A controlling strategy for an active vision system based on auditory and visual cues

Miranda Grahl, Frank Joublin, Franz Kummert

2010

Preprint:

This is an accepted article published in Int. Conf. on Artificial Neural Networks (ICANN). The final authenticated version is available online at: https://doi.org/[DOI not available]

A controlling strategy for an active vision system based on auditory and visual cues

Miranda Grahl[1], Frank Joublin[2], and Franz Kummert[1]

[1]Cor-Lab, Bielefeld University, D-33615 Bielefeld/Germany
mgrahl,franz@cor-lab.uni-bielefeld.de
http://www.cor-lab.uni-bielefeld.de
[2]Honda Research Institute Europe GmbH, D-63073 Offenbach/Germany
Frank.Joublin@honda-ri.de
http://www.honda-hri.de

Abstract. It is still an open question how preliminary visual reflexes can be structured by auditory and visual modalities in order to recognize objects. Therefore, we propose a new method for a controlling strategy for an active vision system that learns to focus on relevant multi modal aspects of the environment. The method is bootstrapped by a bottom up visual saliency process in order to extract important visual points. In this paper, we present our first results and focus on the unsupervised generation of training data for a multi-modal object recognition. The performance is compared to a human evaluated database.

Key words: adaptive learning, active vision, object recognition

1 Introduction

Active vision starts from retinal filtering and is understood as a process that actively interacts with the environment in order to control the gaze towards relevant aspects like objects. Most object recognition systems suffer from training with hand annotated data resulting in an inflexibility regarding spontaneous changes. So far, little work has been done in the computational modeling of an object recognition process, which automatically extracts a structure of auditory and visual cues in order to gain an object representation. Object recognition in an online learning scenario features a wide range of challenges. This means to build up a system that incrementally learns the structure of a demonstrated object. The ability to enhance the visual sensitivity on repeated exposures to multi-modal sources like movements of the mouth and speech requires the integration of bimodal signals. In addition, this requires a mechanism which selects stimulus driven relevant visual and auditory features in an initial learning phase in order to define unsupervised learned classifiers. Walter and Koch [1] propose a model that links a bottom up attention model to an object recognition system with an attentional modulation. This approach focuses on visual perceptual properties of the environment. In order to maintain object constancy for an object recognition, Newell [2] proposes that a constancy can be achieved by a

multi sensory representation and refers to the interaction with haptic cues. Xiao [3] studies the effect of task irrelevant sound on the oculomotor system. The analysis with different pitch deviants shows that the smooth pursuit ability increases with an increasing of the pitch. Lehmann [4] investigates the influence of past audio-visual object representations on an unimodal object recognition task. The criteria of memory performance and accuracy are improved if an object has been perceived in both modalities. Molholm [5] also suggests that an audio-visual representation leads to a faster and more accurate object detection performance and hypothesizes that auditory input modulates the processing in regions of the lateral occipital cortex. This challenges to find features that link the auditory and visual part of an object. Furthermore, a system needs to discriminate relevant information from irrelevant information automatically. Roy et. al [6] addresses the problem of finding significant features for the learning of auditory and visual cues between objects and speech. This approach uses the mutual information as clustering criterion and selects images and speech segments according to their mutual information maximization. A few approaches have been suggested to estimate audio-visual correlations [7], [8]. In contrast to this methods, our approach researches the correlation of auditory and visual properties from an active vision perspective and therefore focuses on space variant regions. In section 2, we present a system architecture for the control of an active vision system. Section 3 focuses on the correlation of auditory and visual properties with respect to center activity. A conclusion about the performance of unsupervised generation of training data is given in section 4.

