Learning from Incongruence

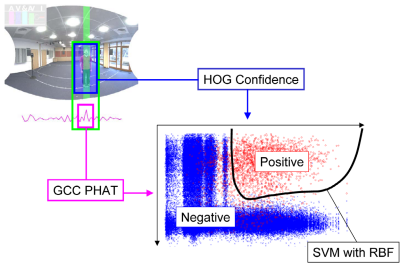
Tomáš Pajdla, Michal Havlena, and Jan Heller

Abstract We present an approach to constructing a model of the universe for explaining observations and making decisions based on learning new concepts. We use a weak statistical model, e.g. a discriminative classifier, to distinguish errors in measurements from improper modeling. We use boolean logic to combine outcomes of direct detectors of relevant events, e.g. presence of sound and presence of human shape in the field of view, into more complex models explaining the states in which the universe may appear. The process of constructing a new concept is initiated when a significant disagreement – incongruence – has been observed between incoming data and the current model of the universe. Then, a new concept, i.e. a new direct detector, is trained on incongruent data and combined with existing models to remove the incongruence. We demonstrate the concept in an experiment with human audio-visual detection.

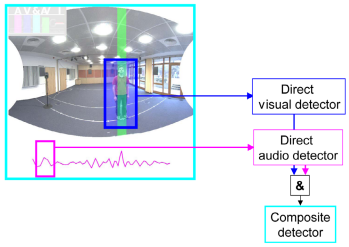
1 Introduction

Intelligent systems compare their model of the universe, the "theory of the universe", with observations and measurements they make. The comparison of conclusions made by reasoning about well established building blocks of the theory with direct measurements associated with the conclusions allow to falsify [1] current theory and to invoke a rectification of the theory by learning from observations or restructuring the derivation scheme of the theory. It is the disagreement – incongruence – between the theory, i.e. derived conclusions, and direct observations that allows for developing a richer and better model of the universe used by the system.

Tomáš Pajdla · Michal Havlena · Jan Heller
Center for Machine Perception, Department of Cybernetics, FEE, CTU in Prague,
Technická 2, 166 27 Prague 6, Czech Republic
e-mail: {pajdla, havlem1, hellej1}@cmp.felk.cvut.cz



(a) The direct audio-visual (A/V) human speaker detector



(b) The composite audio-visual (A&V) human speaker detector

Fig. 1 (a) The direct audio-visual human speaker detector constructed by training an SVM classifier with a RBF kernel in the two-dimensional feature space of GCC-PHAT values (*x*-axis) and pedestrian detection scores (*y*-axis) for different positive (*red circles*) and negative (*blue crosses*) manually labeled examples [6]. (b) The composite audio-visual human speaker detector accepts if and only if the direct visual detector AND the direct audio detector both accept (but possibly at different places) in the field of view. See [6] or an accompanied paper for more details.

Works [2, 3] proposed an approach to modeling incongruences between classifiers (called detectors in this work) which decide about the occurrence of concepts (events) via two different routes of reasoning. The first way uses a single *direct* detector trained on complete, usually complex and compound, data to decide about the presence of an event. The alternative way decides about the event by using a *composite* detector, which combines outputs of several (in [2, 3] direct but in general possibly also other composite) detectors in a probabilistic (logical) way, Figure 1.

Works [2, 3] assume direct detectors to be independent, and therefore combine probabilities by multiplication for the "part-membership hierarchy", resp. by addition for the "class-membership hierarchy". Assuming trivial probability space with values 0 and 1, this coincides with logical AND and logical OR. Such reasoning hence corresponds to the Boolean algebra [4]. In the next we will look at this simplified case. A more general case can be analyzed in a similar way.

The theory of incongruence [2, 3] can be used to improve low-level processing by detecting incorrect functionality and repairing it through re-defining the composite detector. In this work we look at an example of incongruence caused by the omission of an important concept in an example of audio-visual speaker detection and show how it can be improved. Figure 1 and Table 1 illustrate a prototypical system consisting of alternative detectors, which can lead to a disagreement between the alternative outcomes related to an event.

