Howard et al. (1996) analyzed response activities of motion processing areas in the presence of deviating motion stimuli. He additionally found an increased activation in areas that are associated for speech processing.

The reviewed approaches give evidence for the processing of audiovisual stimuli with respect to temporal synchronization. In experimental brain research, a further feature constitutes the semantic congruency of an audiovisual object representation. Learning at the semantic level presumes that we have already classified auditory and visual properties of objects and can derive an object percept from it. This aspect is important because it may give indications whether a correlation occurs at a very early or rather late processing level. Such evidences are found by Hein et al. (2007) who studied semantic congruences of audiovisual objects. Their experiment relied on complementary audiovisual object representations that are generated by artificial stimuli or are known to subjects. An example for a known object is an image of a dog and a familiar related sound such as barking. In case of object incongruency known stimuli such as a dog were presented with an atypical sound like the 'meow' of a cat. In summary, subjects showed an increased response activity to unfamiliar and incongruent object representations in brain areas that are associated with learning processes, e.g. the *inferior frontal cortex*. In contrast, cortical areas that are linked to storage processes such as the *posterior superior temporal sulcus* and the *superior temporal gyrus* are activated in the presence of known congruent objects.

Such alignments of audiovisual features require that object representations are properly stored and available on demand for an artificial system. Therefore, the computational storing process of audiovisual objects should be designed in such way that a recognition of them is possible.

2.2. Object Learning and Attention Models

The development of computational attention models for eye movement simulations is based on the mathematical representations of filtering processes. The goal for example is to specify interesting regions in images in order to utilize them for automatic object learning. The advantage of this method relies on a reduction of object features, so that only relevant object characteristics are left for a learning process. An attention model constitutes an ideal candidate for an autonomous learning system, because processing focuses on the essential aspects of the available visual information. Such a model was developed by Itti et al. (1998) and Itti et al. (2003), which relies on the principles of *stimulus-driven*

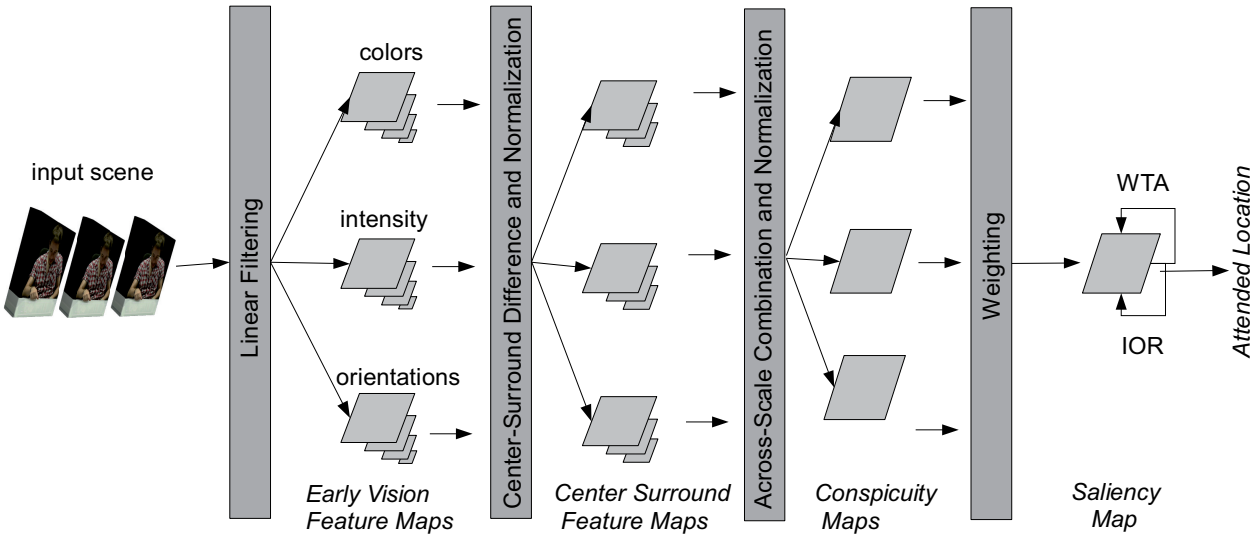

Figure 2.1.: Sketch of a bottom-up attention model (adapted from Itti et al. (1998)). The scene is decomposed in conspicuity maps for color, intensity and orientation. A weighting step linearly combines them into one saliency map whose maximum response location can serve as a new gazing point. An inhibition of return (IOR) mechanism suppresses already selected gazing points.

saccades (see Fig. 2.1). The basic idea is described in Koch and Ullman (1985), who proposes the mechanism of selective attention that is inspired by the *Feature Integration Theory* (Treisman and Gelade, 1980).

The *Feature Integration Theory* relies on the premise that visual attention is directed to those aspects which exhibit similarities across different visual characteristics. Characteristics can hereby for example be the color, intensity, or orientation of objects. From a biological point of view, the idea is supported by the existence of relevant feature detectors in the visual cortex, sensitive to receptive fields in visual field (*retina*). The response behavior of such detectors decomposes the scene in so-called feature maps. This decomposition is related to a preattentive processing step, because the scene is convolved by low-level feature detectors without specifying an object hypothesis. Finally, the combination of these feature maps leads to a common *saliency map*, which highlights particularly salient points that share common visual characteristics. The points represent potential locations where a robot can allocate its attention. To extract the best candidate, a *winner takes all* mechanism is applied to prune the amount of possible candidates to one location.

This approach can be exploited for a rough detection of objects and does not require any top-down guidance, i.e. solely based on low-level visual characteristics. This is a great advantage if there are objects in the scene, which can be specified by such simple features. A recognition of them is only possible if object properties remain unaltered. But in everyday life, we perceive objects in a more complex way with variations in characteristics like color, perspective, or size. These types of changes may serve as learning signal for a robot to recognize different kind of object properties. For example, the rough detection of the object size may be used to specify object-specific features in more detail. A common method for object size specification is the image segmentation that may rely on simple object cues, such as color or shape. Such region specification offers a larger surface than a single saliency point and can serve, e.g. to calculate regional object characteristics and relations between features.

Such an approach is pursued by Walther and Koch (2006). An appropriate image segmentation is initialized by the model of Itti et al. (1998) and computes so called *proto-objects*. These are the result of an automatic segmentation that is implemented by a spreading mechanism. The extracted *proto-objects* are used to control an existing object recognition system based on the learning of object contour orientation combinations. The spreading mechanism is based on a gradual diffusion step of object-specific features and inhibits scene aspects that are beyond the attended location. Such inhibition mechanism can be utilized more efficiently for a modulation of an object recognizer, since a reliable feature extraction is possible. The proposed model by Walther and Koch (2006) features a one-shot learning method that benefits from the inhibition of areas that lie outside of the object. But this model lacks a proper classification, i.e. it is unknown to the system when a new object is available as well as when to extend the knowledge base with new object knowledge. Such incremental learning of objects is also proposed by Walther et al. (2005). The model is based on a similar image segmentation as the above model with the difference that distinguished *Sift-Features* (Lowe, 1999) are obtained from the computed salient regions. The model matches existing object representations to the internal object state and learns a possible new one.

The principle of inhibition in the modeling of attention systems is not only used to improve feature selection, but also to model object representations. Frintrop et al. (2005) (see Fig. 2.2) proposed a method for a goal-directed search that utilizes learned objects to simulate top-down saccades. A modulation of a visual bottom-up search is realized in the approach by the varying influence of learned top-down information. Here, the object is represented by excitatory and inhibitory information that highlights salient object parts. The weighting of such learned top-down information shows that the mechanism may be used by an artificial system to

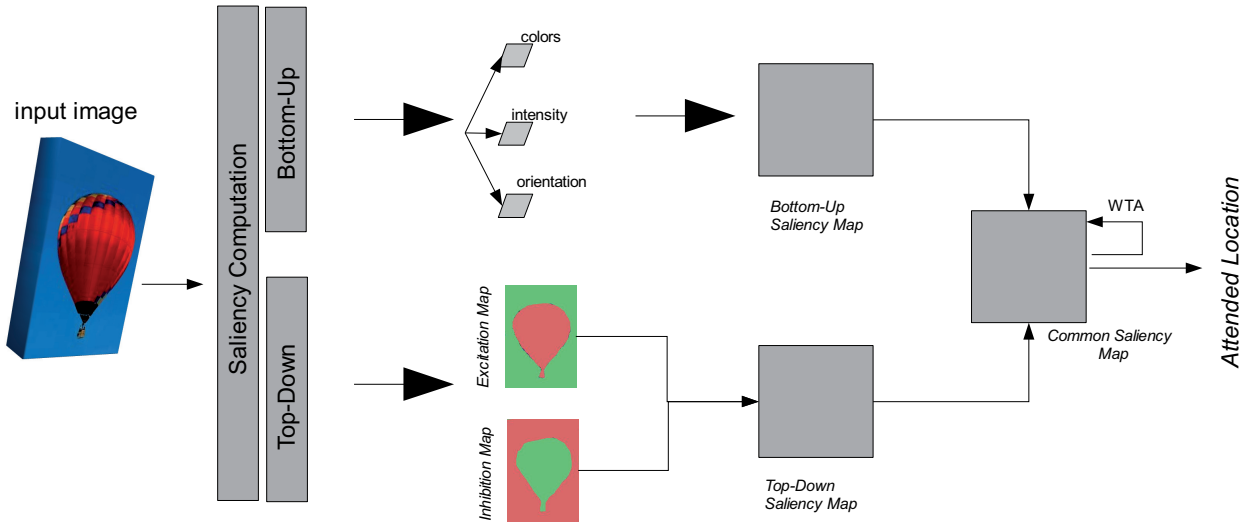

Figure 2.2.: Sketch of a top-down modulated computational attention system (adapted from Frintrop et al. (2005)). The architecture learns top-down knowledge by a decomposition of objects via inhibitory and excitatory information.

suppress reactive eye movements and learn to direct attention to acquired visual information. However, the main difference to the above mentioned approaches above relies in the triggering of learning top-down knowledge. A bottom-up attention model is used to segment only visual information in a specified region of interest and does not define the position of relevant scene aspects. In detail, the determination of potential object candidates is either manually predefined or supplied through the output of an object recognition algorithm. A continuous free exploration of the system does not occur and hence an unsupervised learning of objects is not supported. Additionally, the acquisition of object representations is implemented with an offline learning process. This limits the system with respect to knowledge acquisition, since the system can only focus on previously learned aspects. More precisely, new aspects are ignored since an additional learning of them does not occur. Consequently, the system does not implement incremental learning of top-down knowledge.

The reviewed approaches of Walther et al. (2005) and Walther and Koch (2006) showed that the mechanism of inhibition may improve the selection of object relevant features. In addition, bottom-up attention models can be used to determine relevant regions for object learning. The extracted regions can specify the object itself or describe background information. The approach of Frintrop et al. (2005) combines both information sources on the basis of low-level features that are given

by bottom-up constraints, e.g. color, intensity and orientation. But this way of object modeling is not invariant to object changes and therefore not sufficient for an object learning scenario in which objects can be varied freely in their appearance. Contrary, Walther et al. (2005) extracted a set of *Sift-Features* from regions that are limited by the object region itself as a representative. This step allows an object description that is invariant to rotation and therefore more robust. Hence, it would be useful to find a compromise between these two approaches that allow a robust object representation, e.g. by a combination of *Sift-Features* and the mechanisms of inhibition. However, a foreground and background segmentation as a basis for the calculation of more complex features like *Sift-Features* would simply lead to memory problems. For this, a method may be useful that extracts only those background characteristics that are relevant for an object description in order to limit the number of features.

None of the presented approaches deals with an adaptation to varying object attributes, such as the perspective. A robot should have this ability to automatically expand its knowledge base in order to recognize objects from different views. In part, the approaches show a learning of top-down knowledge on the basis of visual modalities that can be used to control robots' cameras, but not by additional modalities like auditory characteristics.

Learning multimodal object representations presupposes that visual and auditory information are acquired and somehow learned in combination with each other. These aspects are not covered in the above mentioned models. It is shown that bottom-up visual information is beneficial to trigger the learning of top-down visual information. However, without a combination of vision and audition an autonomous system cannot take advantage of acoustic scene features to control the visual attention to objects. In robotics, the concept of multimodality not only plays an important role in the development of learning algorithms, but is also being studied in interaction studies. This aspect will be highlighted in the next section.

2.3. The Role of Multimodal Attention in Robotics

Multimodal attention systems in robotics are developed and used to enable a flexible interaction with humans. However, numerous parameters vary in a human-robot interaction and need to be handled by an autonomous agent. For instance, a free interaction may be characterized by spontaneously shown objects whose functionalities are arbitrarily demonstrated using other modalities such as speech. In addition to the presence of speech and other acoustic properties, hand movements or body gestures can guide the interaction. From an engineering point

of view, this results in a wealth of stimuli with which autonomous learning system needs to deal to finally extract meaningful information for an object representation. This is challenging in many ways, particularly with respect to the engineering of learning and recognition processes of audiovisual objects, since it cannot be determined in advance which objects are shown or at which positions they appear. Such a position cue can be varied freely by a tutor, so it is desirable that a robot has the ability to recognize objects invariant of their locations. In addition, a free interaction between a human and a robot not only comprises the demonstration of objects but also spontaneously appearing scene aspects that do not belong to a learned object. Such spontaneous aspects need to capture smoothly by an artificial agent that may be handled with a suppression mechanism for such irrelevant scene attributes. One approach to the learning of such multimodal object representations can be found in the work of Aryananda (2006). The proposed top-down audiovisual attention system enables the robot to gather spatio-temporal patterns that result from a joint concurrency of audiovisual events. The suggested architecture is tested in an interaction study with humans in order to extract the data stream in visual and auditory segments that have significant coincided over time. This segmentation can provide an important basis for the learning of multimodal objects. In detail, the patterns include information about whether human faces or toys are shown in coincidence with acoustic signals such as speech. Crucial to the design of this attention model is the use of a predefined face recognizer and color histograms for the detection of toys. Additionally, the classification of acoustic properties is defined by a word recognizer that is preceded by an energy-based voice activity detector. The coincidence of both modalities is determined by a correlation calculation with the location cue. More precisely, sound localization in azimuth is integrated on a sensory ego-sphere with the visual stimulus location.

The modeling of attention systems can benefit from online learning due to storing capacity and computational costs. Typically, the system starts with very little knowledge about existing objects and builds up object knowledge step by step. This has the advantage that the system learns to organize perceptual information in an unsupervised way and acquires new knowledge incrementally if necessary. The application of off-line trained classification techniques such as in the above mentioned model has the disadvantage that the system is not able to modify its own object representations which hence leads to an inflexible gazing behavior. A computational approach not based on pre-trained classifiers is examined by Ruesch (2008). This attention model is based on the integration of bottom-up attention models for the auditory and visual modality and can be used to simulate *stimulus-driven* eye movements to audiovisual aspects. In detail, the multimodal integration process is based on the location on an ego-sphere, but in

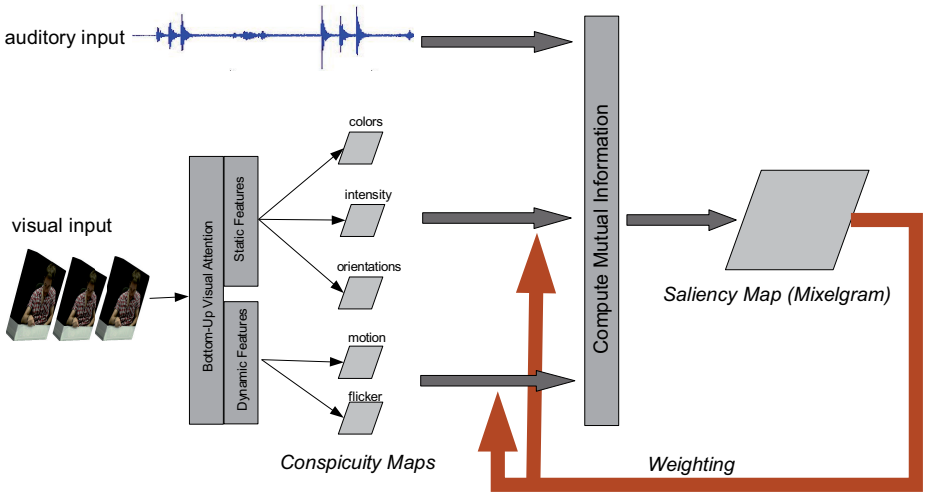

Figure 2.3.: Computation of synchrony based on *Mutual Information* of low level features (static features and dynamic features). The model embeds auditory features (e.g. the intensity) into a bottom-up visual attention model. The calculated *Mutual Information* serves to reinforce or to alleviate the weighting (red arrows) of the respective conspicuity maps. Adapted from Rolf et al. (2009).

contrast to the above approach they localize the sound in azimuth and elevation. Additionally, the architecture is equipped with a habituation model, which is gradually updated with multimodal information. A memorization process such as habituation ensures that multimodal aspects can be focused for a defined time period. After exceeding a habituation value new points can be targeted. To avoid refocusing the same points, the gazing strategy is regulated by a so-called time-decay factor, which inhibits previously attended locations.

Another model, which pursues the modeling of reactive saccades, is the approach of Rolf et al. (2009)(see Fig. 2.3). Unlike Ruesch (2008), the model is based on the study of synchrony between features such as the intensity of the auditory and visual modality. The model is inspired by the *Intersensory Redundancy Hypothesis* (Bahrick and Lickliter, 2000), that relies on the assumption that infants increase their attention to redundant perceptual information, e.g. to a knocking hand. These redundant overlaps in both modalities can help infants to structure their perceptual environment and to acquire knowledge about constructs like speech or objects. For this purpose, the presented model is evaluated in two tutor scenarios: One scenario comprises video data, which shows a parent-parent interaction. The other scenario is a child-parent interaction. The attention model computes synchronous portions of both modalities of the scene from the perspective of the learner and shows an increasing response activity to the infant-

directed learning scenario. The correlation computation is based on the *Mutual Information* (Hershey and Movellan, 1999), which is integrated with a constant weighting term over time to suppress random correlations. The result is a map (*mixelgram*) showing the locations in the camera image in which both modalities particularly coincide with each other. This temporal integration is useful in a robot learning scenario, since gazing behavior may be modeled on aspects that are bright and loud. A proper classification of multimodal objects is missing, since complex objects are not acquired by the model. Instead, the extension of an existing bottom-up visual attention model by Itti et al. (2003) is proposed, which calculates a weighting of conspicuity maps using the proposed correlation computation. This means that in the event of temporary high synchrony of both modalities, dynamic features such as optical flow or movement are weighted stronger in the coincidence computation. The approach shows that the correlation calculation is dominated mainly by features such as the movement of objects and less by object-specific correlations with auditory features. The system learns specific positions of the scene where audiovisual correlation is present. However, the actual classification of audiovisual objects is not integrated and therefore less usable for a voluntary gazing towards multimodal aspects.

2.4. Summary

The study of voluntary eye movements in neurophysiology shows that the mechanism of stimuli suppression in our environment plays an important role. This may be important in robotics in two ways. On the one hand, such a mechanism may enable the robot to track an object continuously since distractor stimuli may be suppressed and thus leading to a stable gazing behavior. On the other hand, this mechanism may be utilized to direct saccades to an object without generating random eye movement. The mechanism of inhibition plays an important role for the generation of endogenous saccades that may biased by already learned visual knowledge. For the development of voluntary gazing behavior this means that an artificial agent needs to acquire visual scene knowledge in an unsupervised way and expand it on demand. Furthermore, studies have shown that the visual information processing is biased not only by visual but also by auditory top-down knowledge, e.g. in object recognition tasks. In particular, the storage and recognition performance of objects is decisively influenced by the temporal synchrony of both modalities. This shows clear indications of how the visual information processing is supported by the auditory modality, and ultimately affects the recognition of objects. This modulation of visual attention control can play a key role in the development of an intelligent gazing behavior for a robot. This principle has not previously been used in robotics, so that there is a lack

of learning processes which simulate voluntary eye movements on multimodal aspects.

The approaches given by Aryananda (2006), Ruesch (2008) and Rolf et al. (2009) aim at modeling multimodal attention systems for artificial agents. The fusion of both modalities is based on the respective position of the observation by the robot, i.e. local audiovisual features are correlated in the visual field of the robot. Such 'location-based' approaches can serve for calculating new fixation points and may equip the robot with an according reactive gazing behavior, but are not sufficient to model voluntary eye movements on multimodal aspects. This can be inferred by the aspect that the re-identification process of learned multimodal knowledge is not modeled, since the approaches lack a causal relationship between auditory and visual object representations. In detail, the modeling of top-down aspects such as speech and its influence on visual attention control is not investigated so far and therefore a specific gaze control for targeting objects using acoustic features remains. However, such gazing behavior is desirable for a robot and challenges the automatic acquisition of scene knowledge on demand. Therefore, it is useful to take developmental aspects into consideration and to investigate the gazing behavior of infants. The genesis of this gazing behavior is rapidly learned in infancy and may serve as an inspiration for the development of an artificial gazing strategy. This aspect is addressed in the next Chapter and serves as motivation for the development of an Active Vision Architecture.

3. A Developmentally Inspired Active Vision Architecture

Infancy research yields insights in how children gradually improve their cognitive skills over the years. The analysis of such a development allows us to understand how we learn to comprehend our world and may finally serve as an inspiration for the development of machine learning algorithms for object recognition. In infancy research, especially gazing behavior is examined to derive principles of learning mechanisms. This Chapter gives a summary of the main methods to analyze infants' object learning competence. In particular, the influence of acoustics on object learning is picked out and discussed with respect to potential benefits for artificial agents. Inspired by findings on infants' gazing, the last section presents an Active Vision Architecture that generates voluntary saccades to multimodal aspects of the environment.

3.1. Object Learning in Infancy Research

Infancy research that deals with visual object learning can be viewed from two perspectives. One research direction focuses on parental behavior towards children while they demonstrate objects to them. The other view examines the gazing behavior of infants during object learning in tutoring scenarios. Typically, a parent-infant interaction is characterized by parents altering gestures and speech when they explain something to their children. Especially, in object learning scenarios it can be observed that parents synchronize their voice and object movements in time to gain the attention of their children. Such an alignment of both modalities is called *multimodal motherese* and was examined by Gogate et al. (2000). The study focused on different groups of children aged 5-30 months and revealed that the synchronization of object movements and speech in caregiver tutoring is strongest when interacting with the youngest age group. Such a modification of parental behavior is not only found in the coupling of motion and speech, but also in sign language communication with deaf children (Masataka, 1992).

From a technical system perspective, it is of interest to detect such redundant source of information to model them in a machine vision system. Therefore, it is

important to more deeply explore how young children develop their visual skills from such multimodal aspects. For this purpose, a review is given of different factors that affect the looking behavior of infants and which aspects may be beneficial for machine learning. Experimental studies on infant gazing behaviors are particularly important at preverbal developmental stages to provide qualitative measures on infant object representations.

In principle, infants' object competencies such as learning or recognition are carried out by the analysis of two capabilities. One capability is defined in terms of discrimination which means whether infants can distinguish between learned objects and others. The other capability refers to the development of visual preference in terms of visual recovery, i.e. whether infants recognize objects they are familiar to. These capabilities are accessed by two basic examination methods: the *Preferential Looking Paradigm* and the *Habituation Paradigm*. Both methods share a learning phase in which infants are familiarized to objects and a testing phase in which infant gazing behavior is analyzed to measure whether objects were learned or not.

3.1.1. Habituation Paradigm

The *Habituation Paradigm* grants access to infants object competencies by measuring whether they can discriminate between new objects and already learned objects. Figure 3.1 shows a framework for an experimental setup for object learning in two conditions with respect to this paradigm. The upper row depicts an example of a unimodal condition (visual only) whereas the bottom row shows one for a bimodal condition (audiovisual). The left and right columns describe the learning and testing phase, respectively. Firstly, in a so-called *Habituation-Phase* (Sirois and Mareschal, 2002; Roder et al., 2000), an object is presented repeatedly to infants and their looking duration to the object is recorded (dashed line). For example, the visual condition may comprise an image of a *flower*. In the bimodal condition, it may be a *hammer* that is striked with a certain *frequency* on the table. Here, the *hammer* serves as visual referent that is learned in association with a specific *rhythm*. The stimuli are presented until the infant shifts its attention to novel aspects in the scene (solid arrow). This attention shift during habituation is assumed to signal that the child gathered sufficient information on. Afterward, a testing phase follows.

