Visual Attention by Audiovisual Signal-Level Synchrony

Matthias Rolf
Dep. Adaptive Machine Systems
Osaka University, Japan
matthias@ams.eng.osaka-u.ac.jp

Minoru Asada
Dep. Adaptive Machine Systems
Osaka University, Japan
asada@ams.eng.osaka-u.ac.jp

ABSTRACT

Current approaches to artificial attention are largely limited to the visual domain. Only some consider audition as a source of information at the same time. Yet, attention is not necessarily limited to a single modality or a mere agglomeration of several modalities in human perception. Cross-modal attention, and its manipulation by cross-modal cues, seems to play a vital role in asymmetric interactions such as a parent tutoring a child. We discuss previous efforts [23] to reflect such perceptual processes with an artificial attention system that considers signal-level synchrony between vision and audition to guide visual attention. Results show that the system is receptive to infant directed cues from parents.

Keywords

Bottom-up Attention, Cross-Modal Synchrony, Social Robotics, Developmental Robotics

1. INTRODUCTION

Both infants and robots need to make sense of multi-modal streams of information when someone wants to teach them the meaning of word or how to do something. For instance, the concept behind a word that comes as an auditory stimulus needs to be associated to a visual stimulus or proprioceptive sensation. Both parts are first of all situated within an entirely unsegmented and continuous stream of information. Making sense of a novel concept therefore requires to *identify* and *segment* the relevant stimuli from each modality among irrelevant information, and to properly *relate* them to each other. In contrast to current robotics system, infants utilize very specific, rich, and multi-level support from their parents when trying to solve this tremendously difficult task. When parents teach their children about the meaning of word or how to do something, they largely modify their behavior compared to adult-adult interactions and establish a highly interactive and multi-modal teaching-process, rather than teaching in uni-directional and uni-modal ways. For instance, recent approaches to word learning take advantage of interactive processes between participants [34]. Gogate, Bahrick and Watson [8] took an experimental approach and investigated how mothers interact with their children when they try to teach them a new word. The authors found out that when mothers were asked to teach a new word for the objects or actions, they moved the objects in temporal synchrony with the new label, hence establishing a *cross-modal* relation beyond the content of each isolated modality. Other modifications of parental teaching compared to adult-directed communication include modifications each of prosody in speech [7] and demonstrated motion [3, 18], both of which are hypothesized to highlight relevant information and allow to "package" [29] them together. Besides parents emitting such cues, it has been shown that infants are well receptive to them. The power of cross-modal binding has been shown for newborn and young infants as auditory stimulation has been found to facilitate the visual attention [17, 14]. Infants as young as two months are sensitive to voice-lip synchrony during speech [16]. Furthermore, recent studies by Zukow-Goldring and colleagues [33] using eye-tracking technology with video confirms that 9 to 15 months old infants prefer looking at objects that are presented in a synchronous word-object condition. Consequently, experiments [9] revealed that 7 month-olds indeed benefit from synchronous stimuli when learning to map a syllable onto an object. The "Intersensory Redundancy Hypothesis" [2] attempts to explain how synchrony of signals can guide infants' selective attention and contribute to the learning process. Signals from different modalities reinforce each other on the basis of amodal properties like synchrony which promotes earlier processing. They thus attract the attention of perceivers and become foreground in contrast to other properties to become background [2]. The infant's initial sensitivity to amodal information such as synchrony – as it has been shown in the study by Gogate and her colleague [8] – provides an economical way of guiding perceptual processing to focus on meaningful, unitary events [2]. Hence, infants' learning is promoted in a cycle of both uni-modal and cross-modal parental tutoring cues, and infants' attentional mechanisms well tuned to these cues.