Three direct detectors and one composite detector are shown in Figure 2(a). The direct detector of "Sound in view", the direct detector of "Person in view", the direct detector of "Speaker", and the composite detector of "Speaker" are presented. The composite detector was constructed as a logical combination of direct detectors evaluated on the whole field of view, hence not capturing the spatial co-location of sound and look events defining a speaker in the scene. See [6] or an accompanied paper for more details.

Table 1 Interpretation of agreement/disagreement of alternative detectors of the "Speaker" concept and their possible outcomes. See text, [2, 3].

	Direct	Composite	Possible reason
1	reject	reject	new concept or noisy measurement
2	reject	accept	incongruence
3	accept	reject	wrong model
4	accept	accept	known concept

Table 1 shows the four possible combinations of outcomes of the direct and composite "Speaker" detectors as analyzed in [3]:

The first row of the table, where none of the detectors accept, corresponds to no event, noise or a completely new concept, which has not been yet learned by the system. The last row of the table, when both detectors accept, corresponds to detecting a known concept.

The second row, when the "Speaker" composite detector accepts but the direct one remains negative, corresponds to the *incongruence*. This case can be interpreted as having a partial model of a concept, e.g. not capturing some important aspect like the spatial co-location if the composition is done by AND. Alternatively, it also can happen when the model of the concept is wrong such that it mistakenly allows some situations which are not truly related to the concept if the composition is done by OR.

The third row of the table, when the direct "Speaker" detector accepts but the composite one remains negative, corresponds to the *wrong model* case. Indeed, this case applies when the composite detector mistakenly requires some property which is not truly related to the concept if the composition is done by AND. It also happens when the composite detector has only a partial model of the concept, e.g. when it misses one of possible cases in which the concept should be detected, if the composition is done by OR.

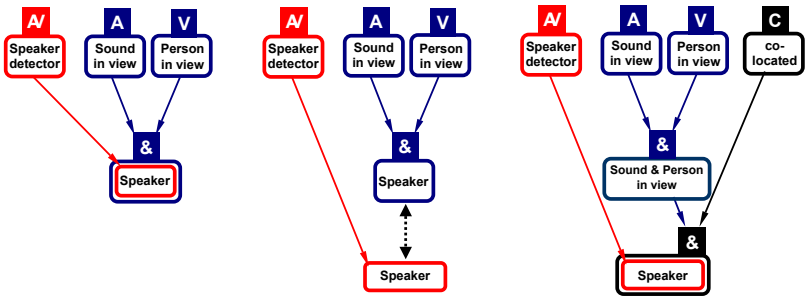
We can see that the interpretation depends on how the composite detectors are constructed. Restricting ourselves to Horn clauses [7], i.e. to making the compositions by AND, we interpret the second and third rows as in [3]. Horn clauses are a popular choice since they allow efficient manipulation, and are used in PROLOG [8].

Assume to have a composite detector C constructed in the form of a Horn clause of direct detectors $D_1, D_2, \ldots, D_n$

$$D_1 \wedge D_2 \wedge \ldots D_n \rightarrow C \tag{1}$$

which means that C is active if and only if all D_i are active. For instance "Person in view" $\wedge$ "Sound in view" $\rightarrow$ "Speaker" is a Horn clause.

With this restriction, a detected incongruence can be understood as if the composite detector missed a term on the left hand side of the conjunction in the derivation rule, which is responsible for falsely rejected cases. It is easy to remedy this situa-



(a) (b) (c)

Fig. 2 (a) Two different detectors are executed in parallel. The direct A/ detector captures concisely all observed data involved in the training. The composite A&V detector aims at deriving ("explaining") the presence of a human speaker in the scene from primitive detections of human look a human sound. (b) Since the composite A&V detector does not properly capture co-location of human look and human sound to represent a true human speaker, a discrepancy – incongruence – appears when the A/ detector accepts but the A&V detector rejects. (c) An improved composite A&V&C detector uses a direct detector of space-time co-occurrence to better model (and hence detect correctly) human speakers.

tion by learning a new concept corresponding to the missing term in the conjunction from the incongruent examples.