In the testing phase, infants' discrimination capabilities are analyzed by means of looking duration. A novel stimulus is added and the gaze duration is recorded again. In the unimodal condition this can be archieved by adding the visual stimulus of a *cloud*. In the bimodal condition, it may be the change of the

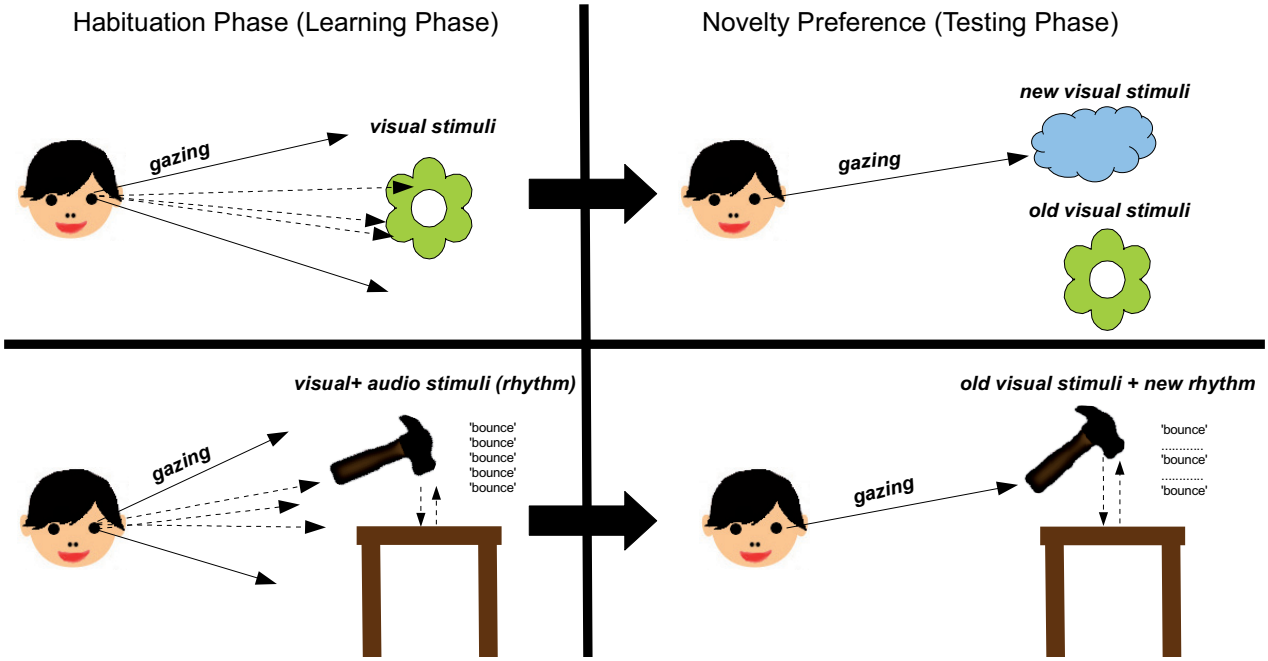

Figure 3.1.: Sketch of the *Habituation-Paradigm* in Infancy Research (from left to right). A habituation-phase is analogous to a learning phase, where decreasing interest results in a shorter gaze duration to the object (see dashed arrows). In a testing phase, a novel object is additionally presented and the infant shows a preference for this object. A new object can comprise a visual object representation or changes in the auditory characteristic like the rhythm that is aligned with a moving object. The illustration of the bimodal condition is adapted from Bahrick and Lickliter (2000).

rhythm with that the *hammer* strikes on the table. The gaze duration to novel objects is examined. The presence of a longer gaze duration suggests that infants can distinguish new introduced stimuli from the learned one. For the unimodal condition that would result in longer gaze fixations to the cloud. In the bimodal condition, a longer gaze duration to the hammer is interpreted by a detection of the changed rhythm. In both cases, this kind of gazing behavior relies on the assumption that infants have learned and memorized objects from the habituation trial and subsequently demonstrate more interest to novel objects by increasing their visual attention to them.

3.1.2. Preferential Looking Paradigm

The *Preferential Looking Paradigm* (PLP) (Golinkoff et al., 1987) provides an analysis of both, the discrimination and the recognition capabilities of infants with respect to objects. In contrast to the *Habituation-Paradigm*, object competencies are accessed by an increasing gaze duration on learned objects, i.e. not

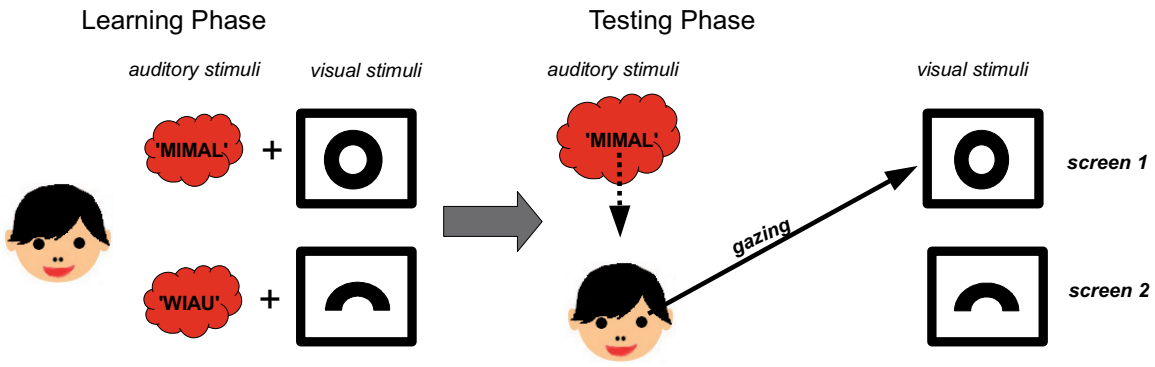

Figure 3.2.: Sketch of the *Preferential Looking Paradigm* in Infancy Research (from left to right). Word labels are complemented with novel objects during a learning phase. One perceptual cue can subsequently guide the visual attention to objects (Allport, 1989).

by novelty preference. Additionally, the PLP is mostly utilized to examine object learning in the presence of other modalities. Other modalities may include word labels or object sounds. One possible procedure is shown in Figure 3.2, where the left hand side depicts the learning phase and the right hand side shows the testing phase. An application of this paradigm may comprise the goal to get insights about infants' word learning capabilities by examining their gazing behavior. For this purpose, novel visual stimuli are presented with various novel word labels in a learning phase. Such complementary representations may comprise relations between objects and word-labels like *circle-'mimal'* and *semi-circle-*'wiau' as shown in the example.

The subsequent testing phase involves the presentation of previously shown stimuli from the learning phase. In this example, both shown visual stimuli (*circle*, *semi-circle*) and one stimulus from the other modality (*'mimal'*). The gaze duration to familiar visual stimuli is examined during the presence of an associative learned word label. If infants show longer looking times to the object *circle*, it is assumed that the infants learned the new label *'mimal'*. This gazing behavior following the label *'mimal'* is based on an associative learning step with the visual referent (*circle*) and an additional discrimination step of the label *'wiau'*. Such preferential looking at familiar objects is interpreted by the guidance of complementary learned modalities that may lead to an increase of infants' visual attention in terms of object recognition.

The introduced methods show that the examination of infants' object competencies can be divided into two models. It is noteworthy that the shown examples just represent possible forms of accessing object competencies and further reduce the experimental scenario to its essentials. In principle, the first model relies on the assumption that an object is learned when children are capable of distin-

guishing it from another object. Such a discrimination ability with respect to audiovisual components is desirable and is particularly important for an automatic learning of objects. This in turn stresses the relation to the significance of an appropriate suppression mechanisms of the incoming sensory stimuli (see Section 2.1). In addition, the model shows that infants are very sensitive to novel aspects. This may demonstrate a way for an autonomous system to act with new object knowledge to specify them for an internal memorization.

The second method, however, investigates object learning from a different point of view in which the learner develops a kind of visual preference. An object is considered as learned when correct visual referents are focused. The model attempts to access infants competencies by recovering familiar visual object knowledge during listening to learned auditory qualities such as labels or sound. Such an associative learning of multiple modalities and the corresponding triggering of the visual information system may inspire the development of an artificial gazing strategy. This mechanism may enable a robot to configure an internal filtering process that benefits from acoustic qualities to move its eyes to objects.

3.2. State of the Art

It is of interest to get further insights into infants' gazing especially in the presence of auditory characteristics and how such characteristics may serve to structure eye movements of a robot. The shown methods are applied in different ways and divide infants' visual competencies into two gazing strategies that may be important for artificial saccade generation in robotics. Therefore, the following section focuses on studies depicted from infancy research with focus on the integration of multiple modalities and their effect on gazing behavior.

3.2.1. Acoustic Cues and Infants' Gazing Behavior

Here, it may be important how infants differentiate between acoustic properties and when they start to associate them with objects. This aspect appears relevant for an artificial vision system to implement a possible association mechanism between different modalities. Such mechanisms have been studied by Cummings et al. (2009), where her experimental work focused on the visual recovery performance of infants. In the study, shown objects were provided either by a verbal label or with a meaningful sound. The set of stimuli comprised the image of a dog and the word label *'a barking dog'* or the sound of the dog barking like *'woof'*. The experiment demonstrated that 15-25 month old infants develop a visual preference in the presence of object specific sounds. On the contrary, the presence of

verbal labels leads also to an increase of visual attention to objects but only for the older subject group. It has been concluded by Cummings et al. (2009) that these participants already gained verbal proficiency in their life. Additionally, it has been shown that infants form object categories in object-label tasks (Balaban and Waxman, 1997; Fulkerson and Waxman, 2007) and treat words differently to tones (Ferry et al., 2010).

A similar result is also obtained in a study of Best et al. (2010). Outstanding in this study is that many different types of visual objects and verbal labels are presented to infants from 16 to 24 month. The results showed that infants are able to focus on objects using a variety of verbal labels and use them to structure their visual environment. In comparison to this, shown objects without labels resulted in a lack of preferential looking. Such stimuli variance is also characteristic in learning situations with robots. A variety of acoustic properties may depend on different object properties and may further vary depending on the tutor. It thus challenges the way in which a robot can deal with them. Such gazing behavior by means of acoustic cues can help a machine vision system to form unified object representations between modalities like the assignment of speech or even tapping a particular visual object.

The presence of auditory properties may not only help a robot to structure visual properties, but also to improve the ability to generalize objects. This characteristic has been studied experimentally by Plunkett et al. (2008). The study showed that infants not only try to associate similarities between modalities, but utilize labels to merge dissimilar appearing objects into same categories. Such a fusion process determined by several modalities can support a machine vision system in considering objects being similar even though their appearance differs.

All studies considered the acquisition of objects to be an associative learning process that relies on static object representations. However, in learning scenarios with robots objects are additionally moved, i.e. they are not always presented at the same location by a tutor. Consequently, it is interesting to see whether infants are able to learn objects by relying on dynamic features such as motion and accompanied sound. More precisely, whether infants gazing behavior continuously relies on associations between dynamic features and rather than on the location resulting from acoustic sources. Such a mechanism may enable gazing regardless of sound sources, but prerequisites that the robot is equipped with a sustaining mechanism that allows him to focus on moving multimodal objects. The ability to locate multimodal objects irrespective of the sound location (or a predefined location) is demonstrated by already 6 months old children (Richardson and Kirkham, 2004), called *Dynamic Spatial Indexing*. For this purpose, the gazing behavior of infants was examined in two studies that are now explained

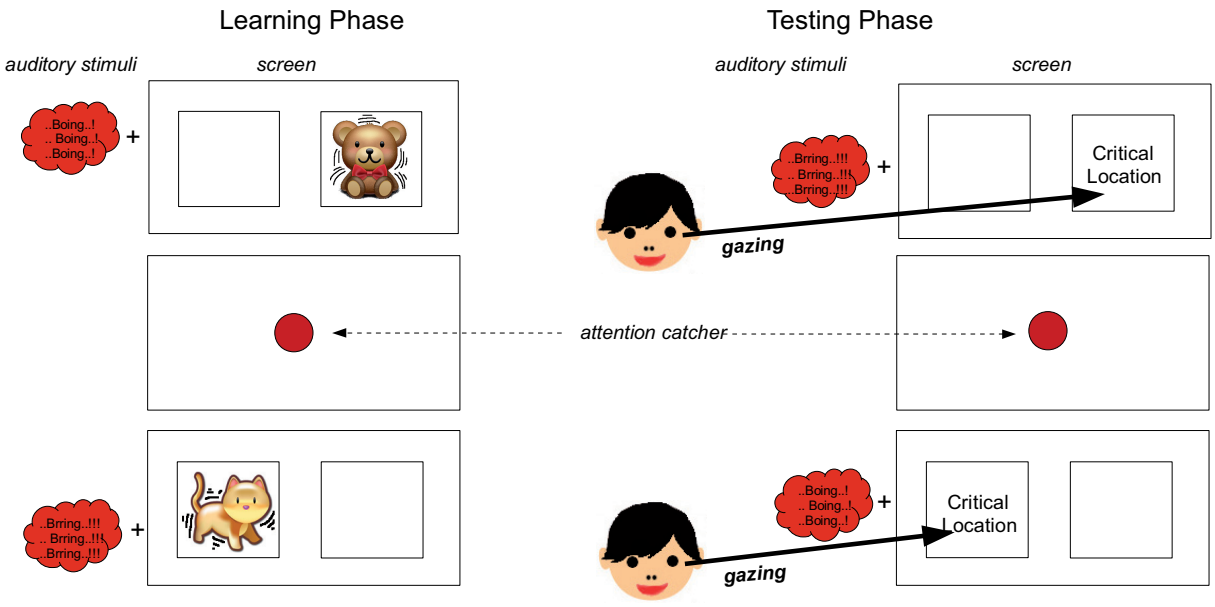

Figure 3.3.: Sketch of *Spatial Indexing* capabilities of infants (from left to right). Sounds are complemented with bouncing objects during a learning phase. Associated sounds can serve for guiding attention to an object location. The sound always appears from the center of the screen. Adapted from Richardson and Kirkham (2004).

in detail. The insights from these experiments can serve as an inspiration for designing a gaze control strategy for robots to attend to multimodal events.

The first study considered an associative learning between object movements and a synchronously played sound to analyze infants gazing strategy irrespective of sound source locations. The procedure of the experiment is depicted in Figure 3.3. In a learning phase, infants are confronted with bouncing toys that are presented in boxes on the right or left side of a screen. Synchronous with the bouncing toys, infants are listening to sounds. These always appear from the center of the screen to test infants gazing to multimodal aspects invariant from the sound location. The stimuli set is presented multiple times. In between, an attention catcher appears to relocate infants attention to the screen center again.

In a testing phase, the infant is looking at a screen that comprises two empty boxes in form of rectangles and is listening to sounds of the learning phase. One of these boxes mark the object position of the learning phase, a so-called *critical location*. The final result showed that infants increase their visual attention to such *critical locations* using sounds that have been associated with them. This gives clear indications that infants are able to generate voluntary saccades to multimodal aspects regardless of sound locations.

This gazing mechanism constitutes a substantial benefit for a machine vision

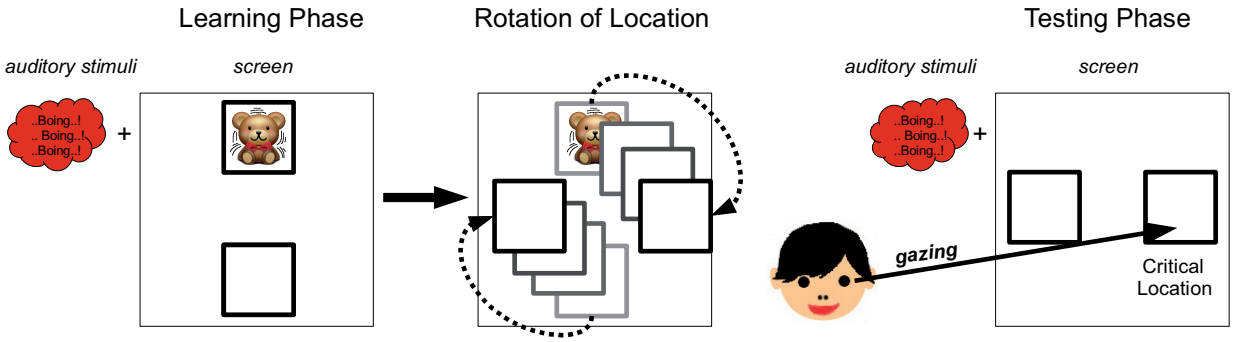

Figure 3.4.: Sketch of *Dynamic Spatial Indexing* capabilities of infants (from left to right). Sound are accompanied by bouncing objects during a learning phase. Subsequently, the object location is rotated. Adapted from Richardson and Kirkham (2004).

system, since the generation of such saccades may enable a voluntary gazing on multimodal aspects. More precisely, the demonstrated gazing behavior relies less on the position cue of dominant features, e.g. sound but rather on learned associations. This fact has been interpreted by Richardson and Kirkham (2004) in terms of forming of visual expectations that are coupled to the presence of a sound. Equipping a robot with such an expectation mechanism may allow a classification of various objects by acoustic properties and further may involve a weighting of learned visual knowledge. However, it is important that a robot not only moves the eyes on learned object positions as it was demonstrated in the previous experiment. Objects may change their locations and this implies that the robot needs to allocate its attention in a more flexible manner which involves a kind of position invariance.

Exactly this aspect was studied in the second experiment where a rotation of multimodal object locations was additionally taken into account. The experiment was designed to investigate whether infants own the ability of *Dynamic Spatial Indexing* (DSI), i.e. use sounds to locate objects, even though the positions of them changed. Such DSI capabilities were tested in an experiment that is schematically illustrated in Figure 3.4. Unlike the first experiment (see Fig. 3.3) an animation is inserted into the learning and testing phase which comprises a rotation of the object positions in terms of the object boxes. After the animation, infants were confronted once again with a screen that showed two empty boxes. Again, a sound is presented from the learning phase and infants' gazing behavior is analyzed with respect to visual preference to the boxes. Overall, the result showed that infants look more frequently at the *critical location* in the presence of an associated sound.

The experiment demonstrated that the learning of multimodal objects may be

implemented by an associative mechanism between different modalities. The learned association can be used to equip a robot with an expectation mechanism that operates with different acoustic qualities and activates corresponding visual object knowledge. Such a development of visual preference may constitute a basis for a robot to obtain a gazing behavior to multimodal aspects.

Overall, the studies showed that infants learn voluntary gazing by relying less on bottom-up stimuli. The execution of such voluntary gazing presumes visual learning processes to gather of top-down knowledge such as object-sound. But also other top-down factors can control the visual attention of infants during the acquisition of object competencies. These may be additional features that are offered by tutors during the learning process such as gazing or pointing gestures to objects. Additionally, the analysis on the interaction of bottom-up and top-down processes may shed light on how much prior knowledge should be defined for an automatic acquisition of audiovisual objects.

Pruden et al. (2006) analyzed this aspect, in particular when infants start to benefit from such top-down cues and when they rely on simple bottom-up stimuli in object learning. Her study examined whether infants learn to associate words with objects. During the learning session, the tutor explicitly gazed to the object. Special to the study was that infants have to assign labels to salient objects or boring objects. Here, salient objects possess conspicuous colors or a sparkle appearance and boring objects were tools with a rather modest appearance. The result showed that infants at the age of 10-month form label-object associations. They develop a preferential looking only in the presence of salient objects. This demonstrates that learning from audiovisual associations is initially driven by visual salience or driven by simple visual features. This may bootstrap object learning for a robot. Additionally, the study revealed that the assignment of words to less salient objects appears later in infancy. Such a late integration of top-down knowledge demonstrates a way to model voluntary saccades that initially rely on bottom-up process such as visual saliency. Later on, top-down knowledge may bias the gaze control which may allow a more voluntary gazing towards objects regardless of their salient visual appearance.

The above mentioned study used word labels and static objects and relied on modal specific aspects of objects. However, it may occur that a tutor alters her speaking behavior and presents objects with different motion amplitudes to the robot. Therefore, a consistent object description and recognition is not guaranteed by modeling an object via individual feature characteristics since they may strongly vary in both modalities. For this purpose, it may be useful to extract features that abstract from their individual characteristics and rather comprise properties shared by both modalities.

This kind of features are called *amodal* and are intensively studied in infancy

research. An example for an *amodal* object property includes the rhythm (see Fig. 3.2 and Fig. 3.3) that appears in both modalities. An object can be moved with a certain rhythm and accordingly generates rhythmic sound patterns. Infants are able to detect such synchronously occurring events and may use them as a learning signal (Bahrick and Lickliter, 2000) that enables them to succeed in difficult discrimination tasks (Bahrick et al., 2010). Such a synchronous presentation could help a robot to extract relevant learning signals, to build associations between modalities, and provide a way to combine them. But an associative learning step that is solely based on rhythm would fail, since rhythm exhibits various tempi characteristics, e.g. hands can sometimes beat faster or slower on a table. This may result in several hand-sound pairings. More precisely, visual objects of the same instance may be associated with different rhythms. However, it is desirable that a robot can generalize such appearances to shift the attention to objects. Such generalization capabilities were studied by Farzin et al. (2009). He showed that infants not only rely on *amodal* properties for learning but are able to detect multimodal events by their equivalence in numerosity. For this purpose, infants aged between 6 and 9 months were confronted with jumping toys and a certain number of tones that were synchronously played in a fix interval. In a learning phase, they are invited to align the number of tones with the number of object jumps. Next, the same toys were presented again without movements as well as the same number of tones differing in their rhythm to examine infants abilities to abstract *amodal* features. The result showed that infants can build associations by using the numerosity of perceptual events and increase their visual attention to familiar objects irrespective of *amodal* information.

The detection of redundant information through algorithmic processes requires the specification of the term *synchronicity*. In particular, it is important to define which characteristic in the multimodal signals could allow to measure synchronicity. Additionally, it is not only decisive to specify appropriate features, but also whether synchronicity is composed of a continuous interplay between different modalities or fits together more likely from the cooperation of single events. What exactly is the concept of *synchronicity* and what is its meaning in infancy research?

In infancy research, *synchronicity* is described as a coherent appearance of features such as object movements or intensity of voice signals. Gogate et al. (2009) studied this aspect on the basis of two months old infants and showed that they are capable of using synchronous onset-offset transitions to discriminate between different syllable-object pairings. Such transitions are extracted from object movements and the energy of the audio signal and relates to infants novelty preference. Implementing onsets and offsets may be a major progress for an automated processing, as this demonstrates a robust feature that abstracts from

the amplitude of the signal. Another important issue which has been examined refers to the compensation of time delays between modalities. Such delayed signals are perceived synchronous in a certain range of time shifts. The range in which such shifts are perceived as synchronous is also known as *intersensory temporal contiguity window* (Lewkowicz, 2000). In general, infants can compensate time shifts up to 350 ms, i.e. this interval can help to compensate inaccuracies from computational processes such as automatic onset/offset calculations.

3.2.2. Learning during Infancy Sleep

The previous methods assumed that children are in an awake state while performing object learning. Of course, this is necessary to analyze the behavior of eye movements during the acquisition of object knowledge. But what happens during the sleeping time in infancy and what benefits can be inferred for an unsupervised learning of objects? In analogy to infants, a robot may operate in an 'awake' or 'sleeping' state. The awake state could correspond to an online phase in which knowledge is acquired by a visual exploration of the scene. In contrast during a sleeping state a robot could perform offline processes in which internal representations are manipulated to enable a deeper processing of acquired knowledge. Hupbach et al. (2009) showed that infants develop enhanced generalization capabilities with respect to predicted patterns in language, when they have slept after a learning phase. It is assumed that infants consolidate learned information during sleep and show an enhanced retrieval performance in long-term memory. Such a consolidation process may be important for an autonomous system with respect to two aspects.

Firstly, this process may allow to organize the object memory of the robot so that it can most efficiently be retrieved. One possibility would be to equip the robot with a mechanism that merges similar objects as one percept and restructures existing internal object representations. The second aspect concerns the accumulation of learned object representations by the system. In children, it was observed that they not only consolidate memory contents during active sleep (Tarullo et al., 2011), but the sleep also plays an important role in the formation of synaptic connections in the brain. Here, the pruning of synapses is of particular interest for developing machine learning algorithms. Such a pruning may be useful to prevent an acquisition of redundant object information and to constraint the growth of the memory.

3.3. A Gaze Control Strategy towards Multimodal Events

In the following an active vision architecture is described, which is inspired by developmental aspects of infants' visual information processing system. The aim of the architecture includes the learning of object representations, so that a robot is able to focus multimodal aspects in its environment. The architecture learns objects in an unsupervised way, thereby integrating information from different sensory modalities. These multimodal associations can subsequently be used for object recognition. In detail, the recognition phase comprises the configuration of acquired object knowledge based on associations that are extracted during a learning phase.

Figure 3.5 shows a scheme of the architecture, where functional components are illustrated by different colors. In principle, the components are classified according to four colors: pink, blue, green and light blue. The pink colored components depict relevant features of both modalities and the correlation calculation between them. Each of the blue components shows the processing of incoming signals in terms of learning models and their application. The light blue portion describes those components that are involved in the overt control of the camera eyes, i.e. where to look next and what points are interesting to track. The green-colored component describes a weighting function that allows the robot to increase the visual attention on multimodal aspects during the processing of learned auditory knowledge.

Overall, the architecture (Grahl et al., 2011) describes a mechanism for the control of saccades, which automatically increases the visual attention to multimodal aspects. For the development of such a control mechanism, it is useful to refer to the principle of generating voluntary saccades by infants such as shown in the experiments of Richardson and Kirkham (2004). More precisely, the integration of top-down knowledge can help an artificial system to focus those objects, whose visual properties are correlated with auditory cues. To give a detailed overview of the architecture, in the following individual components are described with respect to their embedding in the process of attention control.

