

# Automatic Enhancement of Correspondence Detection in an Object Tracking System

Denis Schulze<sup>1</sup>, Sven Wachsmuth<sup>1</sup> and Katharina J. Rohlfing<sup>2</sup>

1- University of Bielefeld - Applied Informatics  
Universitätsstr. 25, 33615 Bielefeld - Germany

2- University of Bielefeld - Emergentist Semantics  
Universitätsstr. 25, 33615 Bielefeld - Germany

**Abstract.** This paper proposes a strategy to automatically detect the correspondence between measurements of different sensors using object tracking. In addition the strategy includes the ability to learn new features to facilitate the correspondence computation for future measurements. Therefore first a correlation between objects of different modalities is computed using time synchronous changes of attribute values. Using statistical methods to determine the dependencies between changes of different attributes it is shown how a multi layer perceptron (MLP) can be used to enhance the correspondence detection in ambiguous situations. <sup>1</sup>

## 1 Introduction

Nowadays high effort is spend on robotic research. One of its subfields mentioned here is sensor fusion. This describes the process of integrating information from different sensor modalities into a multi-modal representation of the environment. Psychological findings provide high evidence that this integration plays an important role for humans and that cross-modal integration of sensory data can influence the interpretation of certain events. In [1] *H. McGurk* and *J. MacDonald* show that the combination of the visual signal of a speaking mac together with a slightly changed audio signal has an effect on the understanding of spoken syllables. Another common example is the ventriloquism effect [2] which labels the effect that the perceived location of the sound source is shifted towards the mouth of the doll of the ventriloquism. According to these findings it is shown, that the integration of audio and visual information can increase the understanding of spoken speech if the visual data of the mouth is also available to the listener.

The McGurk effect provides some evidence, that humans must have a clue about the motion of the mouth and the emitted sound. This knowledge must be acquired during life time, but how does this happen? *Lewkowicz* postulates that intersensory integration is based on the perception of intersensory synchrony [3]. Based on this assumption our proposed strategy first uses time synchronous changes of attribute values to determine a dependency between representations of objects in different sensory fields. Afterwards those dependencies are used to generate training examples for a MLP that facilitates the correspondence computation based on object measurements in ambiguous situations.

---

<sup>1</sup>This work was supported by EXC 277, Cognitive Interaction Technology (CITEC).

The rest of the paper is structured as follows. In section 2 a short overview of the related work is given. It also includes a short description of the ideas behind our strategy. In section 3 an overview about the software and the algorithms is provided followed by a description and evaluation of the experiments in section 4. The paper closes with a conclusion in section 5.

## 2 Related work

Performing fusion of sensory data is often studied and many approaches can be found in the literature ([4], [5], [6], [7]).

As mentioned in the introduction, psychological findings suggest, that inter-sensory synchrony plays an important role in sensor fusion [3]. One example where synchrony between audio and visual signals is used, can be found in [4]. There, *J. Hershey* and *J. Movellan* implemented a system that uses time synchronous changes in visual and audio features to compute a so called mixelgram which describes for every pixel the mutual information between the signals. The mixelgram is then used to find the speaking person in the scene. These time synchronous changes can be detected very easily by just comparing the time-stamps where two or more attributes change in time. Therefore synchrony in time is taken as a first hint to determine the correspondence between different sensor measurements in our strategy.

Other approaches use different statistical learning methods including neural networks. For example time-delayed neuronal networks are used to fuse visual and audio data either for a lipreading system [5] or to search for a speaking person in a scene [7]. Another example where a Bayesian network is used in a smart kiosk environment can be found in [6]. Using those machine learning approaches has the advantage, that functional dependencies between signals can be uncovered automatically by the learning strategy. Unfortunately these machine learning approaches need, most of the time, a supervised learning phase where different training examples are provided to the network. Using the results of the first correlation computation training examples can automatically be generated and used for a machine learning approach that can be integrated into the correspondence computation.

## 3 Approach

The following section is related to the implementation of the strategy.

### 3.1 System

The system implementing our strategy is based on the work presented in [8]. It bootstraps from cases where object instances can be easily tracked by a single sensor and uses this information to extend its representation to a multi-sensor model. Therefore for every abstract sensor there is a subsystem that tracks the objects detected by this sensor (Fig: 1, Sensor - Percept processing - Anchor manager). Attribute values, such as the position or size of an object, that are