There are many possibilities how to do it. A particularly simple way would be to add a single new concept "Co-located" to the conjunction, i.e.

$$\text{"}\textit{Person in view}\text{"} \wedge \text{"}\textit{Sound in view}\text{"} \wedge \text{"}\textit{Co-located}\text{"} \rightarrow \text{"}\textit{Speaker}\text{"} \qquad (2)$$

which would "push" the composite detector "down" to coincide with the direct detector, Figure 2(b).

Somewhat more redundant but still feasible alternative would be to add two more elements to the system as shown in Figure 2(c). A new composite detector "Sound & Person in view" could be established and combined with another newly introduced concept "Co-located" to update the model in order to correspond to the evidence. Although somewhat less efficient, this second approach may be preferable since it keeps concepts for which detectors have been established already.

As suggested above, the incongruence, i.e. the disagreement between the direct and the composite detectors, may signal that the composite detector is not well defined. We would like to use the incongruent data to learn a new concept, which could be used to re-define the composite detector and to remove the incongruence.

In the case of the speaker detector, the composite audio-visual detector has to be re-defined. A new "Sound & Person in view" concept has to be initiated. The composite audio-visual detector is disassociated from the "Speaker" concept, new "Sound & Person in view" concept is created and associated to the composite detector. This new concept will be greater than the "Speaker" concept. Next, a new

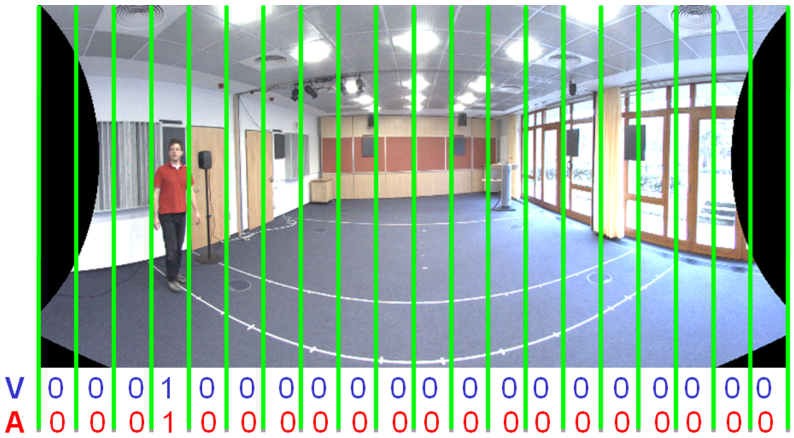


Fig. 3 The field of view is split into 20 segments (bins) and direct audio "Sound in view" detector (A) and direct video "Person in view" detector (V) are evaluated in each bin.

composite audio-visual "Speaker" detector is created as a conjunction of the composite "Sound & Person in view" detector and a new detector of an "XYZ" concept which needs to be trained using the incongruent data. The new composite detector is associated to the "Speaker" concept. The name of the "XYZ" concept can be established later based on its interpretation, e.g. as "Co-located" here.

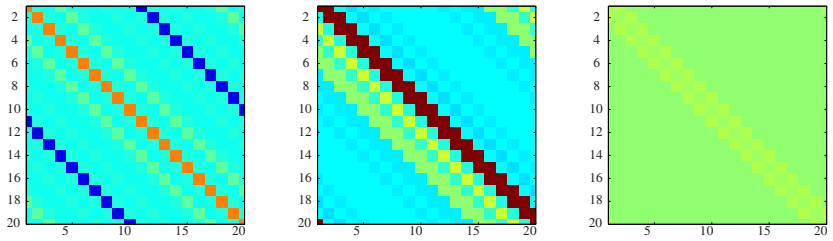
2 Learning "Co-located" Detector

We will deal with the simplest case when the incongruence is caused by a single reason which can be modeled as a new concept. Our goal is to establish a suitable feature space and to train a direct detector deciding the "XYZ" concept using the congruent and incongruent data as positive and negative training examples respectively. As the values of audio and visual features have been used in direct audio and direct visual detectors already and our new concept should be as general as possible, we will use only boolean values encoding the presence of a given event in the 20 angular bins, Figure 3, [6].

First, two feature vectors of length 20 are created for each frame, one encoding the presence of audio events and the other one encoding the presence of visual events, and concatenated together in order to form a boolean feature vector of length 40.