How can robotics systems benefit from such findings? The principle idea is to let the robot take the place of the infant, and let it benefit from the cues sent by human through interaction [31], taking the perspective of developmental robotics [1]. Therefore we have to consider the symbiotic way of interaction between infant and caregiver and replicate the crucial mechanisms on the infant-side in robotics systems [21]. Sev-

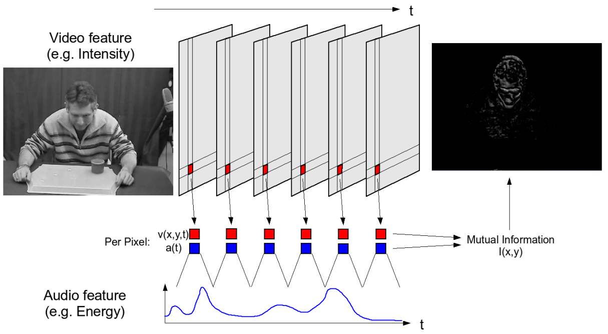

Figure 1: Schematic overview of synchrony detection: a video feature is per pixel compared to a audio feature for a short window in time. Mutual information is computed based on a linear correlation coefficient, yielding a topographic map of synchrony.

eral studies have already pointed out that infant-directed cues also occur when adults teach child-like robots [30, 28]. Here, we discuss previous efforts [23] to make artificial systems *receptive* to such cues, focusing on the exploitation of synchrony between visual and auditory presentations in parental teaching. The main idea is to try to detect synchrony between visual and auditory stimuli at an early level and let it guide visual attention, e.g. preferably attending to synchronous stimuli like an object moved simultaneously with its verbal label. Computational approaches to visual attention have naturally been made mostly based on visual information only, such as in traditional saliency models [13]. While it is increasingly realized that social interaction between robots and humans requires a more multimodal account [32], most approaches limit themselves to a mere agglomeration of modalities (*multi*-modal) instead of exploiting the very *cross*-modal relation between them. For instance, each modality is first processed to obtain uni-modal attention maps before results are merged into a combined map [24, 26, 25]. In contrast, we seek to exploit the cross-modal relation between visual and auditory information in caregivers' teaching at the earliest possible, i.e. signal, level and utilize it for visual attention right away. In the following we will briefly outline our computational account [23] to synchrony detection for attention. Results [23] show that our model is indeed receptive to cues in parent-child interactions in which more synchrony is found than in adult-adult interactions. Thereby the model often points out relevant locations in the visual field in situations were pure visual saliency fails. We finally discuss further (application) perspectives and first successful applications.

2. SYNCHRONY DETECTION

In order to guide visual attention to stimuli that expose high audiovisual synchrony, we consider a stream of video data accompanied by audio. Our method is based on an algorithm proposed by Hershey and Movellan [11]. The algorithm detects temporal correlations (synchrony) between visual features and auditory features. Therefore each image-location (i.e. pixel) is treated separately. The statistical analysis is restricted to a small window in time that is shifted over the audio/video stream. Since each pixel yields independent estimates of synchrony, the result is a topographic map of synchrony. As the final synchrony estimate is a mutual information measure, such maps are also referred to as "mixelgram" (see Fig. 1). Like other approaches to signal-level synchrony, the algorithm was originally developed for statistical sound-source localization (see also [12, 4]). In that scenario, it is assumed that physical sound-sources provide synchronous patterns across modalities. Stimuli that provide synchrony but do not correspond to an immediate sound-source are considered as false positives or disturbances. However, our application context is broader since we want to detect social cues that do not directly refer to physical sound sources. For our purposes, the algorithm has two important properties: Firstly, the algorithm contains no assumptions about the kind of visual (e.g. faces) or auditory stimuli. From the learning perspective this is important since such specific patterns shall be result of, but not a prerequisite for, an overall learning process. Secondly, the algorithm is fast enough to detect synchrony in real time with reasonable video resolutions and sampling rates (in contrast, e.g., to methods based on canonical correlation analysis [4]), which permits usage in a closed interaction loop. The model has already been compared to infants' abilities in synchrony detection by Prince *et al.* [20].

