Deep Gestalt Reasoning Model: Interpreting Electrophysiological Signals Related to Cognition

András Lőrincz Áron Fóthi Bryar O. Rahman Viktor Varga
Neural Information Processing Group, Faculty of Informatics, Eötvös Loránd University
Budapest, Hungary

Abstract

We are to join deep input-output processing *and Gestalt Laws driven cognition under deterministic world assumption. We consider every feedforward input-output system as a* sensor*: including units performing holistic recognition. A mathematical theorem is also a* sensor*: it senses the consequences upon receiving its conditions. Systems seeking consistencies between the outputs of* sensor *are* cognitive units*. Such units are involved in cognition.* Sensor *and* cognitive units *complement each other. We argue that the goal of learning is to turn components of the cognitive system into feedforward holistic units for gaining speed in cognition. We put forth a model for self-training of the holistic units. We connect our concepts to certain electrophysiological signals and cognitive phenomena, including evoked response potentials, working memory, and consciousness. We demonstrate the working of the two complementary systems on low level situation analysis in videos.*

1. Introduction

The thought that consciousness and cognition are the results of unconscious computations has been long proposed as well as debated [10], although this thought is a must if we are trying to build a model for consciousness with computer programs. We put forth the idea that cognition, or '*making sense*' of the world of 'blooming buzzing confusion' [35] exploits *Gestalt Laws* by seeking consistency in both space and time. In order to simplify the formulation of our arguments, we treat sensory neurons, e.g., retinal neurons and tactile neurons on equal footing with feedforward holistic recognition units and we call them '*sensors*'. This way, the word sensor becomes a synonym for feedforward input-output system and disregards the complexity of the 'sensor'.

Feedback is inherently limited by the processing time of the feedback loop and corresponding temporal integration process. Feedback enables adaptivity, but it may become unstable. By contrast, both the time and the stability of feedforward processing are guaranteed, there is no loop delay. Correction needs external inputs and that can be delayed. Fast response requires the construction of more and more sophisticated and precise feedforward input-output systems or '*sensors*' that can respond to complex spatio-temporal structures being external to the processing sensors themselves.

We propose that '*making sense*' (i) is a relatively slow process, (ii) it exploits Gestalt Laws, (iii) develops a growing hierarchy of deeper and deeper feedforward sensors (iv) trained by the outcomes of the '*making sense*' process itself. We note that representations can't make sense. Instead, the input makes sense, if it can be matched by means of the representation (see, e.g. [25] and the references therein). We shall come back to this point later.

We assume that making sense is a consistency seeking process that exploits feedforward computations and eliminates inconsistent outputs. This thought is somewhat similar to the idea of recognition by components put forth by Biederman [4] except that we also consider discrepancies. Component based reasoning corresponds to cognition in our model. Cognition driven training of holistic recognition modules produces deeper input-output systems, shortens processing time and raises cognition to higher levels.

There are evidences that neocortical computation follows Bayes Law already at very low levels [33]. It is experience based consistence seeking, so should that be considered cognition? Also, experiments show that human thinking doesn't closely follow Bayes Law [42]. We think that the resolution may come from assuming (i) a noisy sensor sensory system, (ii) a deterministic environment, like Laplace did [21], and (iii) the optimization of collecting training samples for the holistic sensors. This triplet determines our inference system.

The paper is constructed as follows. We provide background information about cognition related electrophysiological signals in Sect. 2. Section 3 lists those Gestalt Laws that we are to consider (Sect. 3.1). It is followed by the deep network tools and the methodology that we use (Sect. 3.3)

for the analysis of videos when combining the complementing computations (Sect. 3.4). Demonstrative results are presented and discussed in Sects. 4 and 5, respectively. Conclusions are drawn in the last section (Sect. 6).

2. Background information about the brain

We are restricted in space that we try to compensate by mentioning experimental findings and an extensive, but by no means comprehensive list of the references. We start by considering cognition in the context of consciousness in order to exclude unconscious, but Bayesian inferences apparently present in early visual processing [33]. Awareness is a more dubious phenomenon and the different ways one can influence it can shed light on this separation.

Consider the seminal paper on binocular rivalry [24]. It is about (i) responses in the early visual cortex and (ii) the perceptual reporting in monkeys for images that resist binocular fusion and give rise to alternating percepts between two images. The information is available to the brain, neurons are responding to both potential percepts, some of them are modulated by the actual conscious percept, but we are aware only one part of the visual information at a time. The case of ambiguous figures, like the Necker cube [1], is different. We are aware of the full visual information at any given time instant, but the percept alternates. We say that in both cases, conscious observation 'makes sense of the input' as much as it can and the best it can do is to alternate between the potential interpretations.