Top-Down Filter Weighting

In the initial learning phase, the gaze selection of the system is reactive and is defined by *stimulus-driven saccades*. Such saccades are simulated following the model of Itti et al. (1998, 2003) and serve as bootstrapping mechanisms for object learning since they initiate the gazing of some aspects in the world, which

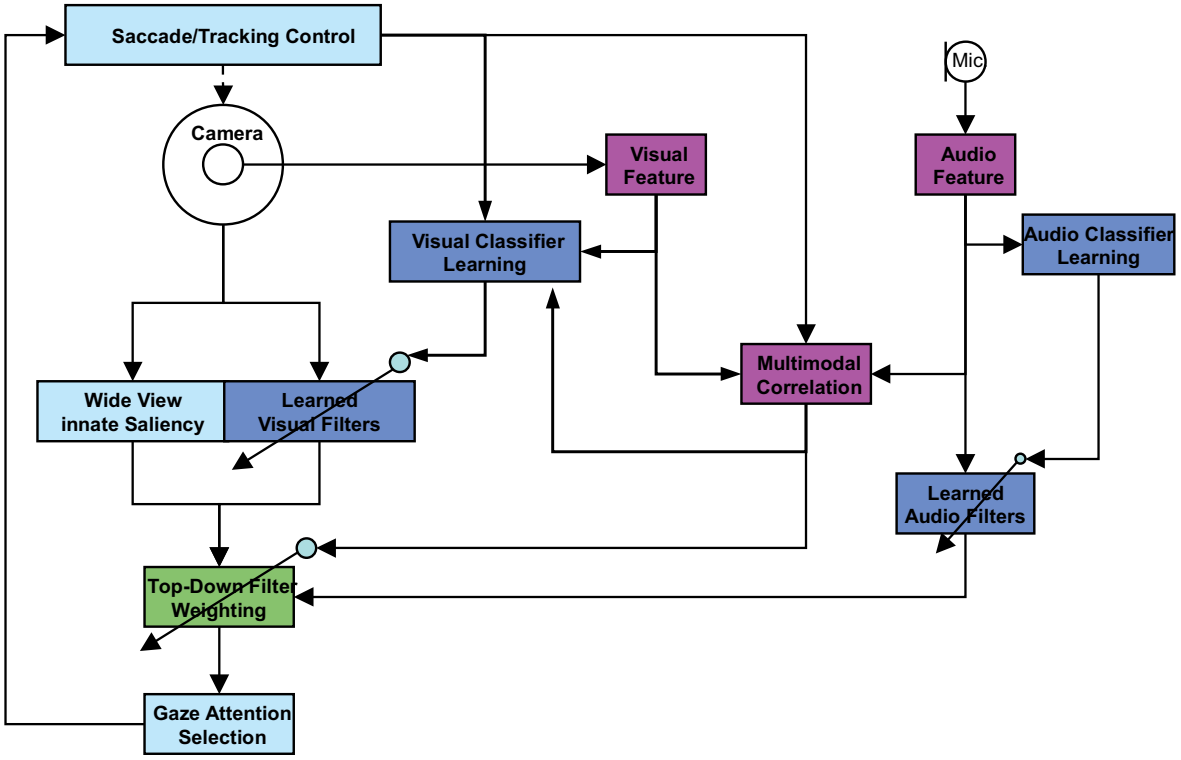

Figure 3.5.: An active vision system for focusing on relevant multimodal aspects.

can serve (when tracked) as a basis to learn visual classifiers for such aspects. In order to generate saccades towards aspects of the world that are correlated to some perceived auditory characteristics, the system first builds new saliency maps whose activity is the result of the filtering of the camera scene by learned visual filters. This mechanism is in accordance to the learning of object preference by means of auditory characteristics and inspired by Richardson and Kirkham (2004). When a sound aspect is perceived that is classified as a previously experienced sound category, the learned saliency maps are modulated by a set of weights that are proportional to the multimodal correlation experienced between the corresponding visual filter and co-occurring auditory features of this sound class. The different weighted saliency maps corresponding to the learned visual filters are then summed up in a globally fused saliency map that is used to select the next gazed point in the world. After learning, the selected gazing point has a higher probability to correspond to a visual aspect of the world that was experienced to correlate with the co-occurring sound.

Saccade/Tracking Control

The learning of voluntary saccades is carried out in a loop, in which visual and auditory knowledge is incrementally acquired and learned in association with each other. The gazing time corresponds to a learning phase. Therefore, each time the camera moves to a new position and tracks the scene for a defined time, the vision field is processed with a set of visual filters. The visual processing is initially determined by predefined filters (motion, flickr, color, orientation, intensity) but is gradually extended by learned visual filters. Each of these filters (predefined or learned) are applied on the visual field to compute a saliency map, respectively. These different saliency maps are then fused to a global one in which a winner-takes-all mechanism is used to determine the next camera gaze position. After a saccade has been performed, a tracking system takes over the control of the camera movement for a fixed time interval by keeping the center of the camera field of view on the point in the world initially chosen by the saccade movement. The duration of the tracking has been chosen to be fixed for simplicity but could easily be made variable depending for example on the confidence with which the tracked point is recognized. The control unit of the camera also records the camera movements and provides this information to the multimodal correlation module.

Multimodal Correlation

During the tracking phase, the system checks if the center part of the visual field ('fovea') correspond to an existing visual class representation by computing a kind of ranking of the classifier responses for this visual aspect. If a class representation fits the perceived visual aspect sufficiently well, it is selected as a candidate for the correlation with possible auditory sources present at that time. On the auditory side, the sounds are classified and a best class is selected based on a similar ranking of the auditory classifier response. In parallel, the time co-occurrence of onsets of motion, (fovea motion energy/local motion) and auditory energy is checked supported from infants learning capabilities (Gogate et al., 2009; Lewkowicz, 2000). When a high correlation and synchrony of such features is detected, the multimodal correlation module creates an association between the corresponding visual and auditory classes. This association is done in form of an association strength matrix which is then provided to the top-down filter weighting module for a modulation of the saliency map fusion process.

Classifier and Filter Learning

The visual classifiers and the visual saliency filters are learned autonomously in an unsupervised way during the camera tracking phase. In theory the auditory classifiers could be learned in a similar way as the visual ones but to keep the system reasonably simple the auditory classifiers were learned off-line in a supervised way in this work.

Visual Classifier and Filter Learning

The learned visual classifiers and the learned visual saliency filters are tightly linked. The former are the basis to compute the later. During a tracking phase, the system first has to decide if the tracked information in the fovea is known or not. To do so, it compares the visual features present in the fovea with reference feature vectors already learned. In case of a good match, the matching response gives the classifier answer. In case of a bad match a new class has to be created and the currently present visual features in the fovea can be used for that. When a new class is created, it can be used to build a saliency filter. This is done by computing the response of the classifier for all positions in the visual field (not only the fovea). By doing so, a saliency map is obtained where maxima correspond to locations in the visual field where this classifier responds strongly. Under the assumption that visual aspects are not highly repetitive in a normal scene (at least the probability is expected to be low) the expected response of the saliency map should be maximum in the fovea but rapidly decreasing as a function of the distance to the fovea. This theoretical response form can be used as a reference and visual features leading to a classifier response strongly differing from this reference. This can be used to increase the classifier discrimination capability by inhibition. Further pruning or fusion mechanisms inspired by infants learning and restructuring capabilities during sleep (Tarullo et al., 2011), can be used to optimize and regulate the growth of the visual filter set.

Audio Classifier and Filter Learning

The auditory classifier has been trained offline on supervised information labeled by hand (e.g. speech or knocking or sounds). It provides an instantaneous classification of the auditory stream based on frequency channel features. The classification response is used to modulate the individual saliency map in their fusion process and to learn the association strength matrix between multimodal object classes. In the auditory branch no filters are built from the classifier since no auditory attention system is considered in this work.

3.4. Summary

In this Chapter, two methods were introduced that are utilized to analyze the gazing behavior of infants in object learning scenarios. Overall, these methods are used for the analysis of the discrimination and recognition performance during object learning. The interaction with other modalities shows that infants form associations and use them to recover objects. In addition, they are able to fixate multimodal objects again irrespectively of the learned location cue by pairing visual expectations with auditory events.

Based on this principle, an architecture was presented that learns associations between objects. Here, the recognition of objects is linked to a configuration of the response behavior of visual classifiers in dependencies of auditory characteristics. This mechanism allows a robot to classify objects by means of auditory events and generates corresponding saccades to multimodal aspects. To provide such a configuration step for a robot, the next Chapter proposes an approach to learn visual objects in an unsupervised manner during the tracking to utilize them in an associative learning step. Besides the learning of visual classifiers, the next Chapter also proposes a method for pruning of redundant scene knowledge.

4. Unsupervised Acquisition of Visual Object Representations

Object learning with an artificial system like a robot is challenging mainly due to two aspects. The first one addresses changes that can occur during the demonstration of an object. This includes perturbations with respect to the object's appearance as well as changes in the environment, e.g. lighting conditions, perspectives changes or variations in reproducibility of tutor actions. The second aspect addresses online learning and unsupervised learning which refer to mechanisms that enable the robot to acquire object concepts during demonstrations. Online learning is necessary to cope with environmental changes and necessitates a serial data processing, where the entire data set is not available a-priori and the processing of it is conducted in a step-wise fashion. The results of a sequential data processing need to be memorized and available for a next decision step during the learning phase. If an artificial system is trained offline, i.e. based on a batch processing of training samples, the training data would have to reflect the environmental complexity the robot has to cope with. But this is not possible, because the situations in which an autonomous robot is utilized are typically not fully known during the design phase. Therefore, many learning mechanisms are based on unsupervised learning and try to adapt object representations during online operation to cope with environmental changes. Objects are mostly presented from different perspectives by a tutor, which means that an object appearance can change from view to view. One way to overcome this problem, i.e. to ensure a view invariant object representation, is to integrate several views into one object representation. This allows an insertion of new object views upon demand and therewith provides a method for online learning. This mechanism hence promotes object constancy - the ability to perceive objects as the same entities despite variations in their appearance.

The extension of object representations by new views requires the ability to decide whether an observation corresponds to an already known object or to an unknown object. In the latter case, it would be inappropriate to extend an object representation; rather a new one has to be created. A spatio-temporal continuity constraint can be used to resolve this ambiguity. More precisely, physics prevents abrupt changes with respect to the positions of objects. This means that consecutive observations at a particular position in space belong to the same object

with a very high probability. In contrast, observations stemming from different positions very likely correspond to different objects. Therefore, spatio-temporal continuity provides a kind of supervision signal for the classification of different observations. It supports the robot in building object hypotheses and thereby enables an appropriate acquisition of object knowledge.

In this Chapter, an approach to unsupervised object learning is proposed. The learning method incrementally acquires object representations for different demonstrated objects and extends them by additional object views if necessary. An object view is extracted with a *one-shot* learning mechanisms. Furthermore, an approach to the organization of an object memory is described. This method removes redundant visual information and finally yields an online learning method with decreased memory requirements.

4.1. State of the Art

In the following, an overview is given about different approaches that aim at learning object representations, in particular during tracking. The implementation of such a learning process is carried out in various ways. One method for object acquisition relates to unsupervised learning, i.e. the system has no prior knowledge about an object and adapts autonomously to changes in object appearances. The adaptation of internal object representations can be carried out online, i.e. the system learns while single object examples are demonstrated to it. Therefore online adaptation is especially important for model learning during tracking. In addition to this, object models in forms of classifiers need to be robust. Robust means to accomplish a high degree of generalization and discrimination capabilities. Since objects may change in their appearance, classifier responses can decline over time because the underlying view models are not representative of the new object aspect anymore. In such situations, spatio-temporal constraints can be use to improve the object model.

Many approaches aim at modeling such classifiers by using positive and negative examples in order to achieve an improved classifier performance (Javed and Ali, 2005; Babenko et al., 2009; Kalal et al., 2010). Classifiers are applied as tracking mechanisms by learning a scene separation, i.e. an autonomous labeling of negative and positive scene aspects. These methods are mostly based on the principle of semi-supervised learning, i.e. classifiers are trained with examples labeled by humans. The emphasis relies less on the selection of learning examples from the scene, but on the assumption that a scene can always be binary separated into an object and a background. Similarities between target objects and nearby information are not considered, even though it may be important to form an object

specific classifier.

4.1.1. Unsupervised Model Learning

One way of unsupervised object learning consists of an adaptive learning of feature combinations for objects. This approach was pursued by Steil et al. (2007) for the learning of filter masks, so called *adaptive scene depended filters* (ASDF). The scene is segmented based on learned feature combinations, i.e. the scene is spatially separated into the object itself and the background information. During demonstration, the tutor presents a fixed number of objects from a fixed number of different views to the vision system, in which objects are always kept in the hand. In addition, a skin-color algorithm is used to remove skin-color-specific elements of the hand. The system learns automatically an object specific filter that may comprise a features space based on color, edge or dynamic features such as motion or object velocity. Such a method is possible in the restricted context of an object-in-hand scenario. However, such a scenario does not cover all situations, since objects such as mouths or faces can not be presented in such a way.

A different method that enables an unsupervised acquisition of faces in human-robot interaction is proposed by Aryananda (2009). In this work the authors make use of the spatio-temporal continuity to extract video sequences from robot social interactions using a predefined face-detector and a face tracking device. The face sequence are then used in an offline process to cluster the faces in different person classes based on a visual similarity measure This approach uses an unsupervised clustering approach but relies on a predefined face detection and tracking that constraints the usability of the approach to face classification. The offline clustering relies on the availability of training data which makes it difficult to use such an approach in a reactive online learning system.

4.1.2. Model Learning and Tracking

Learning object models during tracking is mostly used to estimate possible object positions to enable a continuous tracking of them. Here, a serial processing of image data and an online adaptation is necessary to keep track of object changes. Furthermore, it is important to enable a robot to form object representations with little scene information. Additionally, a vision system needs to decide when to adapt to changing object appearances and therefore needs appropriate criteria. One criterion may rely on an averaging of different object appearances over time to derive possible object constancy. Another criterion may be based on spatio-temporal continuity (Makovski et al., 2008). The maintenance of such a

continuity during object learning can provide a learning mechanism for combining different object views in order to build an unified object percept. Such a combination of different views, so-called view-based models (Riesenhuber and Poggio, 2000), can be used to describe an object invariant from different perspectives and thus finally enables a robust object recognition.

Online Adaptation and Model Learning

One method of online adaptation may comprise the integration of the response behavior of various models over time. For this purpose, Triesch and von der Malsburg (2001) extracted object specific prototypes to track objects. These prototypes describe object features such as position, color, motion or intensity and set an initial configuration for the tracking process. During the tracking, prototypes are adapted in their feature configuration by a time-moving average to cope with object changes. The response behavior of prototypes gives evidences of possible object positions, where an estimation of object locations in case of occlusions is not possible. Object positions are computed from the response behavior of learned prototype statistics, where the prediction is based on velocity estimation and a linear motion model. The learning of such prototypes is restricted to low level cues and a modeling of more complex object properties such as 3D rotation remains an open issue. But such an adaptation is important to ensure the tracking of objects independently of visual changes.

The learning of models during tracking and an appropriate computation of object locations can also be based on adjacent object information. Grabner et al. (2010) developed a *supporter model*, which comprises the modeling of movements characteristics of adjacent scene aspects to use them as a feature for object description. Object positions are predicted based on a voting scheme of neighboring motion characteristics and therefore also allow the tracking of objects even when they are occluded. An automatic acquisition of visual scene knowledge and an on demand learning of them is not considered.

Benefits from Adjacent Observations

Computational object modeling from scratch and the continuous improvement of object classifiers during tracking in many approaches are based on neighboring visual aspects. More precisely, in order to build an object model, visual patches in the center part of a gazed object can be taken as descriptive (or positive) aspects of the object, whereas visual patches beyond the gazed object can be used as non-descriptive (or negative) samples of the object. Such an approach was suggested by Kalal et al. (2010), where the system automatically learns to

structure unlabeled data into positive and negative samples. The structuring of the scene relies on the use of a spatial-temporal constraint. This constraint is derived from object trajectories that result from tracking. In detail, positive samples are those which lie close to the object trajectory. Negative samples are specified by positions beyond the trajectories and therefore binarize the scene in two principal classes. Such a binarization step lacks a differentiated modeling of the scene background. In this approach, a relation in form of similarities between the tracked object and the scene background is neither considered nor related to the learning process. But this may be important for a machine learning system to enable a comparison between learned classifiers with respect to their response behaviors in order to delete redundant information accordingly. Furthermore, this approach can only learn one model for the tracked object and learning of additional object knowledge is not implemented.

However, learning of multiple object classes is important for a robot, since object learning in human-robot interaction is characterized by an unknown number of object demonstrations. This means that a system needs to acquire knowledge about multiple objects to recognize them later on. For this purpose, Javed and Ali (2005) initialized multiple classifiers from image data annotated by humans and subsequently improve the quality of a classifier in a subsequent training phase by a co-training through positive and negative filter responses of the other classifiers. This semi-supervised approach allows learning and classification of objects in one framework and takes dependencies between present classifiers into account. The analysis of response behaviors of multiple classifiers may provide a basis to identify redundant object knowledge and to reduce computational cost by removing them. Nevertheless, the number of classifiers is predefined in this approach which restricts a vision system in the knowledge it can acquire about the scene. A robot not only needs to learn multiple object classes, but also multiple instances of an object. Another approach that aims at learning object classifiers on the basis of positive and negative information was suggested by Babenko et al. (2009). He tried to cope with multiple object instances. The aim of this work was an improvement of object tracking by means of learning multiple instances of focused objects. Although the approach combines different object instances to ensure generalization performance, an appropriate process for learning multiple classifiers is missing.

4.2. A Computational Model for Object Learning during Tracking

In the following, a computational model for unsupervised object acquisition is described. This model refers to the component *visual classifier learner* of the architecture illustrated in section 3.3. First, an overview of the properties of an object representation is given. Subsequently, a method for obtaining a view-invariant object representation is described. This includes decision criteria when to insert new object views or when to learn a new object representation. The learning method further involves a fusion of already learned representations to enhance the recognition performance. Afterward, the computational model is evaluated on image sequences that show different objects based on which the discrimination and generalization performance is illustrated.

4.2.1. System Overview

Figure 4.1 shows the principles of the view invariant object representation. An object representation for object K is described by a classifier $\mathbb{F}_K$ that accommodates different views of an object in terms of view models f_j. In order to design a flexible response behavior of a classifier, it is necessary to keep the individual response characteristics of the respective view models. Therefore, each view model is equipped with a weighting coefficient c_j that is connected to the classifier. This weighting coefficient is derived by the response behavior of the view model in an initial learning phase.

The online acquisition of a view model is done in one-shot, i.e. only based on the current observation. A learned view model comprises one positive feature $\boldsymbol{s}_{j0}$ and negative features $\boldsymbol{s}_{ji}$, $i = 1 \dots n_j$, where the negative information is used to improve the response specificity of the view model. This is implemented by an inhibition of peripheral observations. Therefore each view model f_j endows weighting coefficients w_{ji} that denotes the strengths of the positive (centered information w_{j0}) and negative information (peripheral information) w_{ji}, $i = 1 \dots n_j$. In order to obtain models that are invariant to changes in scale, orientation and illumination, *Sift-Features* (Lowe, 1999) are used for their description. In summary, a three-layered representation of objects is chosen:

Object Classifier A classifier $\mathbb{F}_K$ accumulates different object views f_j and is denoted by a set of view models with $\mathbb{F}_K = \{f_1, \dots, f_m\}$.

Object View Model An object view is defined as a triple $f_j = \{\boldsymbol{s}_j, \boldsymbol{w}_j, c_j\}$ with $\boldsymbol{s}_j = \{\boldsymbol{s}_{j0}, \dots, \boldsymbol{s}_{jn_j}, \}$ and $\boldsymbol{w}_j = \{w_{j0}, \dots w_{jn_j}\}$. It consists of a set of *Sift-Features*

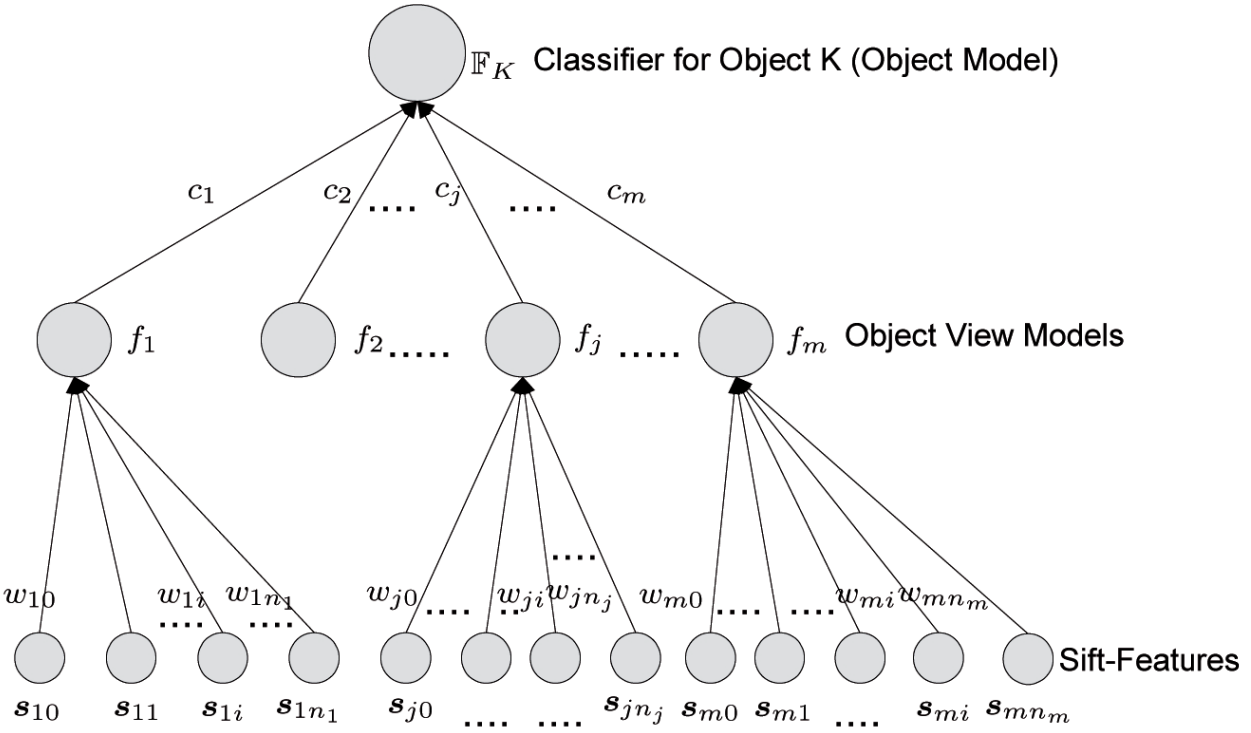

Figure 4.1.: Definition of an object representation. A classifier $\mathbb{F}_K$ (top) comprises different object view models f_j (middle) and equates an object representation (middle) which in turn are represented based on a set of *Sift-Features* (bottom).

$\boldsymbol{s}_{ji}$, learned weights w_{ji} and an extracted normalization coefficient c_j.

Sift-Features A *Scale-Invariant Feature Transform* extracts features for a view model f_j. Each feature vector $\boldsymbol{s}_{ji}$ describes a component of a view model by an orientation histogram and additionally transforms this object characteristic invariant from its scale and orientation. The original *Sift-Feature* computation (Lowe, 1999) localizes keypoints on an image which constitute relevant candidates for an object description. Each keypoint is described by its local neighborhood. For this neighborhood a histogram is computed that contains object characteristics in terms of edge orientations. Typically, a neighborhood is divided into 4x4 regions and eight possible edge orientations are assumed. This results in a 4x4x8-dimensional feature vector $\boldsymbol{s}_{ji}$. To ensure illumination invariance, this vector is normalized to unit length.