The basic mathematical assumption for the statistical analysis is that the values of visual and auditory features originate from a joint probabilistic process. This process is assumed to be stationary and Gaussian for a short period of time. We denote the set of n audio-features over time as $a(t) \in \mathbb{R}^n$ and the set of m video-features for each pixel $v(x,y,t) \in \mathbb{R}^m$. For the synchrony detection, the parameters means μ and (co-)variances Σ are estimated from the video data $\{a(t_k), v(x,y,t_k)\}_k$, where $k \in \{1,\ldots,T\}$ denotes the index of a video frame and T the number of frames in a video. For practical reasons we do not use a discrete time window as in [11] but compute exponentially smoothed estimates of $\mu(x,y,t_k)$ and $\Sigma(x,y,t_k)$ which can be done efficiently in one step per frame. The data from a current frame $(a(t_k), v(x,y,t_k))$ receives a constant weight $\alpha \in]0;1[$ and is recursively combined with the previous estimates of mean and variance:

$$\mu(x,y,t_k) = \alpha \cdot \begin{pmatrix} a(t_k) \\ v(x,y,t_k) \end{pmatrix} + (1-\alpha)\cdot \mu(x,y,t_{k-1})$$

$$\Sigma(x,y,t_k) = \frac{1}{1+\alpha}\left(\alpha \cdot \left(\begin{pmatrix} a(t_k) \\ v(x,y,t_k) \end{pmatrix} - \mu(x,y,t_{k-1})\right)^2 + \Sigma(x,y,t_{k-1})\right)$$

The estimates of (co-)variances $\Sigma(x,y,t_k)$ are then used to express the degree of synchrony between audio and video in terms of mutual information I. Assuming a Gaussian distribution yields an immediate relation that, in case of each only one audio and video feature, can be simply expressed [11] in terms of a Pearson correlation coefficient ρ:

$$I_{A,V}(x,y,t_k) = -\frac{1}{2}\log\left(1-\rho^2(x,y,t_k)\right)$$

$$\rho(x,y,t_k) = \frac{\sigma_{A,V}(x,y,t_k)}{\sqrt{\sigma_A(t_k)\cdot\sigma_V(x,y,t_k)}}$$

Thereby $\sigma_{A,V}$, σ_A and σ_V are the now scalar estimates of variances and the covariance of audio- and video- feature.

The overall result is one mutual information image (mixelgram) per frame. High values of mutual information are

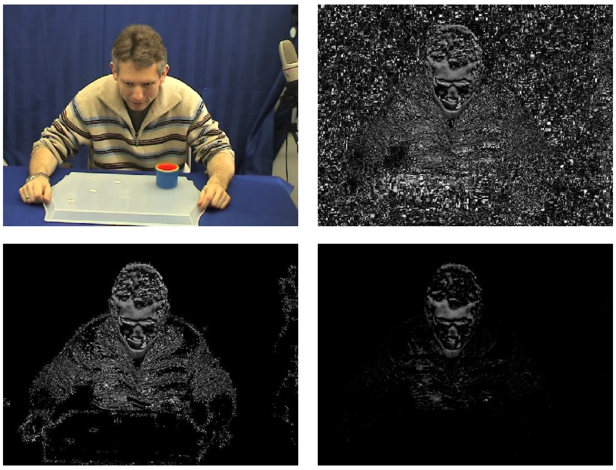

Figure 2: Top left: RGB-frame from a test-video. Top right: Mutual information. Background noise and lighting changes accompanied by peaks in audio can sometimes cause intensive correlation artifacts. White color corresponds to a mutual information of 0.51 ($\rho = \pm 0.8$). Bottom left: Threshold on video variance. Bottom right: Morphological erosion.

visualized with lighter gray scale values (see Fig. 2) and express a high degree of synchrony between audio and video. In the original scenario of sound-source localization, high mutual information reflects a possible sound-source at a certain image location. However, in our scenario we are less interested in such physically causal correlation. Instead, we rather try to investigate the role of synchrony for attention in tutoring. Hence, it is assumed that the model is also perceptive to synchrony induced by the tutoring process itself. Therefore, a mixelgram can directly be interpreted in terms of attention so that image regions with high mutual information receive the highest degree of attention.