Ouhnana et al. [34] studied the contextual effects for both cases. They showed that contextual influence is similar in the two cases suggesting that similar context-integration mechanisms operate under these two different conditions. Concerning our model, it means that awareness and consistence seeking, to some extent, are independent from each other.

In case of binocular rivalry the unobserved part of the input can enter the making sense procedure by reentrant processes that bring information from higher levels to lower levels. They are short, on the order of 100 ms, they can be influenced by fast object substitution masking [11], but the fast magnocellular channels can escape this masking [48], the different feedforward processing and reentrant channels can be dissociated by transcranial magnetic stimulation (TMS) [39], fast and low resolution channels constrain conscious observation [5, 7], similarly to optic flow regularization, the 'regularization' via fast channels affects observation probability, but not observation precision (slower channels) [19], there is a critical time window for binding information from different channels [20], and that the reentrant channels seem critical for visual awareness [18].

These particular features show that timing is critical and conscious observation requires feedback (reentrant) loops. We know from EEG studies that neuronal recruitment for initiating a motion may start about 2 seconds earlier than we think that decision for voluntary movement is made. It is known that delays are also present when movements are triggered and the observation of signals in the supplementary motor area precede the motion by about 200 ms or so, but conscious observation doesn't experience any delay in the motion. In turn, conscious observation compensates for the delays of the observation process as well as for the delays between control instructions launched by the brain and the starting time of the motion. Furthermore, conscious representation is synchronous with sensed but not yet processed inputs having processing time on the order of hundred milliseconds. Thus consciousness must integrate over time windows as claimed in [15] and must also exploit model-based prediction.

Taken together, compensation of time delays and the need of reentrant loops cut the infinite regress of homunculus fallacy short. The fallacy says that representation can't make sense and somebody should make sense of it. Then it asks about the agent that makes sense of the representation, what it is using and if it is a representation as well. Reentrant loops cut short the infinite regress to a reconstruction process: the input 'makes sense' if the representation can reproduce it via the reentrant channels [25]. This thought is of helps in understanding rivalry, too: conscious observation can reproduce only part of the rivalrous input and the non-reproduced part remains the subject of subconscious processing, overcomes the actual interpretation, and is rendered conscious after some time.

A prominent example of fusing multi-modal observations is the McGurk effect [27] in audio-visual integration. Such integration takes considerable time due to the different delays of the different processing (here audio and visual) channels. How is information fused and how would conscious observation emerge under such conditions? A recent paper considers the time course of consciousness, including the transitions between conscious states [2]. According to the authors conscious perception matures by systematically acquiring more and more qualities to the preceding interpretation (version) of the percept. They also mention the view of Herzog et al. [15], who propose that features of objects, for example, are unconsciously analyzed and all features become conscious simultaneously when unconscious processing is 'completed'.

2.1. Event related potentials

Time locked event-related potential (ERP) experiments with human participants may provide information about the different components. The earliest ERP studies already showed such dependencies, and early studies already showed task-related preparation signals prior to the antici-

[1] https://en.wikipedia.org/wiki/Necker_cube

pated trigger, i.e., the zero of the time-locked experiments, too [9]. ERP phenomena are complex and the interested reader is referred to the literature for gaining insights about the nomenclature. We limit ourselves to a few observations.

Vakli et al. [44] have shown that even simple Gestalt information about one component of the body, namely one point of the elbow, increases the P2 component when only face and hands are presented in veridical situation. Component P2 seems to be sensitive to the interplay between holistic and component based processing [44] as well as to deviations from typical configurations and appearances. For example, component P2 becomes smaller for caricatures and for other race faces [47]. In accordance with the view that both holistic and component based information influences the P2 component, face thatcherization[2] in upright headpose delays P2 over the occipito-temporal regions [6].

According to Salti et al. [37], the P3 component of the ERP reflects *conscious perception* and it is not influenced by the level of confidence. Findings of Metzger et al. [29] provide (i) support to this interpretation and (ii) further specification of the P3 component by means of binocular rivalry. They found that the timing of the P3b component is closely related to the timing of the reporting time of the individual about the perceptual change. In turn, the P3b component seems to correspond to changes of conscious perception.