The overall learning method is show in Figure 4.2. The default visual aspects are defined by a set of filters (Itti et al., 2003) that extract a salient point. This point is kept in the central region of the current view during the observation. A new *saccade* is triggered by a timer event and the gaze is recentered on a new

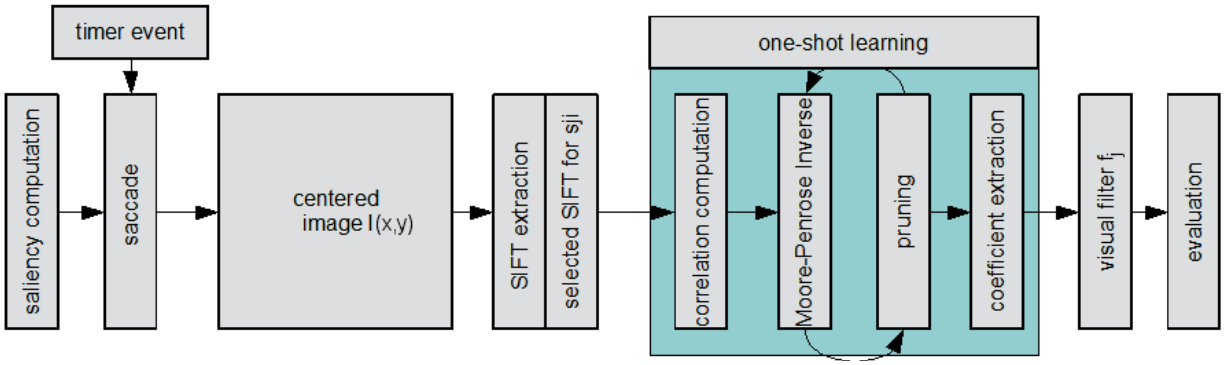

Figure 4.2.: A visual classifier learning process during tracking comprises the selection of nonlinear features and extracts a visual view model f_j with a one-shot learning process. Reprinted from Grahl et al. (2010) with permission of the editor.

salient point. Nonlinear features are extracted by *Sift-Features*. During a fixation period, an object view model f_j is learned by a one-shot learning process from the centered image $\boldsymbol{I}$. On the basis of spatio-temporal continuity constraint, the view model is either fused into an existing classifier $\mathbb{F}$ or serves as an initial view of a new classifier. In the following, the individual processing steps are explained in detail. In the first step, features are extracted from the image. Here, *Sift-Features* are used, since they are known to provide an image description that is largely invariant to changes in scale, orientation, translation, and affine distortions. The Sift-Feature extraction first detects keypoints in an image. In the present model, each pixel of an image is considered to be a keypoint. Next, local orientation characteristics are extracted for each keypoint. This is done by calculating an orientation histogram with respect to a local neighborhood of the keypoint. The resulting Sift descriptors consequently stem from partly overlapping image regions and are centered at the different pixel locations of the image.

4.2.2. One-Shot Model Learning

An image can be described by a set of *Sift-Features*. Therefore, in a first step a set of *Sift-Features* is calculated (see Fig. 4.3). The *SIFT extraction* defines each position of $\boldsymbol{I}$ as a keypoint in order to extract a Sift-Feature $\boldsymbol{s}$. Each centered image results into a set $\mathbb{S}$ of overlapping *Sift-Features* $\boldsymbol{s}$ that describe the local orientation characteristics. After the extraction of $\mathbb{S}$, the *one-shot learning* process of an object view is done by a selection of relevant features $\boldsymbol{s}_{ji}$. Therefore, the center feature and relevant peripheral features for the *j-th* visual view model are defined by $\boldsymbol{s}_{j0}$ and $\boldsymbol{s}_{ji}$. Each $\boldsymbol{s}$ is a nonlinear feature that describes a part of the current observation (image). We can also assume, that a current focused

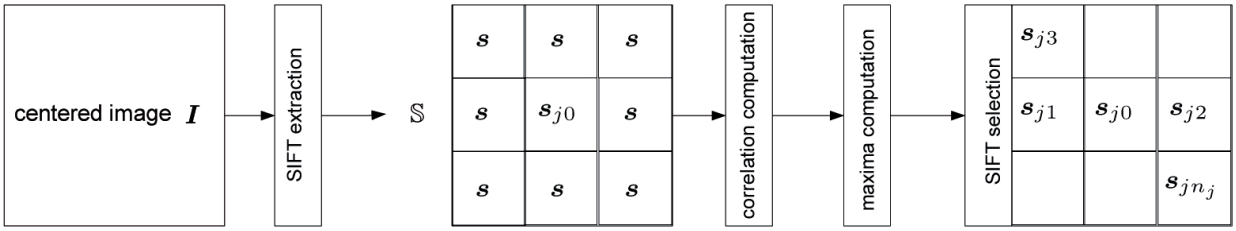

Figure 4.3.: Selection of nonlinear features for a one-shot model learning with an inhibition. The $\boldsymbol{s}_{i0}$ describes the orientation histogram for the center position. Afterwards a *correlation computation* is conducted in order to select negative samples $\boldsymbol{s}_{ji}$ for an inhibition of peripheral observations. The 3x3 decomposition of the image is just for illustration purpose.

object and peripheral observations can be described by those features.

Since we would like to learn a view model for the object that is in the center of the image, the feature $\boldsymbol{s}_{j0}$ might be an appropriate representation for it. However, an object description that is only based on a single feature is error-prone as it would lead to an unspecific response on the image.

More precisely, filtering the image with $\boldsymbol{s}_{j0}$ will not only yield a peak response at the center location, but also at positions that share similar local orientation characteristics. Since such spurious side-peaks should be suppressed, it is not sufficient to describe the object view solely based on the center feature $\boldsymbol{s}_{j0}$. Rather, an inhibition of these side-peaks can be done by additionally including those peripheral features $\boldsymbol{s}_{ji}$ that correspond to the side-peak locations. This inhibition mechanism will finally result in an enhanced specificity of the object model. What remains is a selection of appropriate features $\boldsymbol{s}_{ji}$. The whole process is based on the assumption that the object under consideration is only present in the center of $\boldsymbol{I}$. This assumption is invalid in scenes with repetitive structures or multiple instances of the same object but we consider these cases to be statistically rare. Therefore, let $\boldsymbol{\Phi}_{j0}$ denote the response of the feature $\boldsymbol{s}_{j0}$ on the image. It can be calculated as

$$\boldsymbol{\Phi}_{j0}(x, y) = \boldsymbol{s}_{j0} \circ \mathbb{S}(x, y) \; . \tag{4.1}$$

where $\circ$ is the dot product operator and $\mathbb{S}(x, y)$ the Sift-Feature $\boldsymbol{s}$ extracted from the image location (x, y). $\boldsymbol{\Phi}_{j0}$ constitutes a feature map in which those locations exhibit a large response, where the original image $\boldsymbol{I}$ correlates with $\boldsymbol{s}_{j0}$. To enhance the specificity, $\boldsymbol{\Phi}_{j0}$ is subject to a *local maxima search*. The *local maxima search* corresponds to an extraction of those $\boldsymbol{s}_{ji}$ that exhibit a large response correlation with $\boldsymbol{s}_{j0}$. This means that those peripheral locations (x, y) can be selected in which $\boldsymbol{s}_{j0}$ yields a large response. The response of these locations can be suppressed by inhibiting the feature maps with the response of the feature $\boldsymbol{s}_{ji}$.

This response is obtained by filtering the image $\boldsymbol{I}$ in a convolution-like way with $\boldsymbol{s}_{ji}$ that is explained in the following.

This *selection* step results in a set of nonlinear features that contain one positive $\boldsymbol{s}_{j0}$ for the center position and a set of negative $\boldsymbol{s}_{ji}$ for the inhibition in the periphery. For an inhibition of peripheral scene information, it is required to compute a set of feature maps $\boldsymbol{\Phi}_{ji}$ in order to equip the features with appropriate negative weights. This computation comprises a *correlation computation* that is similar to Eq. (4.1) and results in a set of feature maps represented in the matrix $\boldsymbol{\Phi}_{ji}$: a matrix of the size that corresponds to the number of features $\boldsymbol{s}_{ji} \times$ number of pixels in the image.

$$\boldsymbol{\Phi}_{ji}(x, y) = \boldsymbol{s}_{ji} \circ \mathbb{S}(x, y) \ . \tag{4.2}$$

These feature maps are used for the weight initialization. In order to compute the weighting function of the positive and negative features $\boldsymbol{s}_{ji}$ to the view model f_j, we consider the training of a single layer perceptron where the training samples are the filter responses $\boldsymbol{\Phi}_{ji}$ and the supervised answer of a Gaussian response image $\boldsymbol{G}$ with maximal activity in the center. More precisely, the weights are learned in *one-shot* via the *Moore-Penrose Inverse* + (Haykin, 1999) of $\boldsymbol{\Phi}_{ji}$ with $\boldsymbol{G}$ as an error function:

$$w_{ji} = \boldsymbol{\Phi}_{ji}^{+}\boldsymbol{G} \quad \text{with } \boldsymbol{\Phi}_{ji}^{+} = (\boldsymbol{\Phi}_{ji}^{T} \cdot \boldsymbol{\Phi}_{ji})^{-1} \text{ and } \boldsymbol{G} = exp\left(\left(-\boldsymbol{x}^2 - \boldsymbol{y}^2\right)/\sigma^2\right) \ . \tag{4.3}$$

A weight estimation that is based on Eq. (4.3) can result in positive weights for the response of negative samples $\boldsymbol{s}_{ji}$. Therefore, an additional *pruning step* is conducted for those weights $w_{ji} > 0$. This step removes false positive w_{ji} and corresponding $\boldsymbol{\Phi}_{ji}$ and computes w_{ji} again in order to obtain negative weights with $w_{ji} < 0$.

The entire response $\boldsymbol{Y}_j$ of a view model f_j can be computed according Eq. (4.4). It is defined by the weighted linear combination of the computed w_{ji} and $\boldsymbol{\Phi}_{ji}$, where w_{ji} are the inhibition and excitation weights. Additionally, each object view is equipped with a coefficient $c_j > 0$ that serves as normalization

$$\boldsymbol{Y}_j = \frac{1}{c_j} \cdot \sum_{i=0}^{n_j} w_{ji} \ \boldsymbol{\Phi}_{ji} \text{ with } c_j = \sum_{x,y} \boldsymbol{G} \cdot \sum_{i=0}^{n_j} w_{ji} \ \boldsymbol{\Phi}_{ji} \ . \tag{4.4}$$

During an object fixation visual aspect changes can make an object view model looses its validity and shows a smaller response. This is compensated by a new object view or a completely new object that is not specified by a current classifier. Therefore, it is necessary to introduce a method to measure the response behavior of a model f_j. A possible method relies on a normalization step of the response. For this purpose, a mean average response can be computed that is spatially

weighted. This information enables the specification of the expected response behavior during the initial learning phase. More formally, a derived mean allows a weighting of a view model response such that the expected activity a_j equals 1 to indicate an exact match of an object view model with a currently observed object:

$$a_j \; = \sum_{x,y} \boldsymbol{Y}_j \cdot \boldsymbol{G} \; = \; 1 \; . \tag{4.5}$$

A classifier $\mathbb{F}_K$ accommodates multiple object views, because as much as possible appropriate object views should be presented in order to enable an object classification. Desirably, a classifier flexibly responses to different object views that are obtained by learned model views. One possibility to retain this flexible response behavior is to calculate each view model response independently from the others and subsequently sum them to obtain the overall classifier response. Therefore, it makes sense to compute a weighting coefficient for each view model f_j.

To do so, during tracking the response of f_j on the centered image is calculated via Eq. (4.4), where the overall response $\boldsymbol{Y}_{j^*}$ of the classifier $\mathbb{F}_K$ is given by the maximum response of the integrated views:

$$a^{'}_{j^*}(K) = \sum_{x,y} \boldsymbol{G} \cdot \boldsymbol{Y}_{j^*} \text{ with } \mathbb{F}_K = \boldsymbol{Y}_{j^*} = \max_j (\boldsymbol{Y}_j) \; . \tag{4.6}$$

The resulting center activity $a^{'}_{j^*}(K)$ is compared to a_j estimated from a filter in the initial learning phase. A new filter is inserted, if the current $\boldsymbol{Y}_{j^*}$ does not fulfill the object hypothesis, i.e. $a^{'}_{j^*}(K) < \theta_1$, where θ_1 is a threshold. This means during the observation of an object the method integrates multiple views f_j into one visual classifier $\mathbb{F}$.

4.2.3. Model Learning during Tracking

During a saccade an *evaluation* step is conducted. Its purpose is to decide whether already learned object models describe the current observation sufficiently well or whether a new model has to be learned. The first strategy evaluates the center activity $a^{'}_{j^*}(K)$ and decides if an already learned classifier $\mathbb{F}_K$ for object K is used for the description of the currently observed object. If so ($a^{'}_{j^*}(K) \geq \theta_2$), the classifier is extended by the current view model $\mathbb{F}_K = \mathbb{F}_K \bigcup \{f_j\}$. If multiple classifiers $\mathbb{F}_I$ respond above a defined threshold θ_2, they are combined into one classifier, i.e. $\mathbb{F}_K = \bigcup_I \mathbb{F}_I$. In the case that no classifier is available for the currently observed object ($a^{'}_{j^*}(K) < \theta_2$), a new visual classifier is defined by $F_{K+1} = \{f_{new}\}$.

4.2.4. Evaluation

The evaluation of our method is based on a video sequence that shows a person who demonstrates a cup stacking task. A saccade movement is determined from a saliency map that uses color, orientation, motion and intensity features. An inhibition of return leads to a gaze selection that has not been attended before. A new saccade is triggered every second and partitions the demonstrated task into sequences of tracked objects. The mechanism used for object tracking is based on a modified version of (Triesch and von der Malsburg, 2001). Those sequences that have more than one object in the center, were removed from the data set in order to analysis the principal response behavior of the proposed computational model. In the next Chapters, the analysis bases on image sequences that contain more than one object in the center.

For the extraction of *Sift-Features*, we resize the images from 525x525 pixels to 159x159 pixels. A Sift-Feature describes each pixel in 8 orientations for 4x4 spatial bins extracted from a region with a size of 25x25 pixels. The Gaussian kernel $\boldsymbol{G}$ is computed with $\sigma = 0.05$ (see Eq. (4.3)).

One-Shot Model Learning

At first we evaluate the performance of a one-shot learned visual view model f_j and the insertion of sub views f_j into one classifier $\mathbb{F}_K$ during an *evaluation* step. The performance of both are compared to that of a simple visual view model f_{j0} that does not use the inhibition of peripheral side peaks. In other words f_{j0} only relies on $\boldsymbol{s}_{j0}$ with $w_{j0} = 1$. For this, we show an example for the one-shot learning method and an example for the insertion of new view models during the tracking of a hand. An example of the one-shot learning method is shown in Fig. 4.4. The depicted images 0-7 correspond to the feature maps $\boldsymbol{\Phi}_{j0}$ and $\boldsymbol{\Phi}_{ji}$ belonging to the positive feature $\boldsymbol{s}_{j0}$ and the negative features $\boldsymbol{s}_{ji}$. The image 0 shows the response of a simple visual model f_{j0} and corresponds to the feature map $\boldsymbol{\Phi}_{j0}$. Image 8 shows the learned visual model with an inhibition of peripheral regions, e.g. those that corresponds to the face or the cup.

The feature map depicted in image 0 serves as a basis for the *maxima search* and results from Eq. (4.1). The green square marks the positive sample $\boldsymbol{s}_{j0}$, whereas red squares mark negative samples $\boldsymbol{s}_{ji}$ used for the inhibition. The corresponding feature maps $\boldsymbol{\Phi}_{ji}$ are shown in images 1-7. By combining image 0 with the

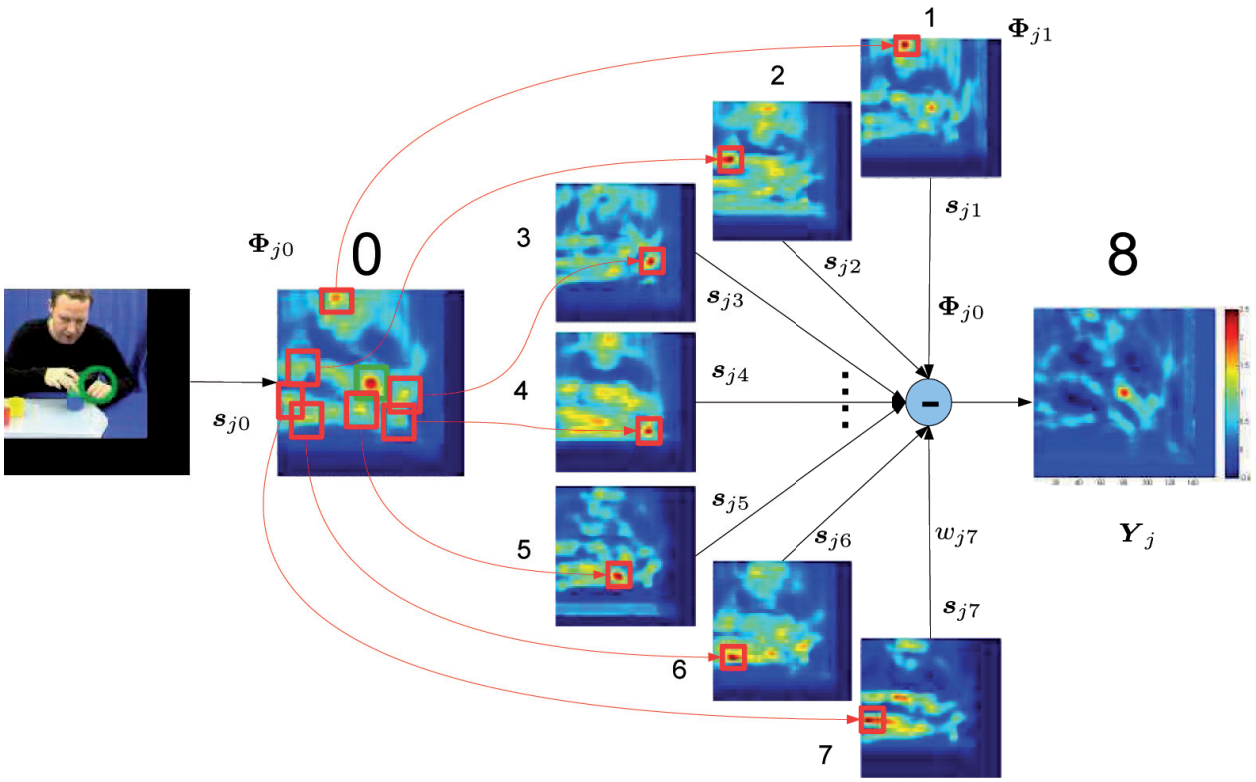

Figure 4.4.: Example of a learned one-shot hand model (from left to right) resulted from a linear combination of one positive feature s_{j0} and negative features s_{ji} extracted from peripheral observations. Φ_{j0} and Φ_{ji} are the corresponding feature maps. The black border on the right and bottom part of the input image is due to the simulation of camera movements on a fixed size pre-recorded video stream.

suppression by images 1-7, we obtain the overall model response $\boldsymbol{Y}_j$ depicted in image 8. Thereby, view model f_j uses features weights w_{ji}= (0.57,-0.08,-0.08 ,-0.11,-0.07,-0.02,-0.03,-0.10), a center activity a_j=1 and a normalization coefficient c_j=0.11. The first entry of w_{ji} corresponds to the weighting term for the positive sample response $\boldsymbol{\Phi}_{j0}$. The remaining negative weighting terms are used for the negative sample responses $\boldsymbol{\Phi}_{ji}$. The comparison of both view models (image 0 and image 8) demonstrates that the inhibition with negative samples extracted from the periphery enhances the model specificity, so that it is mainly responsive to the center observation.

Figure 4.5 shows the integration of different view models f_j with respect to changing views of a hand. The insertions are marked with enlarged pictures (from left to right). The phases 1 and 2 mark the center activity $a'_{j*}(K)$ before and after the insertion as well as the corresponding saliency maps $\boldsymbol{Y}_{j^*}$.

In a first step f_1 is learned in one-shot. Subsequently, three additional view

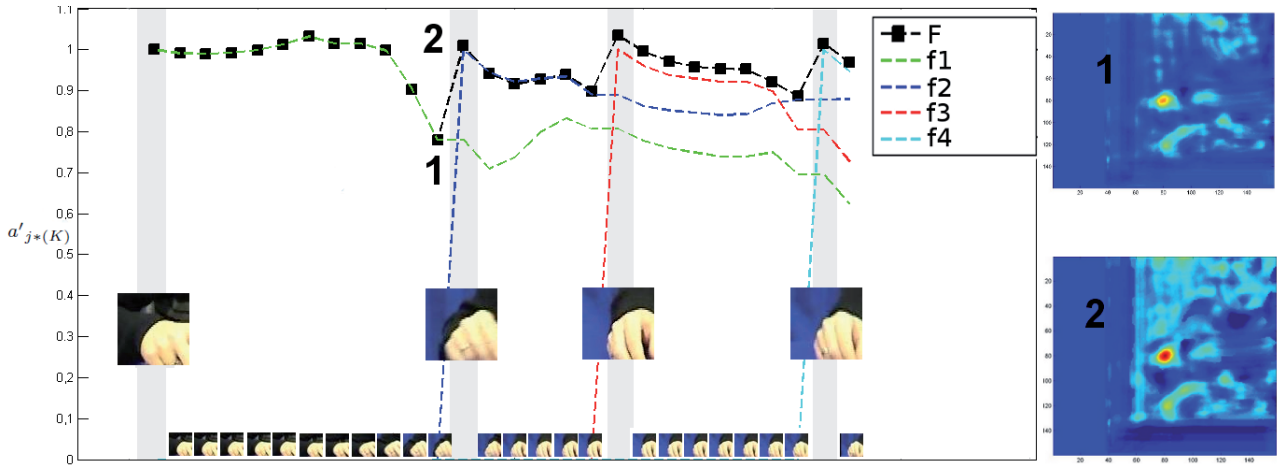

Figure 4.5.: Example of insertions (gray) of f_j during tracking into one visual classifier $\mathbb{F}$. Image 1 shows the response $\boldsymbol{Y}_1^*$ (below the threshold $\theta_1 = 0.9$). Image 2 shows the corresponding $\boldsymbol{Y}_1^*$ after an insertion of f_2. Reprinted from Grahl et al. (2010) with permission of the editor.

models are gradually inserted into the classifier $\mathbb{F}$. Phase 1 shows that the current model f_1 is no longer valid, since a'_{1*} decreases. In phase 2 the center activity a'_{1*} is improved by the insertion of f_2. The modification of the classifier again yields to a specific response at the observed hand (see images 1 and 2).

In order to evaluate the performance of our approach, we compare the different averaged center activities of f_{10}, f_1, and the resulting classifier $\mathbb{F}$ that integrates different views. The models f_{10} and f_1 are derived from the first image of the tracked sequence. The performance is tested on a data set which contains *true positive (tp)* and *true negative (tn)* hand samples. The top left images in Fig. 4.6 shows the 33 *tp* samples, whereas the top right images show the 22 *tn* samples without the appearance of a hand. The bars in Fig. 4.6 depict the average response of the three models to these samples. The result illustrates that the inhibition model f_1 shows a better performance than the simple view model f_{10}. The simple model f_{10} shows a high response to both data sets, which results in many false positive detections.

In contrast to this, f_1 and $\mathbb{F}$ also show a large response to *tp* samples, but suppress their activity for objects of other classes (*tn* samples). Both show a high activity to different hand views. Thereby, $\mathbb{F}$ shows on average a higher activity with respect to *tp* samples and therefore performs better than f_1.

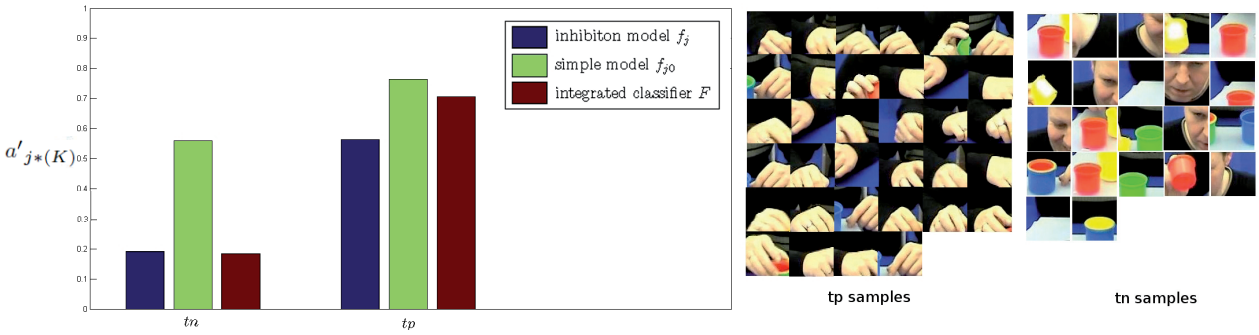

Figure 4.6.: Comparison of a simple visual model f_{10} (green), an inhibition model f_1 (blue), and a classifier $\mathbb{F}$ (red) consisting of several object views. Bars mark their average response $a'_{j*}(K)$ to true positive (tp) and true negative (tn) samples of hands. The test samples are depicted on the right. Reprinted from Grahl et al. (2010) with permission of the editor.

Fusion of Visual Classifiers

In a first step, we show the fusion process of two visual classifiers $\mathbb{F}$. The evaluation is based on six fixation sequences that capture a left and a right hand. In a second step, we apply our learning method on image sequences that exhibit different objects. The fusion of visual classifiers is depicted in Fig. 4.7 (from left to right). The small image patches at the bottom always show the starting positions of the observation. The different phases of the learning method are depicted. Different color bars show the decision phases of the model learning process during a fixation that are explained below the figure. The corresponding classifier responses $\boldsymbol{Y}_{j*}$ are shown on the right hand side. The black and blue lines show the activity course for the left and the right hand. In phases 1 and 2 learned models are fused to improve the visual classifier for the left hand. In phase 3 a new model for the right hand is extracted. In phases 4 and 5 already learned models are again fused for a separate recognition of both hands. In phase 6, the classifiers for the left and the right hand show a similar activity $a'_{j*}(K) > \theta_2$ and are fused into one visual classifier. The resulting classifier response is shown in image 6. The classifier is now able to detect both hands. The response behavior with respect to the absence of hand models is evaluated in the following section and based on image sequences that shows different objects.