2 System Architecture

In the following, the architecture [9] (fig. 1) is described with respect to the shown components. At each time when the camera moves onto a new position and tracks the scene for a defined time, the field of vision is processed with visual filters. The central region of the observed scene is correlated with auditory cues. The visual filtering is initially determined by predefined filters (6) and results into a saliency map. In relation to the new position, the movement is defined by a saccade logic (7) that calculates the center position by using the saliency map. The extraction of the most important point defines the camera movement. The moment of the movement is determined by a timer logic (8) that defines a new saccade and the system reevaluates the scene center. During the track of the scene, the system separates the acquired sound in active and non active audio segments. In a first learning phase the audio signal is not classified and is not accessible for the saliency computation with a weighting by auditory classifiers. The properties of the active audio segments are correlated with visual properties of the camera center and serves as criterion for retaining auditory and visual segments. The correlation computation provides a basis to extract visual and auditory segments in order to cluster them (11, 12) and to prepare classes for learning of an auditory classifier and for additional visual saliency filters. The correlation computation (9) is carried out during the tracking and

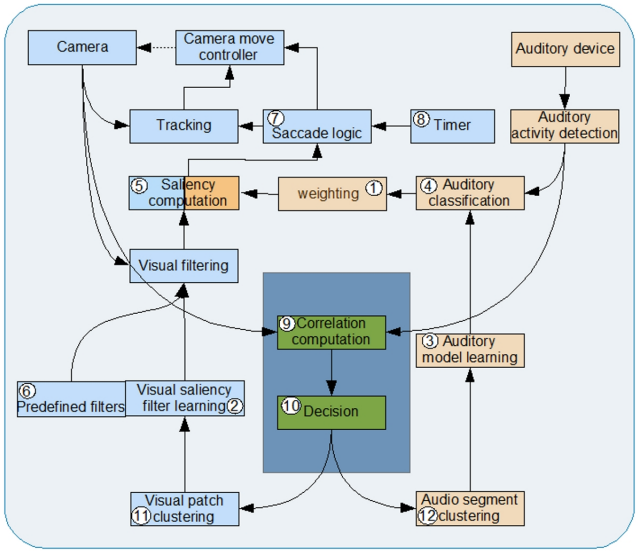

Fig. 1. An active vision system that attends on multi-modal relevant aspects. In the initial phase, the gaze selection of the architecture is reactive and the controlling strategy is defined by [10]. This model extracts visual salient points of the environment. An object classification based on auditory characteristics requires an associative learning of visual and auditory concepts (e.g. mouth/speech, hand/knocking). Therefore a learning of additional visual saliency filters (2) and auditory classifiers (3) are proposed. The association of learned concepts is defined by a weighting (1) of the relative importance of *visual saliency filters* that determines the saliency computation.

results in a statistic that provides indications about the mutuality of visual and auditory properties. The decision (10) of retaining single segments depends on the observed mutuality and extracted visual regions and auditory segments that have been correlated in particular. In this paper, we focus on part 9 and 10.

3 Selection of relevant visual and auditory segments

The proposed selection mechanism serves as a criterion for what to learn and suppresses the processing of noise with respect to missing coherence of auditory and visual information. Hershey et. al [7] define temporal synchronous observations of audio a_t and visual signals $v(x,y)_t$ as *Mutual Information* I (1). Both events at timestamp t are drawn independently from a joint Gaussian process with variances $\hat{C}_{A_t}$ and $\hat{C}_{V(x,y)_t}$ belonging to a joint covariance matrix $\hat{C}_{AV(x,y)_t}$.

$$I(x,y)_t = -\frac{1}{2}\log\left(\frac{|\hat{C}_{A_t}|\cdot|\hat{C}_{V(x,y)_t}|}{|\hat{C}_{AV(x,y)_t}|}\right) . \tag{1}$$