Filtering. Pearson's correlation and mutual information as measures of interdependence between audio and video indicate the significance of a relation between both modalities. The significance of the signal itself is first of all not taken into account. In fact, most pixels in an image are usually static apart from noise, thus providing no significant change over time. Nevertheless those pixels can cause high correlation just by chance (see Fig. 2). We proposed a two-stage filter process to exclude insignificant visual stimuli and noise. The first stage excludes pixels without activity. As measurement of activity we use the variance over time on each pixel. If the variance on a pixel is below a specified threshold $\mathcal{T}_V$, mutual information is set to zero. Figure 2 illustrates the effect: large areas of stationary background are filtered out. Still, there is notable noise in regions that must be considered to be active. This noise results in single, outstanding pixels with high mutual information (Fig. 2, bottom left). These single pixel distortions are effectively handled by the second filter stage: a morphological erosion. Each pixel value is replaced by the minimum value of its direct neighborhood. Thereby, single outstanding pixels are completely erased, while massive regions of mutual information are retained (Fig. 2, bottom right).

3. EXPERIMENTS

The major goal of our experiments [23] was to investigate synchrony in a social learning scenario in terms of child-directed communication. It is important to note that the data for our analysis encompasses a contingent interaction since we analyzed parental behavior during a real situation with their children. In this situation, they continuously reacted and adapted to their child. The basic hypothesis is that during a demonstration, parents provide additional learning cues by synchrony. The hypothesis is tested by comparing the degree of audiovisual synchrony between adult- and child-directed communication.

Materials and Procedure. We investigated 184 videos showing 48 participants, teaching either their adult partner or their child how to do one of four tasks (stacking cups, using toy blocks, ringing a bell, using a salt-shaker). The data stems from the original video corpus [21, 18], which contains videos of 66 parental couples interacting with their children. The infants' age ranged from 8 to 30 months. After excluding trials with disturbances such as interaction with the experimenter, 192 videos were selected for the analysis. They were equally distributed over 4 tasks, each with 12 parental couples in 4 runs. Eight further videos had to be excluded due to missing or corrupted audio tracks or annotations, yielding a final number of 184 videos available for analysis. In the study, both parents interacted with their child and with an adult. The first run was an adult-child interaction, in which one parent (randomly selected) and her or his child sat across the table. The parent was instructed to demonstrate the function of the objects to the child. Here, the parent was free to teach either the word, the action, or both (those two acts were in fact mostly inseparable in the collected data). We asked to move the white tray and to give the objects to the child only after the demonstration. The child was attending to the demonstration and interacting with the parent. In a following adult-adult interaction, the same parent was asked to demonstrate the object to her or his partner. In the third run, the second parent demonstrated the objects to the child. In the fourth run, the same parent demonstrated the objects to an experimenter.

Measurement and Features. In order to assess the overall synchrony of a demonstration towards either adult or infant, we first evaluated the average mutual information for each point in time in a video. Thereby the averaging only took place across pixels in a videoframe that were not excluded by the variance threshold:

$$\begin{aligned} S(t_k) &= \frac{1}{|\mathcal{P}(t_k)|} \sum_{(x,y)\in\mathcal{P}(t_k)} I(x,y,t_k) \\ \mathcal{P}(t_k) &= \{(x,y) : 0 < I(x,y,t_k)\} \end{aligned}$$

Results were then averaged over time, and normalized against a respective measurement in which the original audio track was replaced by white noise.

$$\mathcal{S} = \frac{1}{T}\sum_{k=1}^{T} S(t_k) \qquad \mathcal{S}^{rel} = \frac{\mathcal{S}}{\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}^{noise}(i)}$$

Values $\mathcal{S}^{rel} > 1$ indicate a higher degree of synchrony being detected than expectable by chance (e.g. noise), and hence a significant synchrony between speech and visual body or object movement.