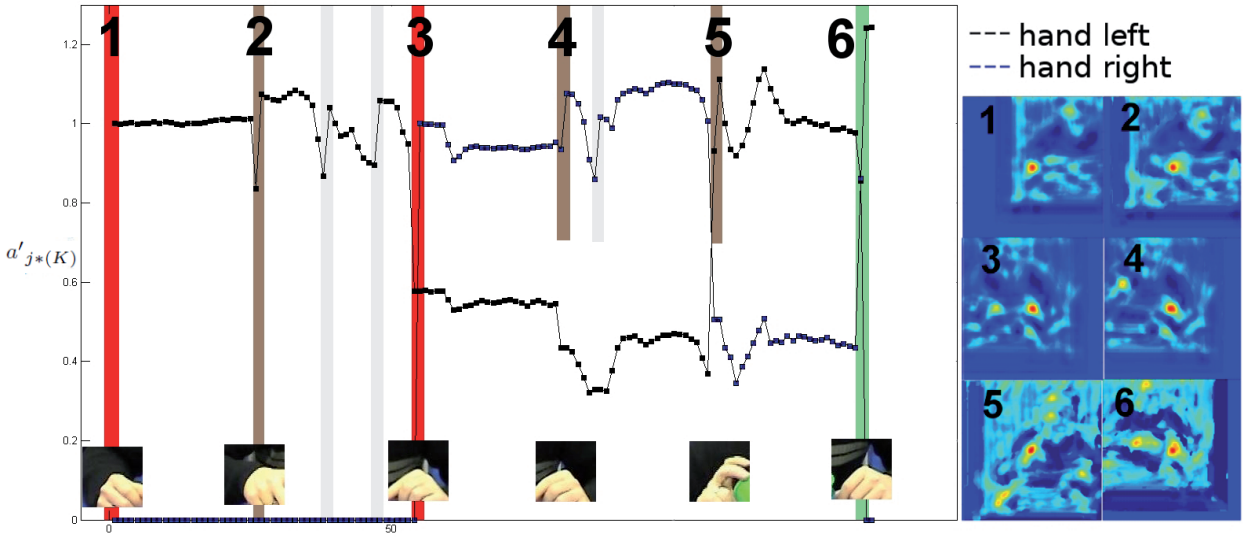

Figure 4.7.: Fusion of visual models during tracking and corresponding $a'_{j*}(K)$ with $\theta_1 = 0.9$ and $\theta_2 = 0.75$. The red bar marks the learning of a completely new visual model, brown depicts the fusion of a newly learned visual model with an already existing one, and gray shows an insertion of new views. A green bar marks the fusion of two existing models with a newly learned model. Reprinted from Grahl et al. (2010) with permission of the editor.

Learning Classifiers for Multiple Objects

The classifier learning during the tracking of different objects is depicted in Figure 4.8 (from left to right). These objects are cups, hands and faces. We evaluated 15 sequences that resulted in 6 visual classifiers. The bars in Fig. 4.8 show the activity $a'_{j*}(K)$ of the learned visual classifiers during a fixation, whereas images at the bottom show the corresponding locations. As can be seen, the learning process incrementally adds new visual classifiers. The classifiers for the right and the left hand slowly converge into one classifier. The responses of other filters show less activities in case of hand fixations. Cups and faces are also learned. Their responses are specific to cups and faces, respectively. This means they do not yield an activity for other objects (e.g. a hand).

4.3. Pruning of Acquired Visual Information

In the computational framework presented in the previous section, a new one-shot model is either inserted into an already learned visual classifier $\mathbb{F}_K$ or it defines a completely new classifier. This decision is carried out after each saccade and

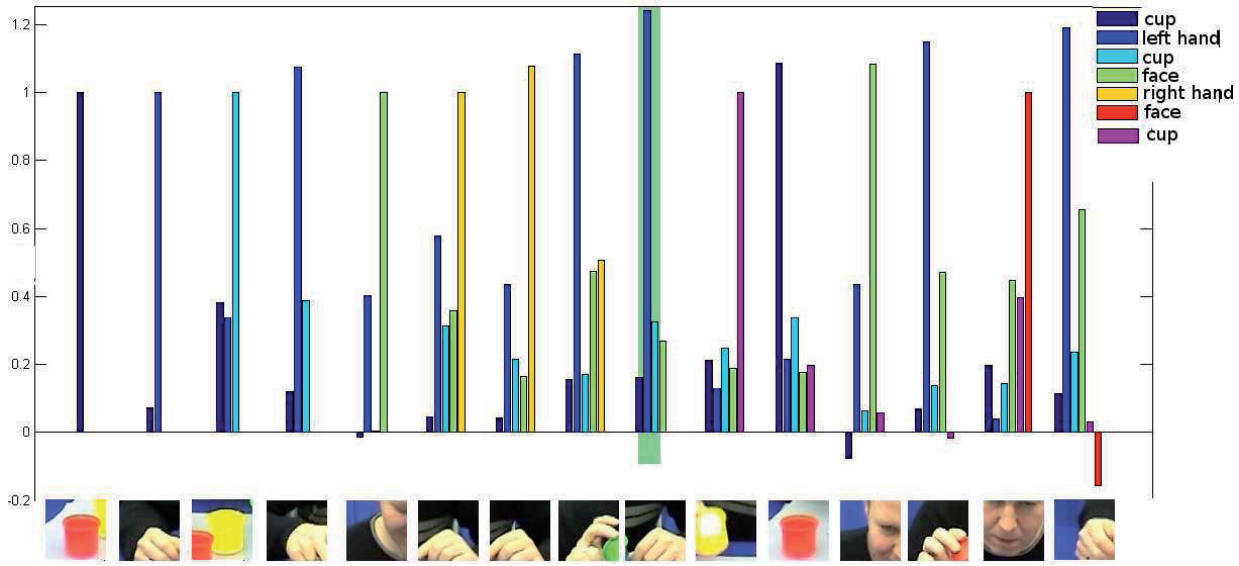

Figure 4.8.: Activity of different learned visual classifiers during scanning the scene with θ_1=0.9 and θ_2=0.75. After a fusion (green bar) of the 'left hand' classifier with the 'right hand' classifier, the resulting classifier shows a high response characteristic to both hands. Reprinted from Grahl et al. (2010) with permission of the editor.

depends on the responses of already stored object views to the current observation. As a consequence, the number of classifiers initially increases as long as new observations are identified. However, over time it converges since more and more observations can be described by already acquired classifiers. In contrast to this, the number of view models (and therewith also the number of *Sift-Features*) linearly increases during the exploration of a scene because with each saccade a visual model is inserted. This implicates unfavorable memory costs and processing time. In order to overcome this restriction, the following section examines two strategies. The first strategy modifies the model fusion step described in Section 4.2.3. It re-uses learned visual classifiers in order to minimize the number of models and features. A second strategy relies on a *combination* of a retrieval of visual information and an additional pruning of positive and negative features s from the current memory.

4.3.1. Re-Use of learned Visual Classifiers

In section 4.2.3, we proposed a fusion of learned visual classifiers. The proposed strategy thereby assumed that each one-shot model is integrated in the current object memory. This leads to an immense increase in the number of stored object

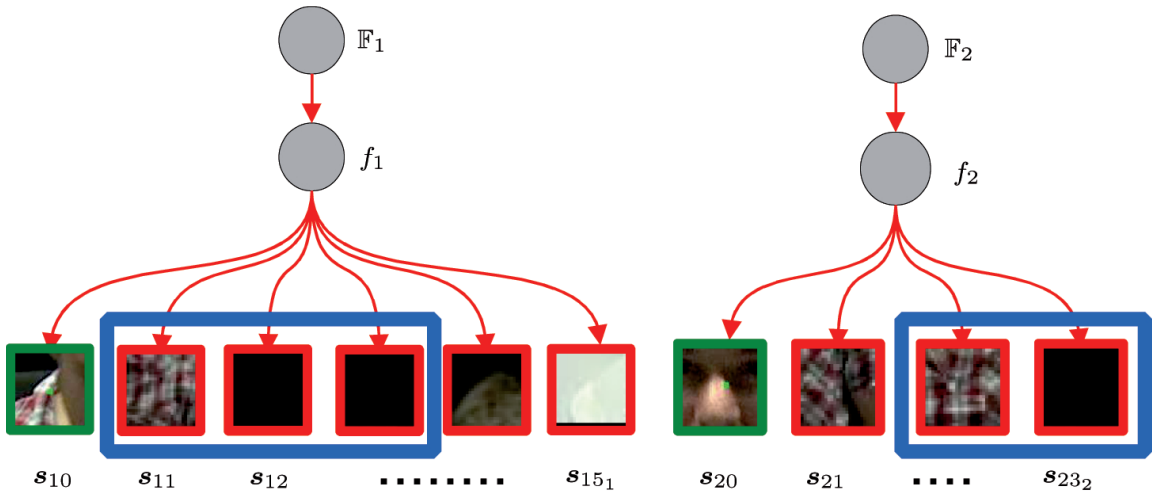

Figure 4.9.: An example of two visual view models f_1 and f_2 which partly rely on similar features. Positive features are marked green and negative features are marked red. Similar responding *Sift-Features* s are colored blue.

views and corresponding *Sift-Features*. In the following, the proposed method is modified, such that it re-uses already learned visual classifiers $\mathbb{F}_K$ without a concatenation of similar responding visual models f_j. If several models are responding to a current observation, the most representative model is used for the learning process. A new visual classifier $\mathbb{F}_{K+1}$ with $\mathbb{F}_{K+1} = \{f_{new}\}$ is only created, if no existing classifier adequately captures the current observation, i.e. the center activity is $a'_{j^*(K)} < \theta_2$ instead of always creating a new model. As this strategy does not incorporate new object views into existing classifiers, each classifier is solely composed of a single object view.

4.3.2. Pruning of Positive and Negative Features

The second strategy addresses a pruning of features $\boldsymbol{s}_{ji}$ on which the object views f are based on. Therefore, Figure 4.9 first demonstrates the problem of the linear increase in the number of features during the learning process. It is a result of redundant occurrences of features within the individual object views. The example shows two classifiers $\mathbb{F}_1$ and $\mathbb{F}_2$ consisting each of a single view f_1 and f_2, respectively. The two models represent a person's face as well as a part of the neck. Thereby, the models rely on partly similar features (colored blue). The view f_1 further shows a feature overlap within itself. In order to reduce the number of features s, redundant ones should be removed from the object memory. We consequently propose an organization of the object memory as shown in Fig. 4.10. Key to the framework is that different views can share

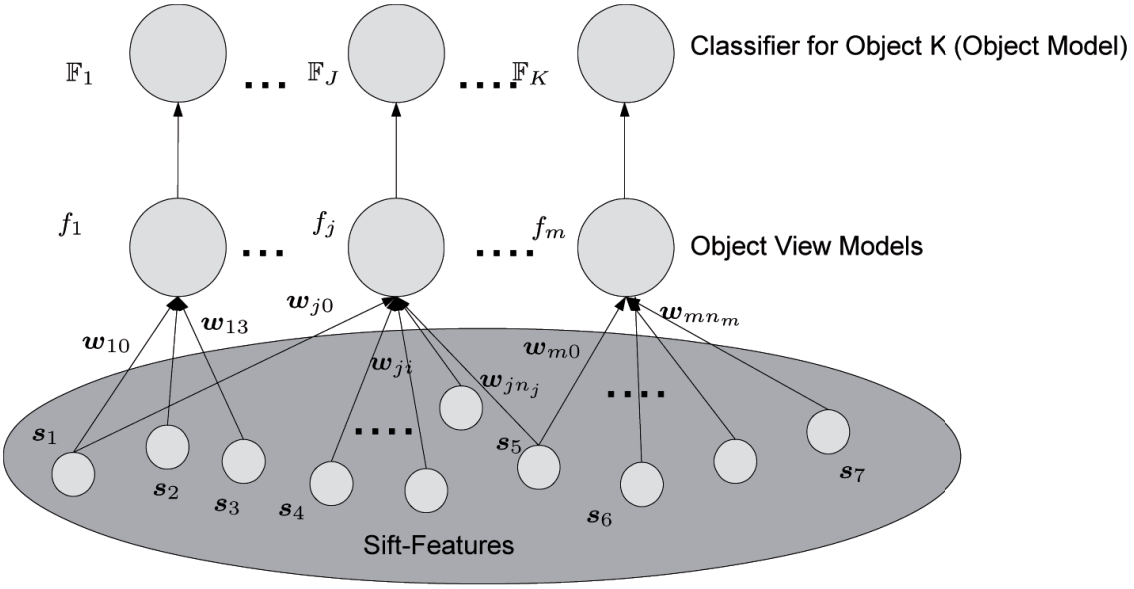

Figure 4.10.: Scheme of the object representation: Different view models f_j of different classifiers $\mathbb{F}_K$ can share features s.

features, which differs from the previously proposed system (see Fig. 4.1). Here, we additionally restrict classifiers to consist of only one view since we focus on the examination of redundant response behaviors between learned object models. The proposed framework could be easily extended to cope with multiple views per classifier and is applied in the next Chapter. An example of the pruning process is show in Fig. 4.11, where redundant Sift-Features are reduced within a classifier $\mathbb{F}_1$ that comprises three features. After a pruning step, the memory is reduced to the two features s_1 and s_2, where the third feature s_3 is further represented by the feature s_2 that additionally keeps the weight w_{12}. The maintenance of feature specific weights is important, since each weight is learned in the context of a specific view model. Otherwise the learned view model would decrease its specificity and this effect is not desirable for an object classifier. An example of the new memory scheme is shown in Fig. 4.12. It illustrates the organization of the feature set before and after a pruning step. On the left, the scheme displays 10 *Sift-Features*. They are used by the object view models depicted in Fig. 4.9. After a pruning step, the feature set is reduced to seven filters which are partly shared by both classifiers. The accompanied weights w_{ji}, that resulted from the *Moore-Penrose Inverse* for each view model f_j, are kept. This means that the response of a view model f_j to the current observation can still be calculated as a weighted superposition of feature maps (see Eq. (4.4)). The introduced memory scheme hence assigns a set of weights to each Sift-Feature according to its use in

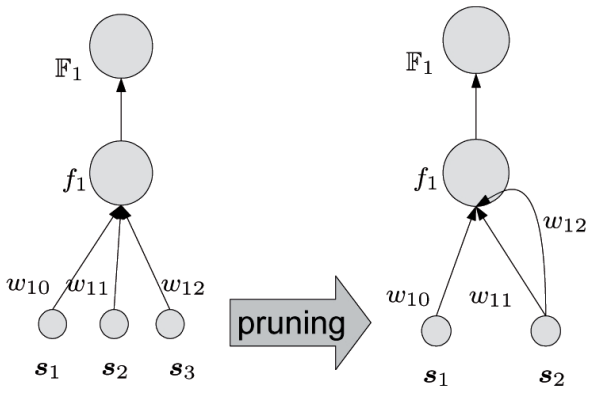

Figure 4.11.: Example of the object memory before and after pruning. The number of features is reduced from 3 to 2 features.

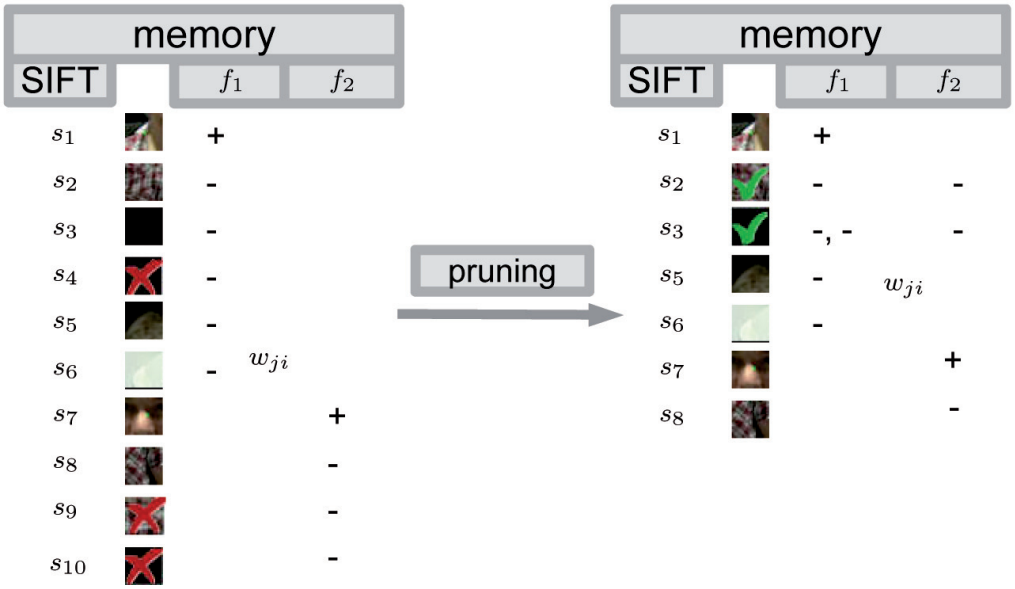

Figure 4.12.: Organization of features $\boldsymbol{s}$ of the visual view models depicted in Fig. 4.9 in a memory-based scheme. The accompanied weights $\boldsymbol{w}_{ji}$ (+,-) are retained that are extracted during the one-shot learning process. Red crosses symbolize the removal of redundant features and green check marks show the retained samples after pruning. The features $\boldsymbol{s}_4$,$\boldsymbol{s}_9$ and $\boldsymbol{s}_{10}$ are pruned. They are further represented by the features $\boldsymbol{s}_2$ and $\boldsymbol{s}_3$ such that f_1 and f_2 can share them.

the different view models f. Any feature can consequently be used as a positive or a negative sample within the different view models.

An appropriate pruning of redundant features requires a *similarity computation* that estimates the similarity between the responses of the features. To do so, the corresponding feature maps $\boldsymbol{\Phi}$ can be compared. In order to cope with changing object appearances, the similarity measurement is further smoothed over time, i.e.

a running average of the similarity between feature maps is used as a criterion for feature pruning. The feature pruning itself selects similar responding candidates to remove redundant features in the object memory.

Activity Detection of Feature Maps

In order to enhance the response specificity of feature maps $\mathbf{\Phi}_j$ irrespective of environmental changes, the maps are threshold by their averaged activities. This results in feature vectors $\mathbf{\Phi}_j^a = [\boldsymbol{\phi}_{j,1}^a, \boldsymbol{\phi}_{j,2}^a, \ldots, \boldsymbol{\phi}_{j,l}^a]^T$ composed of l elements with $l = x \cdot y$ (see Eq. (4.7)). Thereby, $\mathbf{\Phi}_j^a$ is normalized to have length $\|\mathbf{\Phi}_j^a\| = 1$, where $\|\cdot\|$ denotes the Euclidean vector norm with $\|\boldsymbol{e}\| = \sqrt{\sum_i \boldsymbol{e}_i^2}$.

$$\boldsymbol{\phi}_j^a(x,y) = \begin{cases} 0, & \text{if } (\boldsymbol{\phi}_j(x,y) - \frac{1}{l}\sum_{xy}\boldsymbol{\phi}_j(x,y)) \leq\ 0\ . \\ \frac{(\boldsymbol{\phi}_j(x,y) - \frac{1}{l}\sum_{xy}\boldsymbol{\phi}_j(x,y))}{\|\boldsymbol{\phi}_j(x,y) - \frac{1}{l}\sum_{xy}\boldsymbol{\phi}_j(x,y)\|}, & \text{otherwise} \end{cases} \tag{4.7}$$

Similarity Computation of Feature Responses

A similarity between feature maps $\mathbf{\Phi}_j^a$ is organized in a similarity matrix $\boldsymbol{A} \in \mathbb{R}^{M \times M}$, where M is the total number of *Sift-Features*. Thereby, an element $a_{m,n}$ corresponds to the measured similarity between two feature maps $\mathbf{\Phi}_m^a$ and $\mathbf{\Phi}_n^a$. The compared feature maps correspond to the m-th and n-th *Sift-Features* $\boldsymbol{s}_m$ and $\boldsymbol{s}_n$ that are currently stored in the memory scheme (see Fig. 4.12). The similarity between two feature maps is calculated based on the normalized Euclidean distance between them.

$$\hat{a}_{m,n} = 1 - \frac{\|\mathbf{\Phi}_m^a - \mathbf{\Phi}_n^a\|}{\sqrt{\|\mathbf{\Phi}_m^a\| + \|\mathbf{\Phi}_n^a\|}} \tag{4.8}$$

In order to capture spontaneous changes of $\boldsymbol{A}$ that may result from occlusions or changing object appearances, the *similarity computation* is smoothed over time. Each entry of $\boldsymbol{A}$ is updated at each tracking step according to Eq. (4.9). Thereby, each current similarity $\hat{a}_{m,n}$ is weighted with the learning rate $\lambda{=}\frac{1}{t}$ according to its temporal presence in the averaging process. This means, in the initial phase a similarity $\hat{a}_{m,n}$ is strongest weighted, whereas later similaries are less taken into account.

$$a_{m,n}(t) = a_{m,n}(t-1) + \lambda \cdot [\hat{a}_{m,n}(t) - a_{m,n}(t-1)] \tag{4.9}$$

Candidate Selection and Pruning

During each gaze fixation, a set of similar responding features $\boldsymbol{s}$ are selected that are used for a pruning step. To do so, the computed similarity matrix $\boldsymbol{A}$ is examined to find redundant responding features based on their response similarity. To decide which features can be removed or rather be representative for the further object learning, these features need to be selected from the object memory. This can be implemented by those candidates that reveal a large similarity $a_{m,n}(t)$.

Since we would like to prune those features that are similarly responding in other view models, we select those features that exhibit a large similarity value with $a_{m,n}(t) \geq \theta_3$, where θ_3 is a threshold. In order to facilitate the pruning process, we restrict the computation on the lower triangular matrix of $\boldsymbol{A}$. As the object memory is incrementally updated, a straightforward pruning of features s_n is computed in a loop, where each m-th and n-th entry of $\boldsymbol{A}$ is examined. The feature pruning is continued for so long as no similarity value exceeds the threshold θ_3 and is implemented with:

while $\max_{\boldsymbol{A}} \geq \theta_3$ **do**
  $\boldsymbol{s}_m \leftarrow \boldsymbol{s}_n$
  for $i = 1 \rightarrow M$ **do**
    $\boldsymbol{A}(i, n) = 0$
  end for
end while

This process enables a pruning of several $\boldsymbol{s}_n$, such that they can represent by only one feature $\boldsymbol{s}_m$. Redundant features are consequently removed after an assignment.

4.3.3. Evaluation

In the following, the performances of the different pruning strategies are assessed. Thereby, the method proposed in Section 4.2.3 serves as a baseline against which the newly introduced strategies are compared. In addition, the strategy 'learning without inhibition' comprises object view models without consideration of peripheral scene knowledge and is compared to the new introduced methods. The method presented in Section 4.3.1 will be termed 're-use' strategy, since it reuses the object view models that already have been acquired. In contrast, the term 'combination' strategy refers to the method which removes redundant SiftFeature (see Section 4.3.2). The latter method is further evaluated with respect to the influence of the activity detection mechanism. This is done by comparing

two simulation runs. The first run uses the activity detection ('combination w/ AD'), whereas the second one does not make use of it ('combination w/o AD').

The evaluation is carried out in two phases - a training and a testing phase. In the training phase, the computational framework does not have any prior knowledge on the objects. This means that learning starts from scratch, but continuously acquires object knowledge over the course of development. The training phase, thus, allows an investigation of the framework's learning dynamics. In contrast, the testing phase does not involve learning. We rather use the knowledge acquired during training and assess the system performance on a set of samples that have not been used for training. This allows an investigation of the discriminative power as well as the generalization capability of the developed representations.

Training Phase

The evaluation of the pruning methods during a training phase is based on a video sequence that shows a person who is knocking and speaking [†]. The training set comprises 100 simulated gaze fixations of the scene. After the saliency computation, the data set is manually partitioned into images that contain mouths, hands, or other fixations (e.g. eyes), respectively. Overall, the data set contains 4 fixations on mouths, 11 fixations on hands, and 85 fixations on other parts of the scene.

Figure 4.13 depicts the development of the number of *Sift-Features* during training for each strategy, respectively. As can be seen, the baseline strategy as well as a classifier learning with an inhibition results in an approximately linear increase of the number of features. This is due to the fact that the method continuously memorizes object views as well as their underlying features. On the contrary, the number of features converges for the other pruning strategies. Overall, the plots demonstrate that the pruning strategies are beneficial for reducing memory requirements. For all pruning strategies the object memory initially increases. This is due to the fact that the system does not rely on prior knowledge on the objects and hence has to memorize the observations. However, over time the 're-use' and the 'combination' strategies use already acquired information for the description of new observations. They consequently have to extend their object

[†]Saccade movements are determined from a saliency computation that is computed by Itti et al. (2003). The frequently triggering of new saccades serves to increase the number of new classifier creation. The images are resized from 1079×1919 pixels to 221×393 pixels that corresponds to the resizing factor used before. The Gaussian Image $\boldsymbol{G}$ is adapted to odd numbers of image length and width and computed with $\sigma = 8$. An insertion of new visual classifier is defined with $\theta_2 = 0.75$ during a saccade. The candidate selection from the similarity matrix is conducted with a threshold $\theta_3 = 0.85$.

memory only if existing models do not cover the new observations appropriately. Hence, the number of features does not linearly increase, but rather increases sub-linearly at a much lower rate over time. The object memory (in terms of the number of features) is significantly reduced as compared to the baseline method. The 're-use' strategy finally results in a medium amount of features, whereas the best results are obtained using the 'combination' strategy. The latter method is particularly beneficial when activity detection is excluded. In this case, the number of features just slowly increases and remains almost constant after the gaze fixation number 30.