This approach suffers from a constant time averaging and hence temporal changes of I are not adapted. Therefore, we use the method proposed by Rolf [8]. We investigate in the analysis of the audio energy a_t and motion activity v_t defined by the difference of intensity images with respect to center activity $w\left(x,y\right) = exp\left(\left(-x^2-y^2\right)/\sigma^2\right)$. For $\hat{C}_{V(x,y)_t}$ we estimate a threshold $\hat{\gamma}_t$ (2) during the tracking:

$$\hat{\gamma}_t = \hat{\gamma}_{t-1} + \beta \cdot (\gamma_t - \hat{\gamma}_{t-1}) \text{ with } \gamma_t = \sum_{x,y} w \cdot \hat{C}_{V(x,y)_t} . \tag{2}$$

Those regions that don't exhibit significantly a large variance of v_t are removed for a further correlation computation:

$$\hat{C}_{V(x,y)_t} = \begin{cases} 0, & \text{if } \hat{\gamma}_t > \hat{C}_{V(x,y)_t} \\ \hat{C}_{V(x,y)_t}, & else \end{cases} \tag{3}$$

Active acoustic segments are obtained by applying a fixed threshold. After each tracking step k the observed mutuality is summarized with $I_k = \sum_{x,y} w \cdot I\left(x,y\right)_t$. In order to select relevant visual and auditory events $\hat{r}_{v,a}$ that have been correlated, I_k is evaluated after each tracking sequence (4). Firstly, the thresholding step ensures that visual and auditory information are removed that obtain low correlation activity. The threshold θ_1 is adapted by removed correlation measurements during the whole observation of the scene. The filtering includes a second threshold θ_2 and is determined by a randomization step Ir_k computed in parallel. For this a is drawn from a normal distribution with σ and μ estimated from origin active acoustic segments.

$$\hat{r}_{v,a} = \begin{cases} 1 & \text{if } I_k > max(\hat{\theta}_1, \theta_2) \\ 0 \text{ and } \quad \hat{\theta}_{1_k} = \hat{\theta}_{1_{k-1}} + \beta \cdot (\theta_{1_k} - \hat{\theta}_{1_{k-1}}) & \text{else} \end{cases} \tag{4}$$

If the estimated I_k yields a higher mean value as the estimated θ_1 and the randomized correlation θ_2, than the visual and auditory information are selected as relevant. Otherwise the correlation is caused by noise.

4 Results

The manual classification contains the separation between relevant and not relevant information of the sequences. The next saccade movement is determined by a saliency map that is computed by color, motion and orientation. An inhibition of return leads to a gaze selection to locations that have not been attended before. A new saccade is triggered each second. The *Mutual Information* I is weighted with $\alpha = 0.05$ and w is defined with $\sigma = 0.1$/cut-off $= 0.5$. The threshold for motion activity $\hat{C}_{V(x,y)_t}$ and $\hat{\theta}_1$ are adapted with $\beta = 0.5$. In order to analyze the performance of our approach, we use a dataset that is manually classified by humans into relevant and irrelevant patches. The dataset is recorded under laboratory conditions and shows a speaking person. The dataset comprises sequences

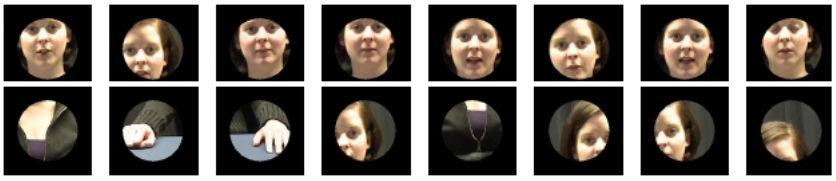

Fig. 2. Example dataset of relevant and irrelevant information (upper/bottom row). For the evaluation, the sequences are extracted in a predefined step with our algorithm. They contain always the last image and the sound information from start to the end of the tracked scene. The image view is restricted according w. The labeling criterion is defined by the appearance of redundant information of both modalities. This means if the sequence contains a mouth and is coherent with speech, the sequence is marked as relevant. Otherwise the sequence is marked as irrelevant. We conducted our analysis on 105 tracking sequences k. Sequences without any sound activity are removed from the dataset. By the manual annotation, we get 35 relevant combinations of auditory and visual information and 70 not relevant combinations.