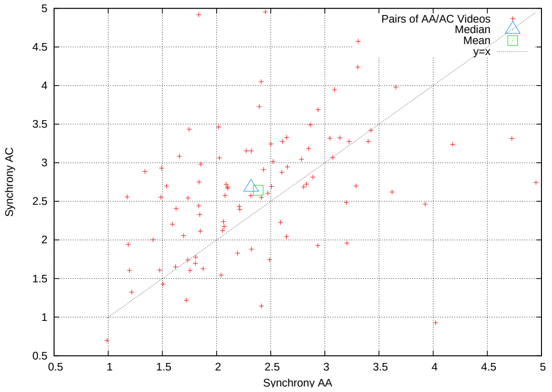

Figure 3: Synchrony results for gradient-strength as feature, $\alpha = 0.05$ and $\mathcal{S}^{rel}$ as measure. Synchrony in each adult-adult video is plotted against the synchrony in the corresponding adult-child video.

		Median		
Settings		AC	AA	Significance level
int	0.1	3.48	2.96	0.005
int	0.05	3.86	3.09	0.001
int	0.02	2.73	2.57	–
grad	0.1	2.31	2.00	0.001
grad	0.05	2.68	2.32	0.001
grad	0.02	2.18	1.96	0.1

Table 1: Comparison between adult-directed and child-directed communication for intensity images and Sobel-based gradient strength as features and different values of α. The median for AC-synchrony is significantly higher than for AA in all settings.

As audio-feature, we used audio-energy (as i.a. used in [11, 15] or similarly RMS values in [20]). As video features, we used image intensity (grayscale values) and gradient-strength images alternatively, which were found [22] to provide the best discrimination between synchronous and asynchronous stimuli. We tested both intensity and gradient-strength images with three temporal smoothing factors $\alpha \in \{0.02, 0.05, 0.1\}$. The variance threshold was defensively chosen and fixed at $\mathcal{T}_V = 5.0$ for both features.

Results. The goal of this experiment is to compare synchrony in adult- and child-directed communication. For this purpose, each video showing an adult-adult (AA) interaction is compared to the corresponding adult-child (AC) video. Figure 3 exemplary shows the synchrony results for gradient-strength feature and $\alpha = 0.05$. Each point in the plots corresponds to a pair of an AA and AC video, where the synchrony in the AA video is plotted on the x-axis and the synchrony in the AC video on the y-axis. The first observation is that all except for three videos gained synchrony values above 1.0. That means that the video signals gained higher mutual information with the original audio track than with audio noise. Hence, a real synchrony could be detected. For a direct comparison between AA and AC conditions, the main diagonal (i.e. $x = y$) is shown in the plot. A point above this diagonal indicates that more synchrony is found in the child-directed interaction than in the corresponding adult-directed situation. Indeed most points (here 62 out of 92) lie above the diagonal. Both median and mean show higher synchrony for the child-directed situation. For this parameter setting, the median synchrony is 2.32 for AA videos and 2.68 for AC videos. The significance of this effect was tested with a two-tailed sign test. The sign test between paired random variables $(a_i, b_i)_{i=1..N}$ thereby tests the null hypothesis $H_0 : P(A < B) = P(A > B) = 0.5$. Here the null hypothesis is that synchrony in AC has the same probability to be higher or lower than in the corresponding AA situation. On the dataset presented here, this null hypothesis can be rejected with high significance (error probability $p < 0.001$). The effect can be reproduced across diverse parameter settings (see Table 1). With respect to the different interaction tasks, the wooden brick scenario shows a significant ($p < 0.01$) trend towards more synchrony in child-directed communication. For all scenarios the median of AC synchrony is higher than the AA median, indicating that the effect is rather task-independent.

An additional observation in the results is that synchrony in child-directed situations is positively correlated with the synchrony in corresponding adult-directed situations. Due to individual differences parents tend to produce high synchrony in AA situations, when they also produce relatively high synchrony in AC situation. We assessed this effect with the Spearman rank correlation coefficient. Analogous to Pearson's correlation, it indicates positive correlation with values between 0.0 and 1.0, but is more robust to outliers. For the settings shown in Figure 3 Spearman's correlation is 0.480. The effect shows to be significant w.r.t. the null hypothesis that the variables are uncorrelated ($p < 0.01$ with a two-tailed t-test). Also this effect can be reproduced across several parameter settings. Thereby a positive correlation is also found within each task.