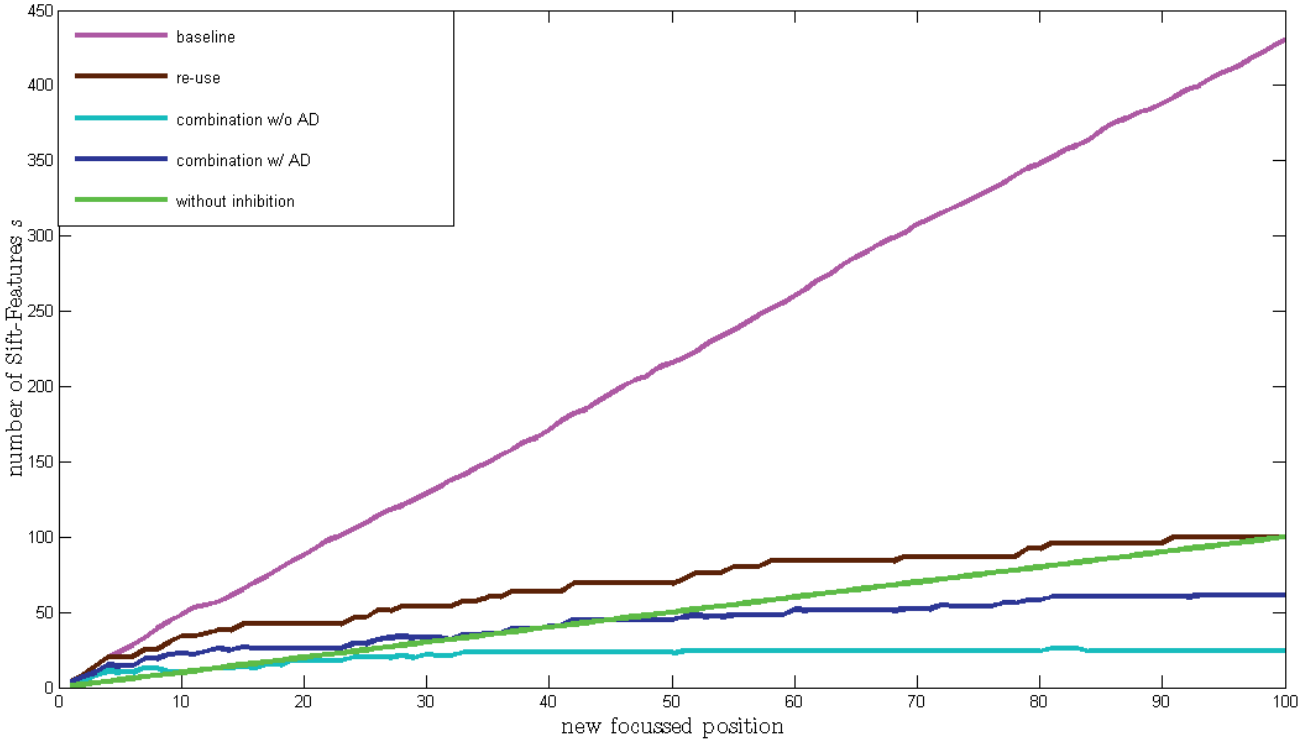

Figure 4.13.: The number of *Sift-Features* as a function of the number of gaze fixations during training. The different plots depict the results of the system when different pruning strategies are applied.

Table 4.1 illustrates the number of the overall learned models and learned mouth and hand models for each strategy. Additionally, the corresponding number of *Sift-Features* are depicted. As can be seen a comparison of the ratio between the number of models and features of the baseline and the 're-use' strategy does not show a significant difference. In fact, the 're-use' strategy reduces the number of view models but this only implicitly influences the reduction of features. In contrast, the 'combination' strategy tries to minimize the number of features, where an activity detection leads to a larger number. This is due to the fact that the system retains principle features of the learned models. The non-linearity

applied during AD increases the specificity of the individual feature responses. This in turn decreases the pair-wise similarity between the feature maps. A subsequent pruning consequently removes only such features that significantly overlap with other features. This is in contrast to not using AD, where smaller overlaps are sufficient to induce a feature pruning. This in turn leads to a false pruning of features and yields to classifiers that lose their discrimination ability. Hence, the number of models significantly decreases since this strategy misses the learning of novel objects. In contrast to this, the incorporation of AD yields a larger number of features and models. This, however, might be beneficial with respect to the recognition performance of the system. This is evaluated in the next part.

strategy	# models f_j	# Sift s_i	# f_j mouths	# f_j hands
baseline	100	430	4	11
without inhibition	100	100	4	11
re-use f_i	25	100	2	4
combination w/ AD	29	61	2	6
combination w/o AD	18	24	3	2

Table 4.1.: Number of overall extracted visual view models f_j and according number of Sift-Feature s after the training phase.

For all pruning strategies, the number of hand and mouth models is decreased. This illustrates that the system dynamically adapts the number to re-use classifiers during the training phase. Objects like hands that reveal a large variability in their appearance are described accordingly with several models. In contrast, simple objects like mouths are described with considerable fewer models. The incorporation of AD yields a larger number of hand models. This is due to the fact that this strategy captures the variability of object appearances in contrast to not using AD.

Testing Phase

In the following, we investigate the discriminative power as well as the generalization capabilities of the system. The evaluation of the learned visual models f_j is performed on a set of images that show objects in the image center. These objects have not been seen before by the system and are obtained from the same saliency computation used in 4.3.3. The obtained objects are manually divided into three groups: lips, hands, and other object appearances (see Fig. 4.14). The same classification is manually conducted for the learned visual view models, i.e.

(a) samples: lips (b) test samples: hands (c) test samples: e.g. bracelet, shirt

Figure 4.14.: Test data set: The left figure depicts samples for lips (a), the middle one for hands (b) and the right one samples e.g. for a shirt and parts of a background (c). The test samples are extracted in a predefined step with a bottom-up saliency mechanism. The data set contains always 20 samples for each appearance.

two classes are build for lip and remaining view models as well for hand models. Remaining models comprise object representations e.g. parts of a shirt or parts of the table.

It is desirable to obtain relevant view models for lips and hands that give specific responses to learned visual information and also detect their absence. Therefore, the system's generalization capabilities are depicted with the true positive *tp* and the true negative error rate *tn*. The *tp* measures the activities of lip and hand view models in the presence of appropriate objects. In contrast, the *tn* gives evidences about the response behavior of the remaining view models in the presence of objects that differ from lips and hands. Further, it is reasonable to learn lip and hand view models that suppress their activities in the absence of the learned information: This is evaluated in terms of the false negative error *fn*. Such a discriminative response behavior is also desirable for the remaining models and means to learn models that show a low response in the presence of objects such as lips and hands. This response behavior is analyzed in terms of the false positive error rate *fp*.

The generalization capabilities and discrimination power are analyzed according to the average center activities a'_{j*} of the extracted visual models f_j in reference to the presence and absence of objects. The performance characteristics are depicted in Table 4.2 and 4.3 which describe the average responses over all test samples for the respective maximum responding classifiers. As can be seen in Table 4.2 all strategies show a high degree of generalization capabilities illustrated by the *tp* and *tn* rate, with the exception of the *combination w/o AD* . This is due to the fact that lip view models partly have been incorrectly replaced by

background filters and consequently show a lower response in the presence of lips. The sensitivity to learned object information is lost which is demonstrated in the low *tp* error. This result is presented repeatedly in the analysis of the discrimination ability. The lack of using AD indicates that the selectivity for lip and for remaining view models are no longer given. The performance is similar to learned view models, which miss the inhibition of peripheral observations. However, the incorporation with AD demonstrates the highest discrimination capability. This is due to the fact that principle filters remain during the pruning. A further reason for the high performance could be accounted by the larger amount of preserved filters compared to the strategy without using AD. This ratio is not confirmed in comparison to the other strategies, i.e. the amount of the filter is not crucial, but rather the selection of most representative filters.

The recognition performance for learned hand view models is illustrated in Table 4.3. The generalization ability of the different strategies is similar to that of learned lip view models with exception of the *combination w/o AD*. This shows a similar *tp* rate compared to the other approaches. This is due to the fact that hand objects are characterized by more structural variance which may lead to similarities in overlap with objects. In principle, the activity characteristics of hand view models can be depicted by the *fn* rate, which is on average higher in all strategies in comparison to learned lip view models. The discrimination performance is dominated by remaining view models without incorporation of AD that demonstrate the lack of selectivity due to incorrect filter pruning. The *fn* shows that the learned view models do not retain any object selectivity that is similar to the models without benefit from inhibition. Consequently, in both strategies the learned view models continuously respond. This is also confirmed by the *fn* rate of those models that lack of an inhibition and the AD that lead to low discrimination capabilities.

model	strategy	t_p	t_n	f_p	f_n
	without inhibition	0.98	0.97	**0.89**	**0.77**
	baseline	0.93	0.87	0.68	0.38
lips	re-use f_i	0.92	0.81	0.68	0.31
	combination w/ AD	1.00	0.84	**0.60**	**0.29**
	combination w/o AD	**0.66**	0.95	**0.77**	**0.62**

Table 4.2.: Classification performance of learned lip view models by means of the maximum responding filters in average.

model	strategy	t_p	t_n	f_p	f_n
	without inhibition	0.92	0.97	**0.86**	**0.80**
	baseline	0.87	0.88	0.70	0.59
hands	re-use f_i	0.85	0.79	0.69	0.53
	combination w/ AD	0.81	0.83	**0.68**	**0.54**
	combination w/o AD	**0.81**	0.92	**0.92**	**0.64**

Table 4.3.: Classification performance of learned hand models by means of the maximum responding filters in average.

4.4. Summary

In this Chapter, a method for visual model learning during tracking that enables an incremental learning of multiple object classifiers was proposed. The approach allows a model learning from scratch and is initialized by a visual bottom-up attention model. In addition to the learning of multiple classifiers, an integration mechanism is introduced to combine different object views. A classifier is defined as a linear combination of positive and negative samples that improves the discrimination capabilities during object learning. A spatio-temporal continuity constraint is used as supervision signal for integrating different views into a classifier. During scene exploration, the acquisition of object knowledge results in a linear increase of the number of Sift-Filters. Therefore, a pruning strategy removes appropriate redundant filters to avoid problems in storage capacities. The removal is based on a pair-wise similarity measurement of filter responses. The incorporation of a non-linearity during activity detection yields a maintenance of discrimination capabilities and only representative filters are kept by the computational model.

The model developed so far has been based only on the visual modality and accordingly generates responses to learned objects. In detail, potential object locations are computed by learned object-specific filters and are usable for an artificial agent to generate saccades. However, a generation of voluntary saccades to multimodal aspects requires the involvement of auditory characteristics and a classification of filter responses by acoustic features. This classification may implemented by a weighting scheme so that only relevant object filters are active in the presence of certain audio signals. The implementation of an appropriate weighting scheme is the focus of the next Chapter.

5. Learning Voluntary Gazing towards Multimodal Events

As we have seen in Chapter 3, children are able to develop a visual preference during their first year of life with respect to multimodal scene objects. They do this by an associative learning of audiovisual object properties based on which auditory characteristics can be subsequently used during visual discrimination tasks. The implementation of such a gazing behavior may enable artificial agents to learn from audiovisual scene aspects and use them for a gazing strategy towards multimodal aspects. This chapter focuses on a method for artificial vision systems by which acoustic scene characteristics can guide voluntary gazing by means of a top-down integration. To do so, a method is proposed that enables an unsupervised associative learning of relevant audiovisual object properties. Secondly, learned object associations are used as a weighting scheme to bias the visual filtering process during attention shifts towards multimodal objects.

5.1. State of the Art

The learning of shared audiovisual object characteristics bears many challenges. One challenge relies in the selection of appropriate features that enable a reliable measurement of audiovisual commonalities by an artificial agent. Many approaches aim at the learning of audiovisual commonalities to perform a speaker localization. Thereby, they assume the existence of scene objects such as lips and speech (Slaney and Covell, 2000; Hershey and Movellan, 1999). These approaches generally ignore important features such as object trajectories generated by tutor interactions as demonstrated by (Delaherche and Chetouani, 2010). Moreover, the main assumption about speech, lips, or hand presence in the scene is only valid in very specific experimental setups and relaxing this assumption would necessarily lead to an autonomous learning of objects present in the scene. This aspect also refers to the learning of audiovisual object properties that needs to be acquired during demonstration, i.e. the visual and auditory stream needs to be processed online to obtain commonalities.

When processing online audiovisual scene information, the processing streams

of the different modalities need to be adjusted in time to allow temporal comparisons. In detail, each processing step such as motion detection or the detection of auditory energy creates different latencies. These must be compensated in order to initialize a learning process that benefits from synchrony in different modalities. This means a learning from audiovisual object properties presumes time compensated filter responses to detect coincidences. Additionally, an audiovisual object can be demonstrated in various ways by a tutor. For example, a demonstrator marks his hand as particular relevant to the robot by presenting it with a slight knocking and with a high motion intensity that results from hand movements. This means an artificial agent has to cope with such 'slight-high' correlated occurrences during object learning (Rolf et al., 2009). Another possibility may rely in a design of a coincidence calculation that abstracts from such 'slight-high' attributes, so that a robot learns whether overlaps of global audiovisual activities exist by means of *acoustic packages* (Schillingmann et al., 2009). This means motion activities of a mouth can be measured in overlap with the presence of speech signals. Similarly, hand movements may overlap with a knocking sound. But mouth movements feature a lower motion intensity than a hand gesture and such position-based approaches may lead to an overshadowing of relevant object movements which hence would not be accessible for object learning.

A learning of coherent audiovisual activities may be carried out via the measurement of motion activities that exactly match the occurrences of accompanied acoustic signals. Unfortunately, this matching criterion is not sufficient to learn audiovisual object presentations such as speech-lips or knocking-hand associations, since they do not always occur in a synchronous way. More precisely, lip movements may appear before speech is actually produced which hence may be audible later for a robot. Therefore, it is of key importance to design a correlation criterion that deals with time shifts, so that asynchronous coherences are detectable (Lee and Ebrahimi, 2011; El-Sallam and Mian, 2011).

In the following, a multimodal attention system is presented which allows a robot to attend audiovisual aspects by using acoustic scene properties. The model enables the robot to learn associations by means of audiovisual features that cope with time shifts. The learned associations support a weighting scheme in terms of a classifier and allows an autonomous learning system to increase its visual attention by means of acoustic properties. The proposed method is based on an associative learning of averaged joint onset activities of visual and auditory features that are derived during object tracking. The overlap measurement is done online, i.e. during object demonstration correspondences between onsets are detected, learned, and finally used to modulate the visual filtering process.

5.2. A Computational Multimodal Attention System

This section gives an overview of a multimodal attention system that may equip artificial agents with a gaze control strategy towards multimodal aspects. Subsequently, an example is shown in which the problem of asynchronous event is described. Additionally, the feature calculation in terms of onset detection is explained. These onsets are next used in an associative learning step to build audiovisual object representations. Finally, the benefits of the resulting associations are demonstrated within the overall system for gaze control. Thereby, the systems' performance is evaluated by means of its discrimination capabilities during associative learning. The generation of voluntary gazing behavior towards multimodal scene aspects is evaluated by using learned object associations.

5.2.1. System Overview

The presented system provides a method for learning voluntary gazing strategy towards multimodal scene aspects. As depicted in Figure 5.1 the method is divided into a training phase and a testing phase. The training phase describes the time interval in which the system learns visual object classifiers $\mathbb{F}_K$ as well as multimodal coherences in terms of associations between visual and auditory features. Subsequently, the testing phase describes the application of the learned associations during the visual filtering process. The result of the classification of the audio signal is used to select appropriate coefficients $c_{a,v}$ from the learned multimodal associations and hence serves as a top-down signal to configure the visual filtering process. Crucial to the testing phase is the combination of several responses $\boldsymbol{Y}_{j^*}$ of visual classifiers by means of their weighting information $c_{a,v}$. The combination of biased model responses allows the calculation of a multimodal saliency map sal_{av} that shows potential object locations which were previously learned in association with the acoustic signal.

5.2.2. Feature Extraction

Figure 5.2 shows an example of tracked lips and the corresponding audio signal. It should motivate the selection of appropriate features for a subsequent learning of multimodal associations. Demonstrated is an image sequence that comprises lip movements (V) and the corresponding speech signal (A). The current gaze fixation of the system is marked by a green dot. For the learning of associations between object motion activities and their acoustic properties, motion dynamics

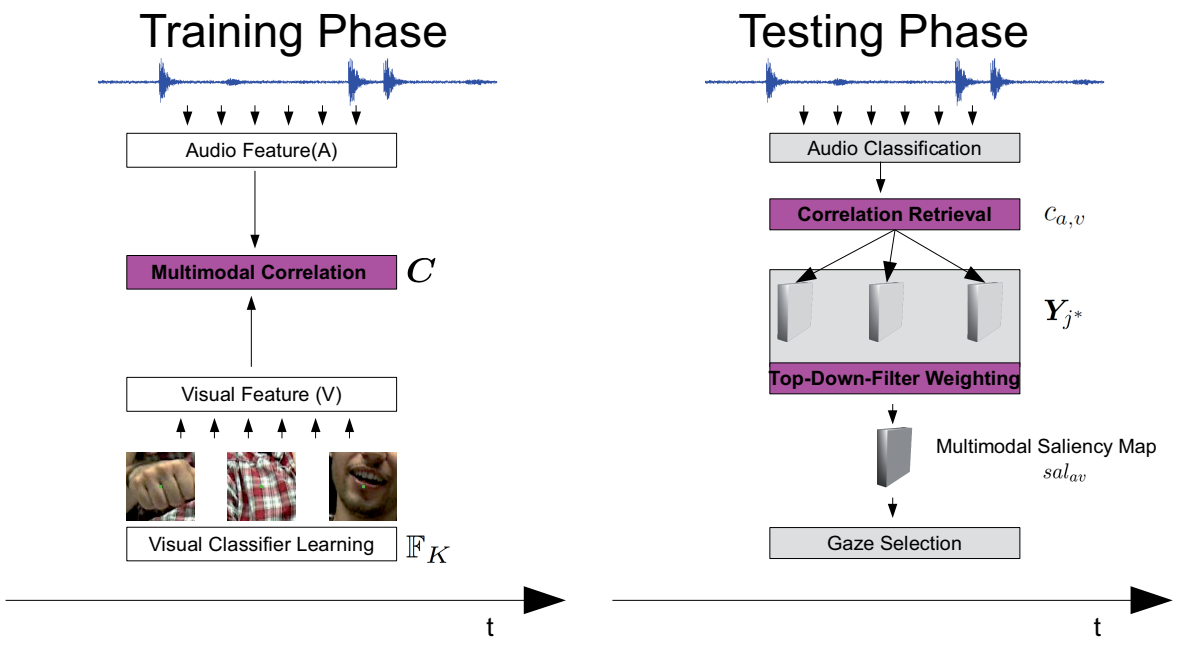

Figure 5.1.: Learning of a multimodal correlation $\boldsymbol{C}$ during object tracking and a computation of an appropriate multimodal saliency map. This saliency map is the weighted sum of visual classifier responses $\boldsymbol{Y}_{j^*}$ with corresponding association coefficients $c_{a,v}$.

can be measured in many ways. On the one hand, such dynamics can be carried out by local motion changes derived from the scene, e.g. such as the ones resulting from the opening and the closing of a mouth. For this, motion can be calculated from local object characteristics. The incorporation of such position-based information may result in motion activities that are caused by the rotation of objects instead of motion produced by the object itself.

In the example sequence, this is the rotation of the head that may shadow motion activities resulting from lip movements. Therefore, the motion extraction from local regions must be conducted invariant of rotations to be ultimately used for a correlation analysis. A further possibility to obtain motion dynamics consists in the exploitation of the object trajectory that may be derived from the tracking signal. The object trajectory is calculated as one pixel position and allows a measurement invariant from positions of neighborhood pixels. This is an advantage over the local motion estimation that is based solely on local neighborhood characteristics.

Figure 5.3 shows an example of the detection of local motion characteristics and the object trajectory that are extracted from the example sequence of Fig. 5.2. In addition, the envelope am of the audio signal $a(t)$ is shown. The envelope am of the audio stream is initially computed by means of its amount with $am = |a|$.

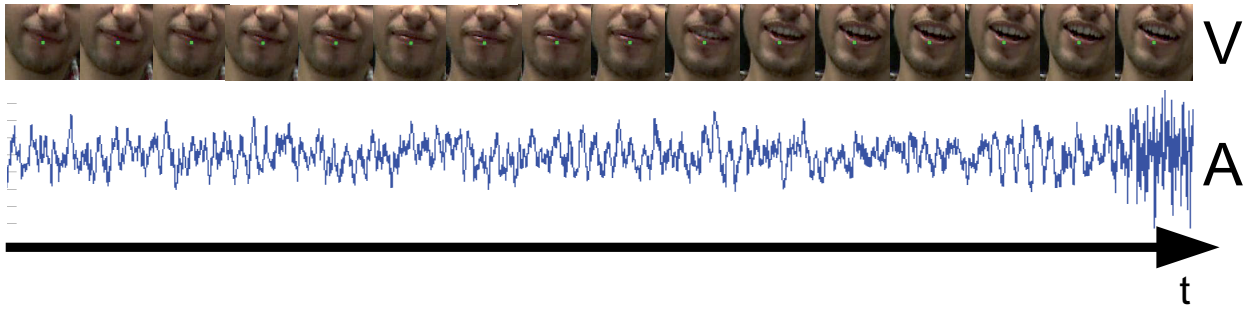

Figure 5.2.: An image sequence that shows lips (V) and the accompanied speech signal (A).

Then the signal is further low-pass filtered with a cut-off frequency of 150 Hz and resampled to 300 Hz. Afterward, the signal is low-pass filtered with a cut-off frequency of 10 Hz to extract relevant gradients of the signal. The extraction of local motion activities over time is denoted with ms and is based on the difference of local orientation information (see Eq. 5.1) of the current object fixation.

$$ms(t) = \sum |\boldsymbol{s}_{j0}(t-1) - \boldsymbol{s}_{j0}(t)| \tag{5.1}$$

The feature vectors $\boldsymbol{s}_{i0}$ describe the orientation histogram of the center object region by a *Sift-Feature* (Lowe, 1999). In contrast to the difference in orientation information, the object trajectory is computed by the difference of relative object positions $p(x, y)$ and is denoted with mt:

$$mt(t) = \sum |p(x,y)(t-1) - p(x,y)(t)| \tag{5.2}$$

This feature exhibits a high value in the case of significant object movements and small values in case of slight movement productions. As shown in Figure 5.3, all three features exhibit different characteristics. Particularly important are the feature gradients which can be exploited to measure onsets. However, the example shows that significant changes in the feature characteristics are typically time-shifted and hence less useful for detecting synchronous events. In addition, motion dynamics of object trajectories compared to changes in local orientations lack a normalization step, i.e. the values are not standardized. This also applies to the auditory signal. A normalization step is important to determine whether a significant overlap between visual and auditory gradients exists or not. For example, significant mouth movements may basically produce less motion dynamics than irrelevant little hand movements. In case of a missing normalization, this may lead to a learning of associations between hand movements and acoustic signals such as speech, although hand movements are not causing the production of speech. Therefore, for an associative learning of audiovisual object properties

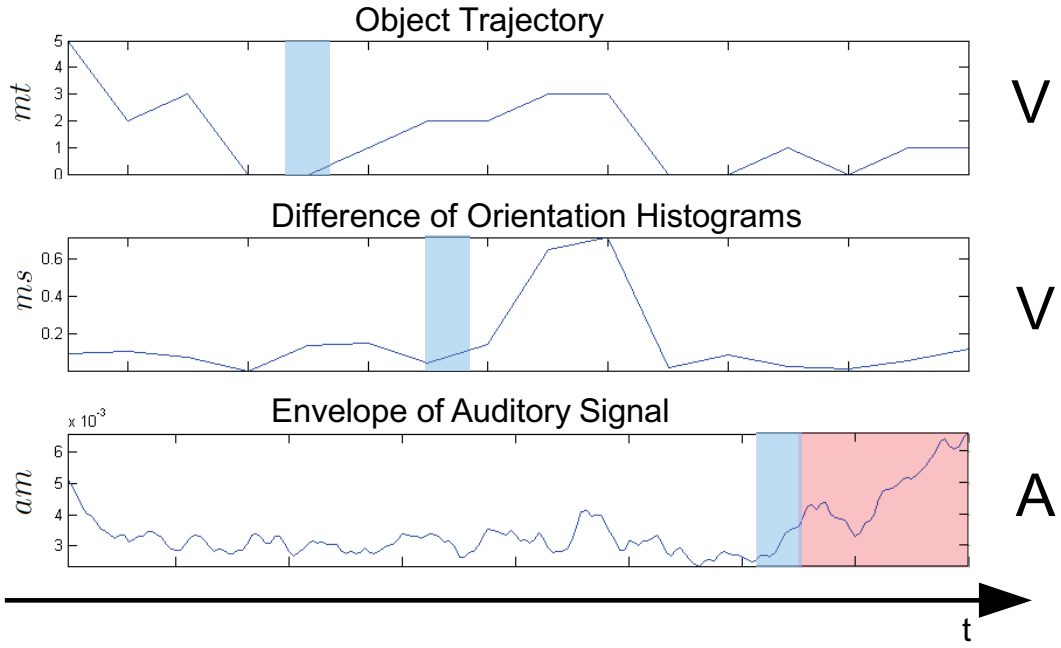

Figure 5.3.: Selected features for the visual domain (V) and auditory domain (A). Blue regions mark significant gradients of the visual features and the auditory feature. The red region marks the entering of the speech signal.

it is necessary to prepare the learning step such that small lip movements events are detected by the system and distinguished from rather unimportant motion events.