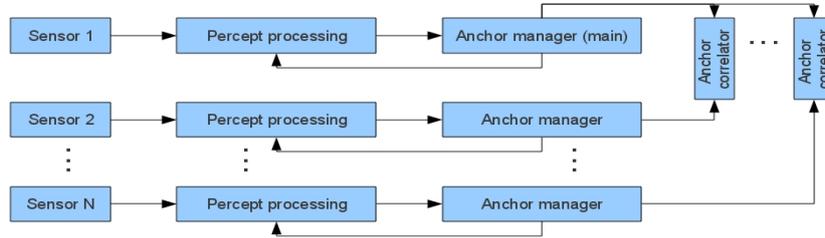


Fig. 1: A schematic drawing of the multi modal fusion system composed of different modules. For simplicity correlation values are only computed between anchors in the main anchor manager and connected anchor managers.

measured by a sensor are stored in attribute histories that are arranged in so called anchors. An anchor links *one* object detected by *one* sensor to a symbolic representation in the system [9]. Whenever new sensor data about an attribute of an object is available and the values exceed a certain difference to the last measurement, a new entry in the corresponding history is added, i.e., attribute histories and anchors are data structures indexed over time. Using the anchors in every tracking subsystem, we provide a general representation for objects in every sensory field.

The goal of our strategy is now to find those anchors in the different anchor managers that link the same physical object to a symbolic representation. Therefore the system computes likelihood values  $l$  that measures if an anchor  $a_j$  in a connected anchor manager belongs to an anchor  $m_i$  in the main anchor manager. The main idea is, that it is possible to collect those likelihood values over time and over different independent hints resulting in an increasing certainty  $c$  that two anchors  $m_i$  and  $a_j$  belong to the same physical object. Let  $l_1 = l(m_i, a_j)$  denote the current maximal likelihood value, then a certainty value  $c$  can be computed using:

$$c = \sqrt{\frac{l_1 - l_2}{l_1} * \frac{l_1 - l_3}{l_1}}$$

$$l_2 = \max_{1 \leq k \leq K} (l(m_i, a_k)); k \neq j$$

$$l_3 = \max_{1 \leq n \leq N} (l(m_n, a_j)); n \neq i$$

### 3.2 Time synchrony correlation

The first correlation computation between different anchors is based on time synchronous changes of attribute values. Therefore we compare attribute histories in main anchors, with all attribute histories in every anchor in the connected anchor manager. As already mentioned entries in a history correspond to time-stamps where the change in a measured attribute exceeds a certain threshold. This threshold should be selected in a way that it suppress changes that occur due to noisy data. The histories are then evaluated and for every entry in one

history, the entry with the nearest timestamp in the other history is searched. To reduce the amount of doubly computed correlation values, we always search the timestamps with the minimal time difference in the history with less entries. Let  $d_{min}$  denote this minimal difference, then a likelihood of correlation based on the timestamps is computed using:

$$l_t^1(m_i, a_j) = \psi_1\left(\frac{1}{\alpha} * \mathcal{N}_1(d_{min})\right)$$

where  $\psi_1$  is a thresholding function,  $\mathcal{N}_1$  is a Gaussian distribution and  $\alpha$  is a normalizing factor so that  $l_t^1(m_i, a_j) \in \{0, \dots, 1\}$ .

### 3.3 Neural network

Using only time synchronous changes to determine corresponding anchors in different anchor managers fails if more than two anchors have simultaneously changing attribute values. Therefore whenever the strategy succeeds in finding anchors that describe the same physical object, the histories of the attributes in the two anchors that change simultaneously are observed. Since time synchronous changes in attribute values denote a dependency between those attributes, entries in the histories can be used to generate training examples for machine learning approaches. In this paper MLPs are used to transform *one or more* attributes in an anchor of a connected anchor manager, to *one* attribute of an anchor in the main anchor manager, hence attribute values in connected anchors are used as training input and the corresponding attribute values of main anchor attributes describe the desired output. A comparison between the attribute values delivered from the perceptrons, and the current attribute values in the main anchor, measured by a sensor, delivers another likelihood value for anchor correspondence. Let  $d$  denote the distance between the computed and the measured attribute values then the likelihood of correspondence is computed using:

$$l_t^2(m_i, a_j) = \psi_2\left(\frac{(1 - e_{rmse})}{\beta} * \mathcal{N}_2(d)\right)$$

Where  $\psi_2$  is another thresholding function,  $e_{rmse}$  is the root mean square error of the network with  $e_{rmse} \in \{0, \dots, 1\}$ , and  $\beta$  is normalizing factor like  $\alpha$ . The overall likelihood value is updated whenever a new change occurs and is computed using:

$$l_t(m_i, a_j) = l_{t-1}(m_i, a_j) + l_t^1(m_i, a_j) + l_t^2(m_i, a_j)$$

## 4 Evaluation

To test the performance of the system we choose the thresholding value for  $\psi_1$  in a way, that all events that lie more than 300 ms apart result in a likelihood