If parental teaching provides learning cues by means of synchrony, it is important to understand *what* these cues actually indicate. Here we discuss some exemplary scenes with respect to spatial aspects on mutual information in the video sequences. Child-directed tutoring was already investigated [18] w.r.t. the spatial distribution of visual saliency [13]. Thereby, a part of the same cup stacking demonstrations towards infants was investigated as used in this work. It was shown that different motion patterns in adult- and child-directed communication caused higher saliency on demonstrated objects in child-directed situations. A comparison between attention via saliency and synchrony can generally be done in two ways: first of all the entire saliency map (or the mixelgram) can be interpreted in terms of *covert* attention [19]. As the potential importance of each image region is encoded in those maps, one can directly compare e.g. face and hand of a subject w.r.t. their importance relative to each other. A more condensed view can be gained in terms of *overt* attention [19]: each saliency map and each mixelgram is reduced to a single attended position – a focus of attention. For saliency maps this is simply the position with the highest value. Thereby we basically used the same saliency configuration as in [18], evaluating intensity, color, orientation, difference images and optical flow by means of Itti and Koch's Saliency map model [13]. In contrast to [11] and [20], we do not find this location within a mixelgram

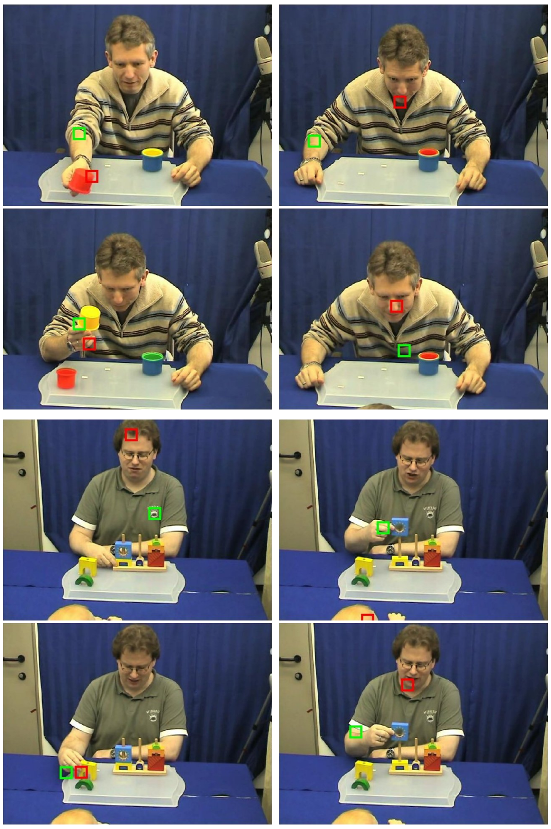

Figure 4: Exemplary maximum positions of mutual information (red) and saliency (green).

by means of a center of gravity since we are not interested in a huge region of synchrony but in a region of high synchrony – whatever size it has. Therefore we apply a 15x15 Gaussian filter to smooth each mixelgram and then detect the position of maximum mutual information.

The analysis of two exemplary videos is shown in Figure 4. The maxima of each saliency and mutual information during parental speech in the child-directed condition are visualized. The first video shows the demonstration of cup stacking. Maximum mutual information is often found on a shown object due a synchronized presentation, but also often on mouth and head due to the inherent synchrony with speech utterances. Obviously, the cups are no source of sound in these situations, but provide synchrony due to the interplay of parents' speech and motion. The highest saliency is often found on the subject's pullover sharply contrasting the background, but also in the subject's action space (in the vicinity of the hands) due to salient movements. The second video shows a demonstration of the wooden bricks. The maximum mutual information is mostly found in the action space, and – contrary to the first video – less often on the face. However, the synchrony is, in some frames, distracted towards e.g. the infant's head, moving into the camera view. Also the saliency maxima are mainly restricted to this action space, but not exclusively to the hands and objects as a maximum can for instance be found on the shirt's sticker. Taking both videos into account, a perfect detection of task-relevant locations can neither be expected from synchrony or saliency, nor from any bottom-up attention strategy. However, we can state that synchrony quite often points toward those locations and is hardly vulnerable to conspicuous modality-specific stimuli like textures or colors, whereas saliency maps are by design.