of images with sound. Figure 2 shows a set of visual patches marked as relevant and irrelevant. As our thresholding criterion (4) contains a random parameter, we repeated our analysis for five times on the dataset. The results are averaged by the number of trails. Compared to the manually annotated data, our automatic approach finds 46 % (table 1) of the dataset that are selected as relevant (tp). Our method classifies 54 % combinations as not relevant. The fn error

Table 1. Evaluation results: The true positive error *tp* and false negative error *fn* describes those patch combinations that are selected as important and unimportant from relevant ones. The false positive error *fp* and true negative error *tn* describes those patch combinations that are selected as important and unimportant from irrelevant ones.

	tp	fn	tn	fp	relevant	irrelevant
relative	0.46	0.54	0.83	0.17		
average total	16	19	58.2	11.8	35	70
d_μ	26	39.4	251.2	217.1		
d_σ	17.6	31.4	165.8	148.4		

shows a loss of training data. This does not implicate an influence of a further clustering step. The most difficult task for an object recognition system consists in the unsupervised description of not relevant information. The tn error is 83 % and shows the effectiveness of our approach. Most irrelevant combinations are identified. Only 17 % are detected as relevant. Hence an unsupervised extraction of valid training data is ensured for a further clustering step. An additional analysis of the spatial distribution of center views of selected and rejected information shows a difference in the different conditions. This similarity is measured

by the euclidean distance d between the centers resulted from a sequence k. The results show that in case of the *tp* the average distance is smaller than in the case of *fp*. This means the accepted visual information from false wise accepted correlation events are distributed to center reference and can not share common features. In contrast to this, the visual fields that are evaluated as *tp* provides a basis for a common feature representation reasoned by the low d_μ.

5 Conclusion

This paper introduces an active vision architecture that is bootstrapped by a visual bottom up process and a correlation computation. A threshold adaptation takes place during the tracking and removes not significant aligned audio-visual events. The results provide a significant discrimination of relevant auditory-visual information. The investigation in the analysis with respect to center reference provides a preliminary clustering and a basis for learning of visual saliency filters.

Acknowledgments. The work described was supported by the Honda Research Institute Europe.

References

1. Walther, D., Koch C.: Modeling attention to salient proto-objects. Neural Networks 19, 1395–1407 (2006)
2. Newell, F.N.: Cross-modal object recognition. In: Calvert G., Spence C., Stein B.E. (eds) The handbook of multisensory processes. MIT Press, Cambridge, 123–139 (2004)
3. Xiao, M., Wong, M., Umali, M., Pomplun, M.: Using eye-tracking to study audio-visual perceptual integration. Perception 36(9), 1391–1395 (2007)
4. Lehmann, S., Murray M.M.: The role of multisensory memories in unisensory object discrimination. Cognitive Brain Research 24(2), 326–334 (2005)
5. Molholm S., Ritter W., Javitt D.C., Foxe J.J.: Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. Cerebral Cortex 14, 452–465, (2004)
6. Roy D.: Learning Audio-Visual Associations using Mutual Information. In: Proceedings of International Workshop on Integrating Speech and Image Understanding, 147–163 (1999)
7. Hershey, J., Movellan, J.: Audio-vision: Using audio-visual synchrony to locate sounds. In: Advances in Neural Information Processing Systems, 813–819 (1999)
8. Rolf M., Hanheide M., Rohlfing K.: Attention via synchrony: Making use of multimodal cues in social learning. IEEE Transactions on Autonomous Mental Development, 55–67 (2009)
9. Grahl M., Joublin F., Kummert F.: A method for multi modal object recognition based on self-referential classification strategies. European Patent Application, No. 09177019.8, pending (2009)
10. Itti, L., Koch, C., Niebur E.: A model of saliency-based visual attention for rapid scene analysis. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)