The shown preparations for an associative object learning are therefore divided in three ways, which are now described in detail. The first aspect relates to the extraction of relevant features, which may be the detection of significant gradients of the signal. This may be provided by information of onset occurrences of object motion and associated onset activities of the acoustic signal. The second aspect concerns the asynchronous occurrences of such onset activities in both modalities, which may give an important evidence to find an appropriate coincidence measurement. One possibility may rely on exploiting the shown asynchronous occurrences. For this purpose, it is useful to artificially delay onsets in their response behavior to cope with the shown asynchrony. Such delays of onsets may be based on normalized onsets and may be implemented by a preceded maxima detection. In summary, there are three steps to extract onsets from the signal: Onset-Detection, maxima-detection and a step for delaying significant onset maxima.

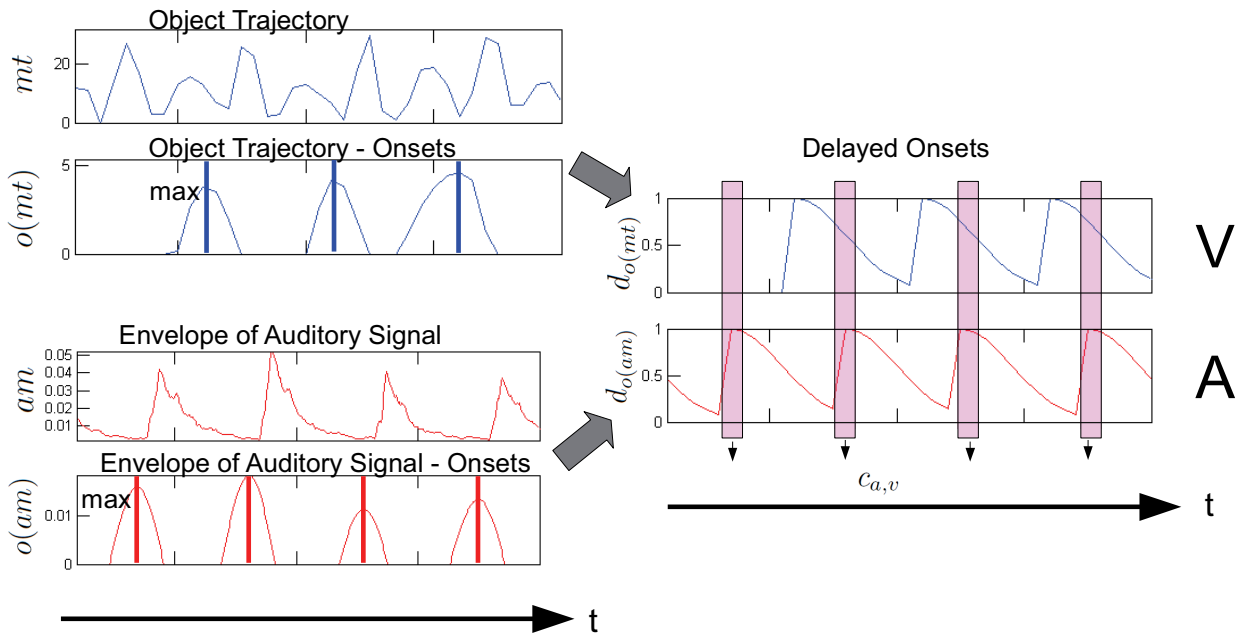

Figure 5.4.: Both, the visual (blue) and auditory features (red) are processed with the convolution operator according to Eq. 5.3 to obtain onsets. The maxima of them are marked and their responses are delayed with a half of the Gaussian operator σ. Plotted are the overlaps between the delayed onset signals, which can serve as a measurement of a correlation in terms of joint onset activities (pink).

5.2.3. Onset Detection

An example of an onset-based feature extraction is shown in Figure 5.4. The example shows motion characteristics of a knocking hand as well as related onsets detected in the acoustic domain. The detection of onsets involve the extraction of relevant gradients of the visual and auditory feature, i.e. the extraction of the beginning of motion dynamics and the beginning of knocking signals. The onsets $o(t)$ can be calculated using the gradients of a signal $x(t)$, for example by the derivative of the signal and are calculated in the following with (Wang and Brown, 2006):

$$o(t) = -G_o'(t,\sigma) * x \text{ with } G_o'(t,\sigma) = \frac{-t}{\pi\sigma^2} \cdot \exp\left(-\frac{t^2}{2\pi\sigma^2}\right). \tag{5.3}$$

$*$ denotes the convolution operator and $G_o'(t,\sigma)$ the first deviation of a Gaussian function. In the example, the convolved signal shows possible onset candidates $(o(am), o(mt))$ that are characterized by a smooth shape. However, the filter responses are still not prepared for a measurement of joint onset activities, since they are not normalized to a unit norm. To do this, maxima peaks are computed

based on the onsets within a time window over three signal points of $o(t)$. The resulted events are not sufficient to measure overlaps for longer time intervals. Therefore, these single peaks are delayed by a Gaussian function during the filtering. The resulting onsets are denoted in the following with d_o. The resulting latencies from different convolution steps are continuously compensated in order to enable an overlap measurement of synchronized filter responses.

5.2.4. Associative Learning and Top-Down Filter Weighting

The learning of associations between auditory and visual features of demonstrated objects is performed independently of the learning of visual models (see Chapter 4). One possibility for an associative learning step consists in the learning of overlaps between onsets of both features, i.e. if both events closely follow each other than the events seems to be associated and are learned by the system. The investigated learning of overlaps assumes that onsets of the auditory domain are linked to onsets derived from visual scene characteristics and finally form a multimodal object representation. Accordingly, a non-represented onset in the auditory domain can be used as a learning signal by which association weights can be adapted.

The classification of acoustic onsets is manually done (see Appendix A), i.e. acoustic properties such as speech or knocking are always correctly recognized in the learning loop by the system. Of course, the classification behavior during auditory processing may decisively influence the associative learning. But the focus of this section refers to gaze control and visual configuration capabilities by assuming a corrects filtering of auditory characteristics.

Training Phase

First of all, the learning of associations is done in terms of a weighting scheme that is denoted with $\boldsymbol{C} \in \mathbb{R}^{A \times V}$, where A is the number of acoustic classes and V corresponds to the number of visual models. Thereby, an element $\hat{c}_{a,v}$ corresponds to a correlation between acoustic delayed onsets $d_{o(am)}$ as well as onsets $d_{o(ms)}$ or $d_{o(mt)}$ derived from visual features. The measurement corresponds to the product of both onsets. Since overlaps can vary from joint onset appearances, it is useful to measure the overlap by means of their averaged frequencies with:

$$\begin{aligned} c_{a,v}(t) &= c_{a,v}(t-1) + \frac{1}{q} \cdot [\hat{c}_{a,v}(t) - c_{a,v}(t-1)] \\ \hat{c}_{a,v}(t) &= d_{o(am)}(t) \cdot d_{o(*)}(t) \text{ with } * \in \{ms, mt\} \\ \text{if } d_{o(am)}(t) &= 1. \end{aligned} \tag{5.4}$$

During visual model learning, the weighting scheme is gradually extended for each object K. The averaged frequency $c_{a,v}$ is updated at that time at which a significant acoustic onset $d_{o(am)}$ is appearing, where q denotes how often $c_{a,v}$ has been updated. This means that at the beginning of training (q=1), the actual correlation $\hat{c}_{a,v}$ has a large influence, whereas the influence decays later in time ($q > 1$). By doing so, the learned correlation $c_{a,v}$ represents the mean of all instantaneous correlations $\hat{c}_{a,v}$ which have been observed so far. After each saccade, the calculation is restarted and the associative learning begins again. The weighting scheme is weakened when auditory onset activities are detected and visual onsets are missing for this event.

Testing Phase

The resulted weighting scheme $\boldsymbol{C}$ configures the gazing behavior of the system towards multimodal aspects. Here, association coefficients are used for biasing the visual filtering mechanism, i.e. the learned coefficient weights bias the corresponding responses of visual classifiers of the system. For example, the presence of signals such as speech or knocking can trigger classifiers for lips or hands. In the testing phase, the learned weighting scheme is combined with corresponding classifier responses, so that a saliency map *sal* is created in which potential multimodal object positions pop out. This weighting step enables the system to perform a voluntary gazing behavior by means of a top-down integration of acoustic properties. The determination of a gaze location towards multimodal objects is computed by a winner takes all mechanism. It extracts the most salient position and defines the focus of attention. The calculation of this gaze position $p(x, y)_{av}$ is computed by means of:

$$p(x, y)_{av} = \max_{(x,y)} sal_{av} \text{ with } sal_{av} = \sum_{i=1}^{M} c_{a,v} \cdot \boldsymbol{Y}_{v_{i^*}} \cdot \boldsymbol{Y}_{a^*} \tag{5.5}$$

The saliency map *sal* combines the responses of the auditory and visual classifiers by a sum weighted with the association coefficients $c_{a,v}$, where $i = 1 \ldots M$ corresponds to the $i - th$ visual model. The overall response $\boldsymbol{Y}_{v_{i^*}}$ corresponds to the responses of the visual classifiers $\mathbb{F}_K$ and $\boldsymbol{Y}_{a^*}$ corresponds to the response behavior of the acoustic classifier. For the classification of acoustic onset activities it is always assumed that $\boldsymbol{Y}_{a^*} = 1$.

5.3. Evaluation

The evaluation for generating saccades towards multimodal aspects is done in two phases. In the first evaluation phase, the discrimination ability of learned

associations based on onset overlaps is analyzed. More precisely, it is of interest to what extent overlaps between visual and auditory onsets are accessible and can be learned by the system. For this, three different objects are tracked and their audiovisual correlations are calculated according to Eq. 5.4.

In a second evaluation step, the overall gazing behavior of the proposed multimodal attention system is analyzed and compared to a reactive attention model (Itti et al., 2003). For this, visual models are learned during tracking according to Chapter 4. During the learning of visual scene aspects, the system aligns audiovisual onsets. Subsequently, a testing phase depicts the gazing behavior towards multimodal aspects by integrating top-down class information such as speech or knocking. To do this, the systems' gazing locations are computed in the presence of different acoustic classes. The learning of object specific models includes an extension of the architecture that is described in Section 4.3.1 and takes object dynamics into account. In detail, new object views are added as soon as object models loose their validity during tracking.

The learning of visual models is based on two different strategies and evaluated in the gazing process. The 'baseline' strategy includes the learning of models without fusing them. The second method corresponds to the 'combination w/AD' strategy and includes the reuse of previously learned visual information as well as the pruning of redundant *Sift-Features*. The model learning with the 'combination w/AD' strategy is carried out to analyze how the removal of redundant object information affects the gazing performance of the system and how the integration of auditory top-down information is still helpful for generating voluntary saccades.

Additionally, it is assessed whether extracted audiovisual object associations are generally useful for a weighting step during the process of visual filtering. For this, the gazing performance is compared to a gazing strategy that does not benefit from learning of relevant associations. Here, the systems' visual response behavior is weighted equally by not associating the visual models with acoustic characteristics. Furthermore, the 'best case' is examined, i.e. the audiovisual associations are set manually for visual models like hands and lips so that they are only biased by means of signals of knocking or speech. For example, the system learns two hand models, the corresponding weights are equipped with values $c_{a,v} = 0.5$ to associate the models to the acoustic class knocking. For three hand models that would result in correlation coefficients $c_{a,v} = 0.33$ so that the sum of them is 1.

Associative Learning by Means of Joint Onset Activities

In the first evaluation phase, three objects are focused by the system and analyzed whether an overlap of onsets in the presence of different acoustic classes is

accessible by the system. The gaze focus of the system is manually set to objects such as the mouth, hand, and to the t-shirt of a person and tracked for 2000 frames. Audiovisual associations by means of joint onset activities are calculated accordingly. The acoustic scene is held constant only varying the observed objects. This results in 31 onsets for speech and in 86 onsets for the knocking signal. These onsets overlap with onsets derived from visual features.

For the evaluation, a video sequence is used that shows a person who is knocking repeatedly on the table. After knocking, the person verbally aligns the number of taps. The video sequence is recorded with 25 fps and visual features are extracted according to Eq. 5.1 and 5.2 with $ms > 0.5$. The difference of temporal orientation information of the object is computed with a 50x50 image patch of the current focused objects that is used to calculate local motion energy.

For the extraction of significant onsets in the audio stream a, the signal is processed with 16 kHz. The envelope of the auditory signal is computed according to Section 5.2.2. It is further convolved with the first derivative of a Gaussian function according to Eq. 5.3. For simplicity, the onset calculation is made in advance and resulted onsets are threshold with t_a= 0.0017 so that $o(am_s) > t_a$ are kept for associative learning. The threshold is defined manually and corresponds to typical onsets of the speech signal.

During object tracking, both visual features mt and ms are computed and convolved according to Eq. 5.3. A short convolution window with σ= 0.05 is used to detect spontaneous gradients. This corresponds to a time window of 403 ms for the processing of acoustic information and to a time window of 440 ms for the detection of onsets of motion dynamics derived from visual features.

In Table 5.1, the resulting association matrices $\boldsymbol{C}$ are shown for three different object models and 2 acoustic classes. The upper three rows describe the association coefficients $\boldsymbol{C_{mt}}$ by using onsets derived from the object trajectory. The lower three rows describe the matrix $\boldsymbol{C_{ms}}$ resulting from onsets in local motion activities of the currently tracked object.

It is evident that the use of object trajectories results in overlaps between all object-sound pairs, where the hand-knocking pair differs from all others. The difference lies in the discriminative behavior of these associations in comparison to hand-speech associations. The corresponding association coefficients show a higher value in case of a hand-knocking occurrence. In all other object-sound pairs, the association coefficients are more or less equally distributed which indicates that onsets derived from the trajectory are not appropriate for association learning. The uniform distribution is justified by the fact that many onsets are continuously produced by the object trajectory. For example, onsets may be a result of the persons' body movements that cause a correlation with the acoustic

object	feature	speech $d_o(am_s)$	knocking $d_o(am_k)$
mouth	$d_o(mt)$	0.46	0.50
hand	$d_o(mt)$	**0.24**	**0.57**
else	$d_o(mt)$	0.52	0.58
mouth	$d_o(ms)$	0.04	**0**
hand	$d_o(ms)$	**0**	0.29
else	$d_o(ms)$	**0**	**0**

Table 5.1.: Learned object associations in terms of joint onset activities.

signals that are less object specific. For a machine learning system, the resulting correlation of audiovisual events is not useful, since mouth-speech associations may be overshadowed by other associations which prevents gazing towards lips in the presence of speech signals.

Therefore, it is useful to analyze the resulting associations using local motion dynamics. In contrast to the use of object trajectories, the use of local motion activity shows a different association behavior with the auditory domain. Object-specific associations such as t-shirt-speech or mouth-knocking reveal a zero correlation, whereas associations such as mouth-speech or hand-knocking show a positive correlation. The positive mouth-speech correlation may be justified by a relevant overlap of speech onsets and onsets of local motion that is caused by lip movements. Particularly, learned associations between the hand model and the knocking signal are conspicuously. Actually, very little motion energy should be produced in terms of difference of orientation histograms while tapping the hand on the table, since its shape is not particularly changing and therefore should result in less local motion activities.

However, the difference of orientation information of hand objects may be the result of changing background information of the currently observed object. For example, this may be the occurring of the table edge before the onset of knocking signal is entering. A further reason for the built association may rely in the production of local motion by forwards and backwards tilts of a hand during knocking activities and hence accordingly be close in time to relevant onset information of knocking signals. These two reasons are not sufficient to use o_{ms} as dominant motion detector for hand movements, since the system can also observe hands while producing knocking sound, that does not tilt. In addition to this, the change of background information is only accessible by the system, if the orientation histogram covers background information, i.e. information beyond the current tracked object. But this two assumptions are not always fulfilled in

a learning scenario and therefore such information is not always accessible by an artificial agent.

Both visual features show advantages and disadvantages for an associative learning step. The use of object trajectories results in a meaningful correlation with respect to hand-knocking associations, but also suffers from the fact that other multimodal associations are not discriminative. On the contrary, the local motion energy generated by the difference of orientation histograms shows a discriminative behavior for relevant speech-mouth associations. However, it can not be assumed that this feature is always reliably detected for hand movements during knocking. Therefore, it may be useful to exploit the benefits of both visual features in order to use them for an estimation of joint onset activities. Resulting correlation coefficients may be combined by a multiplication so that the result is an associative matrix with:

$$\boldsymbol{C_{mt/ms} = C_{mt} \cdot C_{ms}} \tag{5.6}$$

where $\cdot$ refers to an element-wise matrix multiplication.

Voluntary Gazing by Means of Top-Down Information

In the following, the learning method derived from section 5.3 is used to build multimodal object associations during the learning of visual object models. In a second step, the validity of learned associations is examined, i.e. to what extent top-down information such as speech or knocking is usable by the system to generate voluntary saccades towards multimodal events.

On the one hand, it is desirable to equip an artificial agent with a top-down gazing strategy, so that it make use of object-specific sound to direct gaze towards corresponding visual objects, e.g. listening to speech and start to move the gaze towards lips. On the other hand, it is also important that not lip-specific sound leads to a higher looking frequency towards objects different to lips. This gazing performance of the system comprises the generalization ability.

In contrast to this, an artificial agent should also be able to generate a little number of saccades towards objects, when the corresponding learned object-specific sound is absent (i.e. relying on the bottom-up default saliency process). For example when the system is listing to acoustic properties of a knocking hand, this should lead to less frequently looking times towards objects like lips. Additionally, acoustic properties such as speech should lead to less frequently looking times on other objects such as hands or the person's neck. This gazing behavior is measured in terms of the system's discrimination capabilities. For the evaluation, two acoustic classes are examined: speech and knocking and the corresponding targeting behavior towards multimodal objects of the system is analyzed.

	$\#\ \mathbb{F}/f$	$\#\ s$	$\#\ \mathbb{F}_{lips}/f/s$	$\#\ \mathbb{F}_{hands}/f/s$
baseline	15/21	78	1/2/7	2/5/21
combination w/AD	6/11	26	1/2/6	1/2/7

Table 5.2.: Resulted quantities of the visual model learning processes by means of the 'baseline' and 'combination w/AD' strategy.

Training Phase

In a training phase, the system explores a scene that shows a person that is talking and knocking. The system initially explores the scene by bottom-up saccades (Itti et al., 2003) and salient points are tracked for a specific time. The exploration phase starts with an observation on the person's face. A total number of 15 saccades are generated and are tracked for 100 frames. During this period, the system learns visual models and aligns audiovisual events by means of the proposed associative learning step (see Eq. 5.6). During learning the system extracts one lip model and two hand models.

Result

Table 5.2 shows the results after the learning phase and depicts characteristics of the model learning processes. The number of visual models $\mathbb{F}$, the number of integrated view models f, and the number of related stored *Sift-Features* s are depicted from the different learning methods. Additionally, the model quantities are separately shown for lip and hand models. Moreover, Figure 5.5 shows the extracted association matrix $\boldsymbol{C_{mt/ms}}$ that has been learned during the extraction of visual models by using the 'baseline' strategy.

Overall, the results show that the model learning via the 'combination w/AD learned' strategy represents the scene with fewer visual information than the 'baseline' strategy. This not only holds for the total number of visual models, but also for the number of model views and the number of *Sift-Features*. The removal of redundant information by using a pairwise comparison of the nonlinear response behavior of *Sift-Features* results in a reduction of quantities for visual lip and hand models. In particular, the re-use of models shows its effect with respect to the reduction of model views for hands.

The result of the associative learning step shows that objects such as the face and hands yield high association coefficients. This is a result from the combination of visual features to obtain audiovisual associations. Associations for tracked objects such as the arm or parts of the t-shirt lack of coefficients and may be

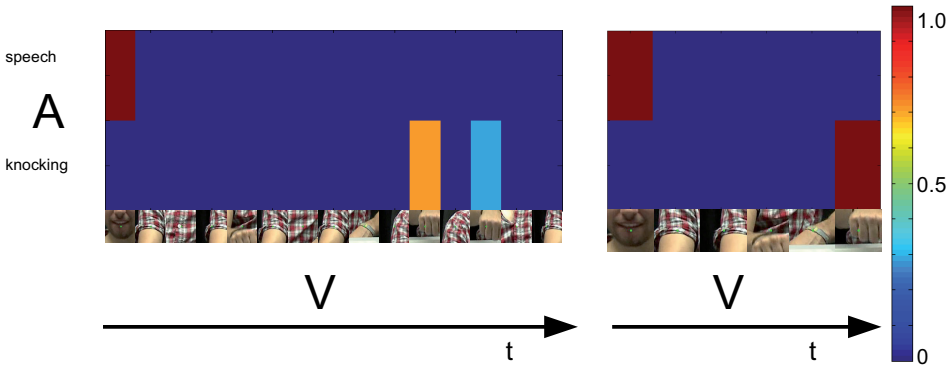

Figure 5.5.: Associative object learning during tracking. The correlation coefficients are normalized to 1. The left hand side shows the learned associations for the 'baseline' strategy. The right hand side depicts the learned associations by means of the 'combination w/AD' strategy.

understood by the missing of motion activity. However, the tracking of lips and hands by the system results into onset overlaps between significant motion onset and auditory characteristics such as speech and knocking onsets.

Testing Phase

In this phase, the generation of saccades towards multimodal aspects is analyzed. To do so, learned associations and their weighting in the processing of visual information are investigated. The classification behavior in terms of multimodal events is studied using the shift of visual attention by the system. For this, the system is confronted again with a scene that shows a person that is speaking and knocking. The occurrence of these two acoustic classes is manually annotated. The data set comprises 1500 frames which include 677 frames with knocking segments and 684 frames of speech segments. Again, the acoustic class 'noise' is not considered here, since an associative learning with objects has not been taken place. The generation of saccades by the system is based on Eq. 5.5, so that the presence of acoustic classes triggers the response behavior of learned visual models. The association coefficients of each acoustic class are normalized to 1 by dividing each coefficient with the overall sum of existing associations for this class. The result of this triggering process is a linear weighting of model responses by the corresponding acoustic class, i.e. a multimodal saliency map that offers locations to shift the focus of attention. The maximum is used as a candidate for refocusing visual attention in order to control the gaze towards audiovisual objects. The resulting foci of attention marks objects that are classified manually to measure the performance of the top-down gaze control.

model	strategy	lips	hands	others
	reactive gazing	**0.2109**	0.0354	**0.7537**
	baseline best case	1	0	0
speech	baseline learned	1	0	0
	baseline equal	0	0.1829	**0.8171**
	combination w/AD learned	**0.9808**	0	0.0192
	combination w/AD equal	**0.2773**	0.5605	0.1622

Table 5.3.: Classification performance in the presence of speech models by means of the maximum response. The gazing probabilities towards objects such as lips, hands and other objects are shown.

Result

Table 5.3 and Table 5.4 show the result of the visual gazing behavior by the system. The visual information processing is triggered by two acoustic classes: speech and knocking and the system's behavior is analyzed in terms of the gazing probabilities towards objects such as lips, hands, and other objects. The 'best-case' is missing for an associative objects-learning during the 'combination w/AD learned' strategy, since only one hand and one lip model is learned and correspondence coefficients are set to the value 1.

At first, the retrieval performance of objects such as lips is examined while the system is listening to speech. The gazing behavior produced by Itti et al. (2003) shows that a reactive gazing is less suitable for the recognition of multimodal objects. This is demonstrated by a less frequent gazing towards lips during the presence of speech. This may be reasoned by features such as flickr and motion that are integrated into the bottom-up attention model to generate saccades towards locations that are dominated by such motion dynamics. Since lips produce very little energy compared to body movements, it is therefore very unlikely that lips are focused and a reactive strategy leads to a frequent gazing towards other objects.