4. DISCUSSION

Our experimental results [23] give a clear indication that child-directed interaction indeed involves a higher synchronization between gestures (or generally movement) and speech. Though this effect was also described by Gogate *et al.*, it is remarkable and encouraging that it can be detected even at signal level, and by means of a computational attention system. It has been argued that cues from child-directed communication help to guide attention towards important parts of either the speech signal or the visual scene [3, 6, 8]. The shown spatial distributions and example frames suggest that mutual information can indeed be used to find relevant image locations. Gogate *et al.* found that object motion is often used synchronously to a word label in multimodal motherese. Though the correlation analysis is performed on an entirely different level, this is consistent with the observation that high mutual information values can be found on shown objects during parental speech.

So far we did not analyze the temporal characteristics of synchrony. It was argued [5] that child-directed speech has i.a. the function to *arouse* and *guide* the infant's attention. Whereas our study focuses on the guidance, it is also likely that synchrony in multimodal motherese is used to arouse the infant's attention when the child is currently not attending to the parent or the task. In that case, an increased level of synchrony might be measurable. Both functions are highly plausible in the context of the Intersensory Redundancy Hypothesis [2] as young infants have been shown to preferentially attend to synchronous stimuli.

Deploying such attention system based on cross-modal analysis already on signal level in an actual closed loop of human robot interaction seems a promising direction for future studies. First studies already pointed out how to successfully gather training data for object recognition [10] from human tutoring recordings by selecting cross-modally synchronous stimuli. However, the approach appears to be potentially useful not only for robots, but also as assistive system for humans with perceptual and attentional deficits, such as Autistic-spectrum disorder (ASD) patients [27]. In conclusion, we can say that audiovisual signal-level synchrony for visual attention might contribute to enabling a symbiotic interaction loop such as between infants and caregivers also between humans and robots.

Acknowledgments

This study has been supported by the JSPS Grant-in-Aid for Specially promoted Research (No. 24000012).

5. REFERENCES

[1] M. Asada, K. Hosoda, Y. Kuniyoshi, and H. Ishiguro. Cognitive developmental robotics: A survey. *IEEE Trans. Autonomous Mental Development*, 1(1), 2009.

[2] L. Bahrick, R. Lickliter, and R. Flom. Intersensory Redundancy Guides the Development of Selective Attention, Perception, and Cognition in Infancy. *Current Directions in Psychological Science*, 2004.

[3] R. Brand, D. Baldwin, and L. Ashburn. Evidence for 'motionese': modifications in mothers' infant-directed action. *Developmental Science*, 5(1):72 – 83, 2002.

[4] H. Bredin and G. Chollet. Measuring audio and visual speech synchrony: methods and applications. In *IET Int. Conf. Visual Information Engineering*, 2006.

[5] R. Cooper, J. Abraham, S. Berman, and M. Staska. The Development of InfantsâĂŹ Preference for Motherese. *Infant Behavior and Development*, 20(4):477–488, 1997.

[6] Dominey, P.F., Dodane, C. Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics*, 17(2-3):121–145, 2004.

[7] Fernald, A. & Mazzie, C. Prosody and Focus in Speech to Infants and Adults. *Developmental psychology*, 27(2):209–221, 1991.

[8] L. Gogate, L. Bahrick, and J. Watson. A Study of Multimodal Motherese: The Role of Temporal Synchrony between Verbal Labels and Gestures. *Child Development*, 71(4):878–894, July/August 2000.

[9] L. J. Gogate and L. E. Bahrick. Intersensory Redundancy and 7-Month-Old Infants' Memory for Arbitrary Syllable-Object Relations . *Infancy*, 2(2):219 – 231, 2001.

[10] M. Grahl. *Focus of Attention on Relevant Multimodal Events*. PhD thesis, Bielefeld University, 2012.

[11] J. Hershey and J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. *Advances in Neural Information Processing Systems*, 12:813–819, 2000.

[12] T. Ikeda, H. Ishiguro, and M. Asada. Attention to clapping - a direct method for detecting sound source from video and audio. In *IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent Systems*, 2003.