Secondly, in order to find dependencies between the different positions of the events, the feature vector is lifted to dimension 820 by computing all possible products between the 40 values. The original feature vector:

$$x_1\, x_2\, x_3\, \ldots\, x_{39}\, x_{40}$$



(a) (b) (c)

Fig. 4 Resulting weights for different pairs of values in the feature vector as obtained by SVM. Numbers on axes correspond to bin indices. Positive weights are denoted by red, zero weights by green and negative weights by blue color. (a) Audio $\times$ audio. (b) Audio $\times$ visual. (c) Visual $\times$ visual.

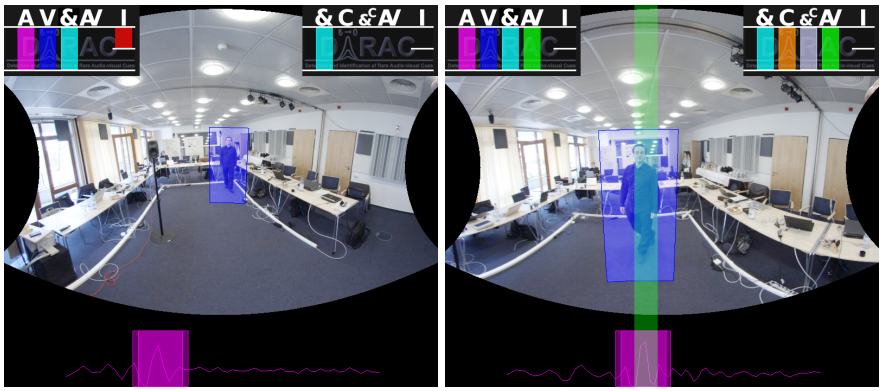
is transformed into:

$$x_1^2\ x_1x_2\ x_1x_3\ \ldots\ x_2^2\ x_2x_3\ \ldots\ x_{39}x_{40}\ x_{40}^2.$$

When the original values are boolean, the quadratic monomials $x_1^2, x_2^2, \ldots$ have values equal to $x_1, x_2, \ldots$ and the monomials $x_1x_2, x_1x_3, \ldots$ have values equal to the conjunctions $x_1 \wedge x_2, x_1 \wedge x_3, \ldots$. SVM training over these vectors should reveal significant pairs of positions by assigning high weights to the corresponding positions in the lifted vector.

We use two sequences to construct the positive and negative training example sets. Each of these sequences is nearly 5 minutes long with a person walking along the line there and back with a loudspeaker placed near one of the ends of the line. During the first approx. 90 seconds, the walking person is speaking and the loudspeaker is silent rendering a congruent situation. During the next approx. 90 seconds, the walking person is silent and the loudspeaker is speaking causing an incongruent situation. In the last approx. 90 seconds, both the walking person and the loudspeaker are speaking which is congruent by our definition as we are able to find a bin with a speaker.

For each frame, the concatenated boolean feature vector is created. The boolean decisions for the 61 directional feature vectors from the audio detector [6] are transformed into the 20 values of the audio part of the feature vector using disjunction when more directional labels fall into the same bin. The visual part of the feature vector is initialized with 20 zeros and each confident human detection output by the visual detector changes the value belonging to the angular position of the center of the corresponding rectangle to one. Feature vectors belonging to frames which yielded a positive response from both the composite and the direct audio-visual detectors, i.e. congruent situation, are put into the positive set, those belonging to frames that were classified positively by the composite but negatively by the direct audio-visual detector, i.e. incongruent situation, are put into the negative set, and



(a) (b)

Fig. 5 (a) The original composite detector falsely detects a speaker in the scene with a silent person and sound generated by a loudspeaker at a different place (response & in the top left corner). The improved composite detector (response $\&^{\mathsf{C}}$ in the top right corner) gives the correct result. (b) Both detectors are correct when a speaking person is present in the field of view.

those belonging to frames with a negative response from the composite audio-visual detector are discarded as such data cannot be used for our training.