In contrast, the application of the learned correlation scheme during the learning of visual models by means of both 'baseline methods ('best case', 'learned')' results in a frequent looking towards lips. A similar adequate performance is also shown by the use of the 'combination w/AD learned' strategy, although redundant information is removed. The comparison to a uniform distribution of the correlation coefficients shows a reversal of the system performance. The speech signal is no longer effective to retrieve lip objects and hence the system retrieves other objects and also hands. Overall, the learned correlation schemes

offer a way to control visual attention to lips by an artificial agent. A performance comparison to a reactive gazing to the 'combination w/AD equal' strategy shows similar gazing behavior towards lips. In detail, a lack of object-specific learned associations leads to the decrease of the specificity of the classification behavior and shows similar performance such as resulting from a reactive gazing behavior.

The triggering of the visual filtering process by knocking signals shows a similar gazing performance such as it was observed by using learned speech-lip associations. Also here, an inversion of the gaze performance is shown when no object-specific associations are incorporated into the gazing strategy. However, the 'combination w/AD learned' strategy results in a lower classification performance compared to speech-lip models. But it clearly shows a better performance than the use of equally distributed associations. This may be a result of the pruning step, since it is allowed to merge similar positive samples. This may lead to a loss of detailed object knowledge of hands. Nevertheless, this strategy shows a better performance in conjunction with learned associations as compared to object gazing resulting from a bottom-up gazing mechanisms.

model	strategy	lips	hands	others
	reactive gazing	0	0.6720	0.3280
knocking	baseline best case	0	**0.9107**	0.0893
	baseline learned	0	**0.8697**	0.1303
	baseline equal	0	0.0937	**0.9063**
	combination w/AD learned	0	**0.7496**	0.2504
	combination w/AD equal	0.0132	0.6750	0.3119

Table 5.4.: Classification performance of knocking models by means of the maximum response. The gazing probabilities towards objects such as lips, hands and other objects are shown.

5.4. Summary

In this chapter, a method was proposed for learning voluntary gazing towards multimodal scene aspects that can be used by an artificial agent. The method is based on a configuration of the visual filtering process and allows a modeling of visual attention with respect to multimodal events independent from the location cue. In detail, learned object associations in terms of a weighting scheme allows a top-down integration of auditory characteristics so that audiovisual objects are focused irrespectively from an initially learned position.

The learning of voluntary gazing is divided into a learning phase and in a testing phase, where the learning phase corresponds to a bottom-up process and serves for the acquisition of audiovisual object associations. An association is computed by joint onset activities of different modalities. It has been investigated in onset activities derived from auditory characteristics with local object motion as well as object trajectories. The combination of both motion dynamics leads to the maintenance of their shown advantages during association learning with auditory scene characteristics. In particular, the discriminative properties of local object dynamics leads to a suppression of multimodal associations in the presence of objects that do not posses sound production capabilities.

During the testing phase, the trained weighting scheme was used to configure the response behavior of acquired visual scene knowledge in the presence of acoustic characteristics. The proposed multimodal attention model yield frequent attention shifts towards multimodal aspects in comparison to a reactive attention model. In particular, the gazing performance does not significantly decrease during top-down filter weighting by means of visual object knowledge that is removed from redundancies. The analysis of a reactive gaze control towards multimodal aspects shows that it is rather dominated by motion dynamics. Furthermore, it was observed that gazing strategies resulting from an equalized weighting scheme lead to a reversal of the system's performance. This aspect stresses the importance of an active configuration of the visual filtering process by means of learned object associations.

6. Summary

In this work, a machine learning method was introduced that allows to control the visual attention of an autonomous system towards audiovisual scene aspects. Underlying learning principles derived from infancy research served as inspiration to model such artificial gazing behavior. One principle comprises the initial visual reactive gazing behavior of infants that makes use of generic visual features rather than object knowledge during scene exploration. With increasing age, infants start to structure their reactive gazing behavior by means of the integration of auditory scene characteristics that may guide visual attention. In particular, dynamic motion characteristics as well as significant contrasts in auditory signals of objects play a key role in this respect. More precisely, infants start to control their visual attention by developing visual expectations in the presence of object-specific sounds (Richardson and Kirkham, 2004). Equipping a robot with such a gazing mechanism may enable it to classify objects by their sound and accordingly learn to control its visual attention to it. In this thesis, this principle of multimodal classification was implemented by an associative weighting scheme so that different learned visual aspects are accessible and biased by an artificial audition system. The integration of top-down knowledge in the visual filtering process by acoustics categories showed that the developed gazing strategy is able to produce gazing more frequently on multimodal scene aspects than a reactive visual attention model. A similar gazing performance was also achieved when the weighting of visual model information was learned on visual models created based on a pruning strategy of redundant scene knowledge.

A major focus of this thesis relied in the investigation of a method for an unsupervised acquisition of object knowledge that can be triggered from acoustic scene knowledge. The acquisition process takes place during the tracking of objects, where a spatio-temporal continuity constraint is used as supervision signal to combine different object views. The learning of object models is characterized by the inhibition of distractors that occur in the periphery of the visual field. The inhibition mechanism enables an increasing object specificity that was shown by improved discrimination capabilities of the system. However, the learning of visual object knowledge leads to a linear increase in the number of object features. For this, a method was proposed that enables an artificial agent to make use of already learned object knowledge in the acquisition process so that redundant object constituents are removed.

6.1. Conclusions

The proposed computational model for learning voluntary gazing is benefiting from audiovisual processing and shows a possibility to control the gaze invariant from learned object positions. The system is able to shift its attention towards multimodal objects even though a demonstrator is freely varying object locations. An initially used reactive gazing mechanism shows that little relevant multimodal information is accessible to the artificial agent. This stresses the importance of gradual learning of voluntary gazing by a top-down integration of additional sensory information. This may be advantageous for an artificial vision system that needs to cope with the challenge to form multimodal object representations with little sensory information. Moreover, the weighting of learned visual knowledge enables a relevant selection of them and thus may reduce time consuming filtering processes for artificial agents.

6.2. Suggestions for Future Research

In the proposed architecture, the duration of the training and testing phases was set manually. More precisely, the entering in the bottom-up and top-down process was externally regulated. Therefore, future research may investigate a flexible design of the duration of both processes to regulate the switching between them by the system itself. One possibility may be to model the system's training duration in terms of the quality of audiovisual objects representations during tracking. The gaze fixation policy may be inspired by findings from the Habituation-Paradigma. More precisely, if the model learning is stagnating for current gaze fixations that might be a criterion to terminate the training phase, since the system acquired adequate object knowledge and needs to develop novelty preference to other visual scene aspects. Such termination criterion may enable the system to regulate the bottom-up process so that new scene aspects are accessible. Consequently, the system would possibly be situated in a learning phase, and this aspect may lead to a shortened training duration and a faster knowledge acquisition.

The suggested method for an unsupervised acquisition of object knowledge is based on an inhibition mechanism that results into a linear increase of features. The implementation of a pruning strategy shows a reduction from a linear to a sub-linear increase. For an artificial agent, it would also make sense if the amount of stored features converges over time. One approach may comprise the examination of already stored positive object samples and their use as negative information for the inhibition process during model construction. Furthermore, it may be beneficial to build new visual scene knowledge by means of already stored

object knowledge, e.g. the selection of stored negative samples for an inhibition of a new learned object model of a current gazed visual aspect.

The proposed architecture was developed to examine the visual gazing behavior by means of acoustic scene properties. A correct classification of the auditory input stream was assumed, i.e. this scene knowledge was not acquired during object demonstration and was defined by external system knowledge given by the designer. In infancy research, there is also indication that infants benefit from synchronous audiovisual object properties to segment their acoustic environment (Hollich et al., 2005) that may bootstrap the unsupervised acquisition of auditory classifiers for an autonomous learning system. Particularly important is the rhythm that mainly influences the concept formation of language in infants development (Mehler et al., 1996). Rhythm may be a possible candidate to form auditory objects for artificial systems, since it abstracts from specific features such as the description in terms of frequency bands. The description of repetitive rhythmic patterns in various frequency bands and their occurrence in combination may constitute one possibility to form an object representation.

Such a description would be a way to create models for speech which may obtain improved generalization and discrimination properties in comparison to low level features given by frequency response activities. Such an extension of the proposed architecture may involve the development of acoustic expectations that are linked to visual object knowledge and may improve the learning performance. More precisely, learning processes of acoustic scene knowledge may be involved in the learning loop of voluntary gazing by means of inhibitory effects during building object associations. For example, the misclassification behavior of auditory models may be exploited by a negative weighting of according visual model response behaviors during top-down processing.

The proposed associative learning step in Section 5.2.4 is based on joint onset activities and showed that onset activities may also be produced from object trajectories as a result of the demonstrators' body movements. The extraction of relevant object trajectories and an appropriate alignment to acoustic properties may rely in the assumption of a multimodal rhythm by means of the correspondence of onset-offset transitions in both modalities. This assumption may enable an artificial agent to reduce produced activities from object trajectories. Additionally, the incorporation of onset-offset transitions may enable the learning of object associations invariant from parameters such as speaker's voice or according lip movements.

Additionally, the learning of object associations in the proposed architecture resulted in a similar correlation behavior of hands models during knocking activities. Similar to the categorization behavior of infants (Plunkett et al., 2008), it may be beneficial for an artificial system to form one object class by means of its

correlation. This may be a way to cope with dissimilarities within the same object instance to arranged multiple object instances in the object memory. Further future work to object memory refers to the shown lack of audiovisual associations. This could be used for a further pruning steps to remove irrelevant multimodal knowledge from object memory.

The shown suggestions are made to improve the gazing strategy for future work. However, infants do not only develop their visual capabilities based on auditory characteristics but also relay on other modalities. Sensory information such as odor or tactile object information influence the learning of objects and the perception of the environment. In particularly, infants use pointing gestures of interaction partners (Woodward and Guajardo, 2002) as learning signals for object discrimination. From the perspective of autonomous learning, it would be interesting to learn suitable models to structure such sensory information and use them for artificial attention modeling.

A. Appendix

A.1. Auditory Onset Classification

The model learning of acoustic classes is a minor focus of this thesis. Therefore the acoustic classes are manually defined and extracted from the auditory stream of the analyzed video sequences. The acoustic classes comprise segments such as *speech*, *knocking* and *noise*. The annotation of the sound sources is conducted with *PRAAT* (Boersma and Weenink, 2009). An example of an annotation is shown in Figure A.1, where the acoustic classes are labeled with an according string. Each time stamp of the sound source is aligned to a defined class such as *knock*, *noise* and *speech*. The annotation of the class *noise* is marked with an empty string. The manual annotation of the auditory stream serves as online

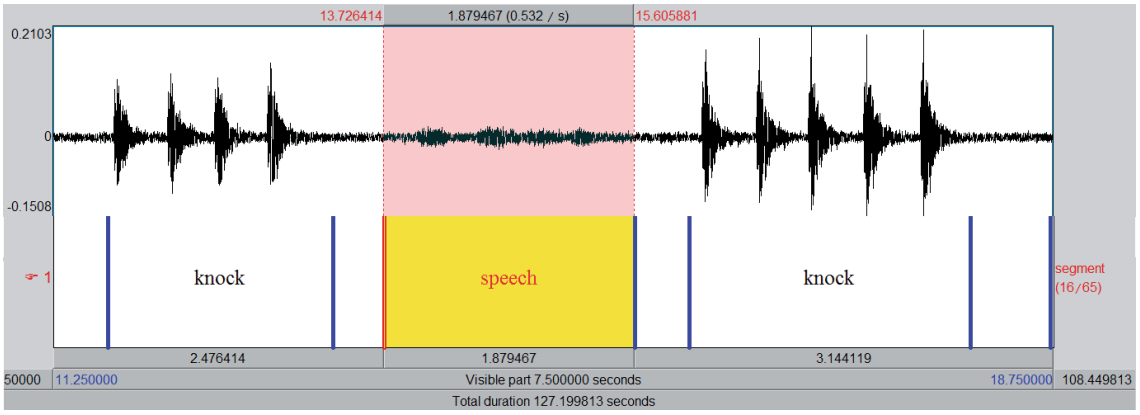

Figure A.1.: Hand annotation of acoustic classes with the software *PRAAT*. Segments for the classes *speech*,*noise* and *knocking* are labeled. The class *noise* is aligned with an empty text grid.

classification of detected onsets. The classified onsets are examined as feature for an associative learning with visual object characteristics.

Bibliography

Allport, A. (1989). *Visual attention.* In: Posner, M. I. (Eds.) Foundations of cognitive science, 631–682, Cambridge MIT Press.

Amso, D. and Johnson, S. (2008). Development of visual selection in 3-to 9-month-olds: Evidence from saccades to previously ignored locations. *Infancy: the official journal of the International Society on Infant Studies*, 13(6):675–686.

Aryananda, L. (2006). Attending to learn and learning to attend for a social robot. In *International Conference on Humanoid Robots*, pages 618–623. IEEE.

Aryananda, L. (2009). Learning to recognize familiar faces in the real world. In *International Conference on Robotics and Automation*, pages 1149–1154. IEEE.

Aslin, R. N. (1988). Anatomical constraints on oculomotor development: Implications for infant perception. *In The Minnesota Symposia on Child Psychology: Perceptual Development in Infancy*, pages 67–104.

Babenko, B., Yang, M.-H., and Belongie, S. (2009). Visual tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 983–990.

Bahrick, L. E. and Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, 36(2):190–201.

Bahrick, L. E., Lickliter, R., Castellanos, I., and Vaillant-Molina, M. (2010). Increasing task difficulty enhances effects of intersensory redundancy: testing a new prediction of the intersensory redundancy hypothesis. *Developmental Science*, 13(5):731–737.

Balaban, M. T. and Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, 64(1):3–26.

Best, C. A., Robinson, C. W., and Slousky, V. M. (2010). The effect of labels on visual attention: An eye tracking study. In *Annual Conference of the Cognitive Science Society*, pages 1846–1851. Cognitive Science Society.

Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer (Version 5).

Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3:215–229.

Cummings, A., Saygin, A., Bates, E., and Dick, F. (2009). Infants' recognition of meaningful verbal and nonverbal sounds. *Language Learning and Development*, 5(3):172–190.

Delaherche, E. and Chetouani, M. (2010). Multimodal coordination: exploring relevant features and measures. In *International Workshop on Social Signal Processing*, SSPW '10, pages 47–52. ACM.

El-Sallam, A. A. and Mian, A. S. (2011). Correlation based speech-video synchronization. *Pattern Recognition Letters*, 32(6):780–786.

Engelken, E. J., S. K. W. (1989). Saccadic eye movements in response to visual, auditory, and bisensory stimuli. *Aviation, Space and Environmental Medicine*, 60(8):762–768.

Farzin, F., Charles, E. P., and Rivera, S. M. (2009). Development of multimodal processing in infancy. *Infancy*, 14(5):563–578.

Ferry, A. L., Hespos, S. J., and Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: an advantage of words over tones. *Child Development*, 81(2):472–479.

Flom, R. and Bahrick, L. E. (2007). The development of infant discrimination in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology*, 43(1):238–252.

Flom, R. and Bahrick, L. E. (2010). The effects of intersensory redundancy on attention and memory: Infants' long-term memory for orientation in audiovisual events. *Developmental Psychology*, 46(2):428–436.

Fogel, A. (2000). *Infancy: Infant, family and society.* Wadsworth Publishing, 4 edition.

Frank, M. C., Vul, E., and Johnson, S. P. (2009). Development of infants attention to faces during the first year. *Cognition*, 110(2):160–170.

Frens, M. and Van Opstal, J. (1995). Spatial and temporal factors determine auditory visual interactions in human saccadic eye movements. *Perception and Psychophysics*, 57(6):802–816.

Frintrop, S., Backer, G., and Rome, E. (2005). Goal-directed search with a top-down modulated computational attention system. In *27th Annual Meeting of the German Association for Pattern Recognition*, volume 3663, pages 117–124. Springer.

Fulkerson, A. L. and Waxman, S. R. (2007). Words (but not tones) facilitate object categorization : Evidence from 6- and 12-month-olds. *Cognition*, 105(1):218–228.

Giard, M. H. and Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Cognitive Neuroscience*, 11(5):473–490.

Gliga, T., Volein, A., and Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old infants. *Journal of Cognitive Neuroscience*, 22(12):2781–2789.

Gogate, L. J., Bahrick, L. E., and Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, 71(4):878–894.

Gogate, L. J., Prince, C. G., and Matatyaho, D. J. (2009). Two-month-old infants sensitivity to changes in arbitrary syllable-object pairings: The role of temporal synchrony. *Journal of Experimental Child Psychology*, 35(2):508–519.

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., and Gordon, L. (1987). The eyes have it: lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, 14(1):23–45.

Grabner, H., Matas, J., Gool, L. V., and Cattin, P. (2010). Tracking the invisible: Learning where the object might be. In *International Conference on Computer Vision and Pattern Recognition*, pages 128–1292. IEEE.

Grahl, M., Joublin, F., and Kummert, F. (2010). A method for visual model learning during tracking. *Australian Journal of Intelligent Information Systems (AJIIPS)*, 11(2):29–34.

Grahl, M., Joublin, F., and Kummert, F. (2011). A method for multi modal object recognition based on self-referential classification strategies. EP20090177019200911125.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*, volume 13. Prentice Hall.

Hein, G., Doehrmann, O., Muller, N., Kaiser, J., Muckli, L., and Naumer, M. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience*, 27(30):7881–7887.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504.

Hershey, J. and Movellan, J. (1999). Audio-vision: Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems*, pages 813–819. MIT Press.

Hollich, G., E., M., Helder, N., and C., P. (2004). Are you synching what i' am synching? In *International Conference on Development and Learning, Poster Presentation.*

Hollich, G., Newman, R. S., and Jusczyk, P. W. (2005). Infants use of synchronized visual information to separate streams of speech. *Child Development*, 76(3):598–613.

Hopfinger, J. B., Buonocore, M. H., and Mangun, G. R. (2000). The neural mechanisms of top-down attentional control. *Nature Neuroscience*, 3(3):284–291.

Howard, R. J., Brammer, M., Wright, I., Woodruff, P. W., Bullmore, E. T., and Zeki, S. (1996). A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. *Current Biology*, 6(8):1015–1019.

Hupbach, A., Gomez, R., Bootzin, R., and Nadel, L. (2009). Nap-dependent learning in infants. *Developmental Science*, 12(6):1007–1012.

Itti, L., Dhavale, N., and Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Annual International Symposium on Optical Science and Technology*, pages 64–78. SPIE Press.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Javed, O. and Ali, S. (2005). Online detection and classification of moving objects using progressively improving detectors. In *International Conference on Computer Vision and Pattern Recognition*, pages 696–701. IEEE.

Johnson, M., Posner, M., and Rothbart, M. (1994). Facilitation of saccades toward a covertly attended location in early infancy. *Psychological Science*, 5(2):90–93.

Johnson, M. H. (1990). Cortical maturation and the development of visual attention in infancy. *Journal of Cognitive Neuroscience*, 2(2):81–95.

Johnson, M. H., Posner, M. I., and Rothbart, M. K. (1991). Components of visual orienting in early infancy: Contingency learning, anticipatory looking, and disengaging. *Journal of Cognitive Neuroscience*, 3(4):335–344.

Kaiser, J., Hertrich, I., Ackermann, H., Mathiak, K., and Lutzenberger, W. (2005). Hearing lips: gamma-band activity during audiovisual speech perception. *Cerebral Cortex*, 15(5):646–653.

Kalal, Z., Matas, J., and Mikolajczyk, K. (2010). P-n learning: Bootstrapping binary classifiers by structural constraints. In *International Conference on Computer Vision and Pattern Recognition*, pages 49–56. IEEE.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227.

Lee, J.-S. and Ebrahimi, T. (2011). Audio-visual synchronization recovery in multimedia content. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2280–2283. IEEE.

Lehmann, S. and Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, 24(2):326–334.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin*, 126(2):281–308.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connection Science*, 15(4):151–190.

Makovski, T., Vazquez, G. A., and Jiang, Y. V. (2008). Visual learning in multiple-object tracking. *PLoS ONE*, 3(5).

Masataka, N. (1992). Motherese in a signed language. *Infant Behavior and Development*, 15(4):453–460.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.

Mehler, J., Dupoux, E., Nazzi, T., and Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infants viewpoint. In James, L. K. D., editor, *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, pages 101–116. Erlbaum, Mahwah, NJ.

Molholm, S., Ritter, W., Javitt, D. C., and Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. *Cerebral Cortex*, 14(4):452–465.

Monosov, I., Sheinberg, D., and Thompson, K. (2010). Paired neuron recordings in the prefrontal and inferotemporal cortices reveal that spatial selection precedes object identification during visual search. *Proceedings of the National Academy of Sciences*, 107(29):13105–13110.

Morrongiello, B., Fenwick, K., and Chance, G. (1998). Crossmodal learning in newborn infants: Inferences about properties of auditory-visual events. *Infant Behavior and Development*, 21(4):543–553.

Mort, D., Perry, R., Mannan, S., Hodgson, T., Anderson, E., Quest, R., McRobbie, D., McBride, A., Husain, M., and Kennard, C. (2003). Differential cortical activation during voluntary and reflexive saccades in man. *Neuroimage*, 18(2):231–46.

Newell, F. N. (2004). *Cross-modal object recognition.* In: Calvert G., Spence C., Stein B. E. (Eds.) The handbook of multisensory processes, 123–139, MIT Press, Cambridge.

Onat, S., Libertus, K., and König, P. (2007). Integrating audiovisual information for the control of overt attention. *Journal of Vision*, 7(10):1–16.

Plunkett, K., Hu, J.-F., and Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106(2):665–681.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1):3–25.

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., and Hennon, E. A. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development*, 77(2):266–280.

Reuter, B., Kaufmann, C., Bender, J., Pinkpank, T., and Kathmann, N. (2010). Distinct neural correlates for volitional generation and inhibition of saccades. *J. Cognitive Neuroscience*, (22):728–738.

Richards, J. E. (2004). *Development of covert orienting in young infants.* In: L. Itti and G. Rees and J. Tsotsos (Eds.) Neurobiology of attention. New York: Academic Press /Elsevier.

Richardson, D. and Kirkham, N. (2004). Multi-modal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology*, 133(1):46–62.

Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, 3(7):1199–1204.

Robinson, C. W. and Sloutsky, V. M. (2010). Development of cross-modal processing. *Cognitive Science*, 1(1):135–141.

Roder, B. J., Bushnell, E. W., and Sasseville, A. M. (2000). Infants preferences for familiarity and novelty during the course of visual processing. *Infancy*, 1(4):491–507.

Rolf, M., Hanheide, M., and Rohlfing, K. (2009). Attention via synchrony: Making use of multimodal cues in social learning. *Transactions on Autonomous Mental Development*, 1(1):55–67.

Ruesch, J. (2008). Multimodal saliency-based bottom-up attention: A framework for the humanoid robot icub. In *International Conference on Robotics and Automation*, pages 962–967. IEEE.

Schillingmann, L., Wrede, B., and Rohlfing, K. (2009). A computational model of acoustic packaging. *Transactions on Autonomous Mental Development*, 1(4):226–237.

Sirois, S. and Mareschal, D. (2002). Models of habituation in infancy. *Trends in Cognitive Sciences*, 6(7):293–298.

Slaney, M. and Covell, M. (2000). Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Advances in Neural Information Processing Systems*, pages 814–820. MIT Press.

Spelke, E. S. (1981). The infant's acquisition of knowledge of bimodally specified events. *Journal of Experimental Child Psychology*, 31(2):279–299.

Spelke, E. S. (1994). Initial knowledge: six suggestions. *Cognition*, 50(1–3):431–445.

Spelke, E. S. (2000). Core knowledge. *American Psychologist*, 55(11):1233–1243.

Steil, J. J., Götting, M., Wersing, H., Körner, E., and Ritter, H. (2007). Adaptive scene dependent filters for segmentation and online learning of visual objects. *Neurocomputing*, 70(7–9):1235–1246.

Tarullo, A. R., Balsam, P. D., and Fifer, W. P. (2011). Sleep and infant learning. *Infant and Child Development*, 20(1):35–46.

Theeuwes, J., Olivers, C. N., and Chizk, C. L. (2005). Remembering a location makes the eye curve away. *Psychological Science*, 16(3):196–199.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Trepel, M. (2003). *Neuroanatomie: Struktur und Funktion.* Urban und Fischer Verlag/Elsevier GmbH.

Triesch, J. and von der Malsburg, C. (2001). Democratic integration: self–organized integration of adaptive cues. *Neural Computation*, 13(9):2049–2074.

Von Hofsten, C. and Rosander, K. (1996). The development of gaze control and predictive tracking in young infants. *Vision Research*, 36(1):81–96.

Von Hofsten, C. and Rosander, K. (1997). The development of smooth pursuit tracking in young infants. *Vision Research*, 37(13):1799–1810.

Walther, D. and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407.

Walther, D., Rutishauser, U., Koch, C., and Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100(1–2):41–63.

Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Wiley-IEEE Press.

Woodward, A. L. and Guajardo, J. J. (2002). Infants' understanding of the point gesture as an object-directed action. *Cognitive Development*, 17:1061–1084.

Xiao, M., Wong, M., Umali, M., and Pomplun, M. (2007). Using eye-tracking to study audio-visual perceptual integration. *Perception*, 36(9):1391–1395.