[13] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[14] P. Kaplan, K. Fox, D. Scheuneman, and L. Jenkins. Cross-modal facilitation of infant visual ïňĄxation: Temporal and intensity effects. *Infant Behavior and Development*, 14(1):83–109, 1991.

[15] Kidron, E., Schechner, Y., Elad, M. Cross-Modal Localization via Sparsity. *IEEE Transactions on Signal Processing*, 55:1390–1404, 2005.

[16] Meltzoff A. and Kuhl P. Faces and speech: intermodal processing of biologically relevant signals in infants and adults. In D. Lewkowicz and R. Lickliter, editors, *The Development of Intersensory Perception: Comparative Perspectives*. Lawrence Erlbaum, 1994.

[17] M. J. Mendelson and M. M. Haith. The relation between audition and vision in the human newborn. *Monographs of the Society for Research in Child Development*, 41(4), 1976.

[18] Y. Nagai and K. Rohlfing. Can Motionese Tells Infants and Robots "What To Imitate"? In *Int. Symp. Imitation in Animals and Artifacts*, April 2007.

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

[20] C. G. Prince, G. J. Hollich, N. A. Helder, E. J. Mislivec, A. Reddy, S. Salunke, and N. Memon. Taking synchrony seriously: A perceptual-level model of infant synchrony detection. In *Int. Workshop Epigenetic Robotics*, pages 89–96, 2004.

[21] K. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann. How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183–1199, 2006.

[22] M. Rolf. Audiovisual attention via Synchrony. Master's thesis, Bielefeld University, 2008.

[23] M. Rolf, M. Hanheide, and K. Rohlfing. Attention via Synchrony: Making Use of Multimodal Cues in Social Learning. *IEEE Transactions on Autonomous Mental Development*, 1(1):55–67, 2009.

[24] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer. Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *IEEE International Conference on Robotics and Automation*, 2008.

[25] B. Schauerte and G. A. Fink. Focusing computational visual attention in multi-modal human-robot interaction. In *Int. Conf. Multimodal Interfaces and Machine Learning for Multimodal Interaction*, 2010.

[26] B. Schauerte, B. KÃijhn, K. Kroschel, and R. Stiefelhagen. Multimodal saliency-based attention for object-based scene analysis. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2011.

[27] L. Schillingmann, M. Rolf, S. Kumagaya, S. Ayaya, and Y. Nagai. Assistance for autistic people by segmenting and highlighting cross-modal perceptual information. In *Annual Conference of the Robotics Society of Japan (RSJ)*, 2013.

[28] L. Schillingmann, B. Wrede, K. Rohlfing, K. Fischer, and G. Sagerer. The structure of robot-directed interaction compared to adult-and infant-directed interaction using a model for acoustic packaging. In *Spoken Dialogue and Human-Robot Interaction Workshop*, 2009.

[29] L. Schillingmann, B. Wrede, and K. J. Rohlfing. A computational model of acoustic packaging. *IEEE Trans. Autonomous Mental Development*, 1(4), 2009.

[30] A.-L. Vollmer, K. S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. J. Rohlfing, and B. Wrede. People modify their tutoring behavior in robot-directed interaction for action learning. In *IEEE Int. Conf. Development and Learning (ICDL)*, 2009.

[31] B. Wrede, K. J. Rohlfing, M. Hanheide, and G. Sagerer. Towards learning by interacting. In *Creating Brain-like Intelligence*, pages 139–150. Springer, Berlin Heidelberg, 2009.

[32] H. Yan, M. H. Ang Jr., and A. N. Poo. A survey on perception methods for human-robot interaction in social robots. *Int. J. Social Robotics*, July 2013.

[33] P. Zukow-Goldring and N. d. V. Rader. Caregiver gestures cultivate a shared understanding: Assisted imitation and early word learning. In *Intermodal Action Structuring*, Bielefeld, Germany, July 2008.

[34] Zukow-Goldring, P. Socio-perceptual bases for the emergence of language: An alternative to innatist approaches. *Developmental Psychobiology*, 23(7):705–726, 1990.