As the loudspeaker position is fixed in our training sequences, we decided to remove the bias introduced by this fact by "rotating" the data around the bins, so each training example is used to generate 19 other training examples before lifting, e.g. a training example:

$$x_1\, x_2\, x_3\, \ldots\, x_{20}\, x_{21}\, \ldots\, x_{39}\, x_{40}$$

is used to generate 19 additional examples:

$$\begin{gathered} x_2\, x_3\, x_4\, \ldots\, x_1\, x_{22}\, \ldots\, x_{40}\, x_{21} \\ x_3\, x_4\, x_5\, \ldots\, x_2\, x_{23}\, \ldots\, x_{21}\, x_{22} \\ \vdots \\ x_{20}\, x_1\, x_2\, \ldots\, x_{19}\, x_{40}\, \ldots\, x_{38}\, x_{39}. \end{gathered}$$

Finally, the feature vectors of 60,320 positive and 41,820 negative examples were lifted and used to train a linear SVM detector [5].

The results shown in Figure 4 can be commented as follows. The most significant result is the dark red main diagonal in the $\mathsf{A}\times\mathsf{V}$ diagram, Figure 4(b), telling us that the positive examples have the audio and visual events in the same bin (or shifted by one bin as one of the neighboring diagonals is red too). Red square at $(1,20)$ is a by-product of the "rotation" as neighboring bins can be separated to different ends of the view-field.

As can be seen in the $\mathsf{V}\times\mathsf{V}$ diagram, Figure 4(c), pairs of visual events are insignificant. The orange main diagonal in the $\mathsf{A}\times\mathsf{A}$ diagram, Figure 4(a), says that

positive examples tend to contain more audio events. This is due to the fact that the only the situation with two audio events present in the training data was congruent, we had no training data with two loudspeakers speaking. The light blue adjacent diagonal is also an artifact of the direct audio-visual detector and "rotation".

To conclude, the just trained detector decides the position consistency of the audio and visual events, so a suitable name for the "XYZ" concept would be the "Co-located" concept. The resulting "Co-located" detector can be used to augment the initial composite "Speaker" detector, Figure 2(a), to produce new "Speaker" detectors, Figure 2(b,c). Figure 5(a) shows the original composite detector which falsely detects a speaker in the scene with silent person and sound generated by a loudspeaker at a different place (response & in the top left corner). The improved composite detector (response $\&^{\mathsf{C}}$ in the top right corner) gives the correct results. In Figure 5(b), both detectors are correct when a speaking person is present in the field of view.

3 Conclusions

We have seen that incongruence can be used to indicate where to improve the model of the universe. It provided the training labels to data to pick up the right training set for constructing a detector explaining the missing concept in the model. Of course, we have just demonstrated the approach on the simplest possible realistic example with a single clause explaining the state of the world by a very few concepts.

The next challenge is to find an interesting, realistic but tractable problem leading to a generalization of the presented approach. It would be interesting to use more general logical formulas as well as to deal with errors in direct detectors.

Acknowledgements This work was supported by the EC project FP6-IST-027787 DIRAC and by Czech Government under the research program MSM6840770038. We would like to acknowledge Vojtěch Franc for providing the large-scale SVM training code.

References

1. Popper, K.R.: The Logic of Scientific Discovery. Routledge (1995)
2. Pavel, M., Jimison, H., Weinshall, D., Zweig, A., Ohl, F., Hermansky, H.: Detection and identification of rare incongruent events in cognitive and engineering systems. Dirac white paper, OHSU (2008)
3. Weinshall, D., et al.: Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. In: NIPS 2008, pp. 1745–1752 (2008)
4. Halmos, P.R.: Lectures on Boolean Algebras. Springer (1974)
5. Franc, V., Sonneburg, S.: Optimized cutting plane algorithm for large-scale risk minimization. Journal of Machine Learning Research **10**, 2157–2232 (2009)
6. Pajdla, T., Havlena, M., Heller, J., Kayser, H., Bach, J.H., Anemüller, J.: Incongruence detection for detecting, removing, and repairing incorrect functionality in low-level processing. Research

Report CTU–CMP–2009–19, Center for Machine Perception, K13133 FEE Czech Technical University (2009)
7. Wikipedia: Horn Clause (2010), http://en.wikipedia.org/wiki/Horn_clause
8. Wikipedia: PROLOG (2010), http://en.wikipedia.org/wiki/Prolog