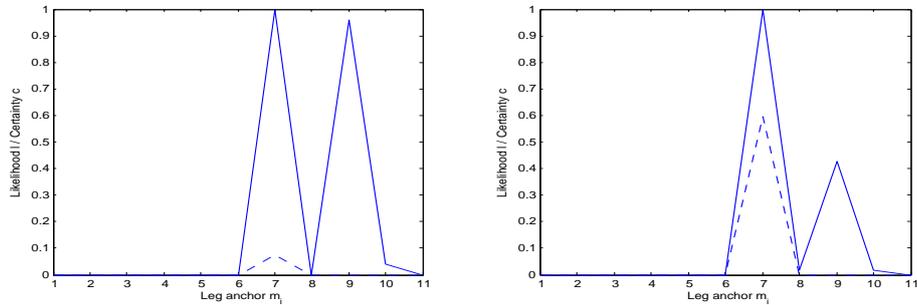


Fig. 2: Likelihood and certainty values (dashed line) that leg anchors belong to one specific face anchor. Left, only synchrony correlation is used while two persons are moving in front of the sensors. Right, same situation but additionally a MLP is used for spatial comparison of anchors. Values are collected after 18 seconds of movement and likelihood values are mapped to  $l \in \{0, \dots, 1\}$ .

value of 0 and  $\psi_2$  sets all likelihood values under 0.5 to 0. The deviation for the Gaussian distribution  $\mathcal{N}_2$  is chosen for every dimension of the mapped attribute separately and with respect to the root mean square error in this dimension. For  $\mathcal{N}_1$  we choose a mean value of 0 and a deviation of 100.

We connected the system with the output of a leg detection algorithm and a face detector (see [8]). The first one is based on a laser range finder and a simple clustering and classification algorithm is used to detect possible legs up to a distance of 5 meters and returns angle and distance values to all possible legs and pair of legs in the laser scan. The latter one delivers position and size information to faces in the camera image, which means, that the coordinates are described in pixel values. The camera is arranged above the laser, both pointing in the same direction.

Whenever a person is moving in front of the sensors and the legs and the face is visible to the corresponding sensor, the likelihood of correlation for these anchors is increasing. If only one person is moving, the time synchronous correlation is sufficient to determine the corresponding anchors. These cases are used to generate training examples for the neural network. Therefore whenever a new entry is inserted in one history and synchronous changes in other histories are detected, then the current values in the histories are used to create the training examples.

Figure 2 shows an example where two persons are simultaneously moving in front of the robot. On the left side of the figure, only the time synchronous changes of attribute values are used, resulting in a small value of certainty (dashed line) that the leg anchor  $m_7$  belongs to the face anchor ( $c = 0.076$ ). On the right side, the same situation is again evaluated, but this time the MLP, trained on 2000 training examples, is integrated in the correlation computation process. It transforms the size and position values of face anchors into polar coordinates that are compared against the position values of leg anchors. This time the correct leg anchor  $m_7$  receives a significantly higher certainty value for correlation ( $c = 0.596$ ).

## 5 Conclusion

In this paper a strategy to determine anchor correspondences is proposed and evaluated on a simple example. The provided experiments show that the strategy works, but that the performance clearly depends on the provided data. The time synchrony correlation is highly dependent on the quality of the attribute histories, i.e. that new entries are only added if the change to the last entry is significant and did not occur due to noisy data. Another drawback is, that it might take some time to collect the training data in highly dynamic environments, because first correlations can only be detected when only single persons move in front of the sensors.

## References

- [1] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–8, 1976.
- [2] D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology : CB*, 14(3):257–62, February 2004.
- [3] D. J. Lewkowicz. The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychological bulletin*, 126(2):281–308, March 2000.
- [4] J. Hershey and J. Movellan. Audio-Vision: Using Audio-Visual Synchrony to Locate Sounds. *Advances in Neural Information Processing Systems*, pages 813–819, 2000.
- [5] D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. *Proceedings of the IJCNN*, pages 289–295, 1992.
- [6] A. Garg, V. Pavlovic, and J.M. Rehg. Boosted learning in dynamic bayesian networks for multimodal speaker detection. *Proceedings of the IEEE*, 91(9):1355–1369, September 2003.
- [7] R. Cutler and L. Davis. Look who’s talking: speaker detection using video and audio correlation. *In Proceedings of ICME*, pages 1589–1592, 2000.
- [8] S. Lang, M. Kleinhagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer. Providing the basis for human-robot-interaction. *Proceedings of the ICMI*, November 2003.
- [9] S. Coradeschi and A. Saffiotti. Anchoring symbols to sensor data: Preliminary report. *Proceedings of the AAAI*, pages 129–135, 2000.