We note that the consistence seeking algorithm doesn't have to be conscious; it may occur at every level. Furthermore, it can be feedback type in case of competition. Conscious consistence seeking involves the working memory and manifests itself in the corresponding ERP signals [12]. The relation between awareness and working memory, i.e., cognition is very complex. Considering that (i) consciousness has only partial access to sensory inputs, (ii) there are time delays between processing channels, (iii) there are critical windows for binding observations, and (iv) awareness requires that reentrant loops 'confirm' the representation, one can understand the question whether "conscious awareness is needed for all working memory processes" [41]. Concerning this question also raised by Aru and Bachmann [2] on the temporal dependencies of conscious processing we can say that it depends on a number of things, e.g., if part of the input is not available for conscious perception and if the inputs are ambiguous. This latter may depend on experience-dependent specialization of the input-output 'sensors' that we shortly mention below.

2.1.1 Developmental learning of face processing

It has been well demonstrated that children gradually develop an expertise in face processing and that this learning process goes in the direction of holistic processing [22]. Studies of Meaux et al. [28] show age related changes. Results point to an increased interest in the eye region together with attentional shift from the mouth to the eyes. They suggest that the developmental dynamics is driven by experience-dependent optimization of face processing for improving social and communication skills. One may conclude that experience related changes foster holistic processing of structures, thus shorten processing time and, in turn, more attention can be paid to subtle differences, such as the eye movements in this case.

[2] https://en.wikipedia.org/wiki/Thatcher_effect

3. Methods

In our work we exploit Gestalt Laws in order to resolve potential discrepancies between the outputs of deep neural network 'sensors'. First, we list the Laws that we apply. Then we turn to the description of our deep tools developed ourselves or by other groups. They serve our demonstrations on how to combine the complementing algorithmic components, i.e., the 'sensors' and the Gestalt Law based consistence seeking procedures. We also describe the videos that serve our demonstrations.

3.1. Gestalt Laws considered.

The number of Gestalt Laws (or Gestalt Principles) is numerous. For a compact review, the interested reader is referred to the literature [43]. The five main laws are: (1) Figure-Ground, (2) Similarity, (3): Proximity, (4): Common Fate, and (5): Closure.

Figure-Ground and Similarity Laws are implicit in the supervisory information used for training the networks. For example, the hand detector was trained on a large number of hands, embedded into some portion of the background, since the detector works by means of bounding boxes.

3.1.1 The Proximity Law.

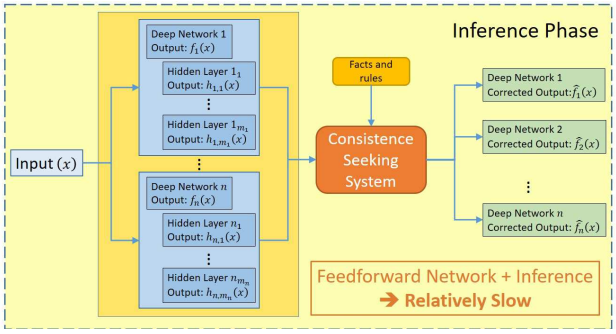
This Law states that elements may be aggregated into a connected (coupled) unit if they are close to each other. We shall use this principle backwards: we know that certain body components such as hand and wrist are joined and will use this knowledge to infer which hand belongs to which wrist, for example.

3.1.2 Common Fate Law.

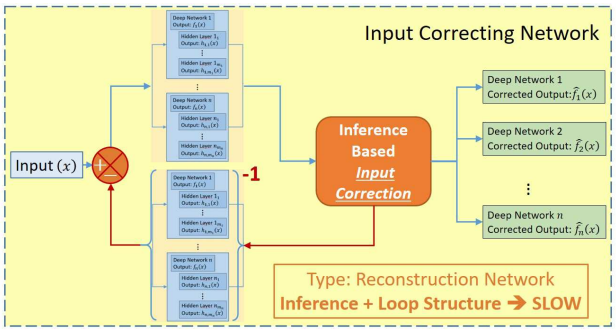
This Law can be phrased as follows: 'What moves together, belongs together' [17]. We will compute the optical flow between frames and will be looking for parts moving together. Optical flow is seen as the motion detector system. We will work on videos that give rise to the 2 dimensional projection of 3 dimensional movements. In such cases, translations, being central features of convolutional network architectures, give rise to simple flows.

3.2. Architectures served by deep networks

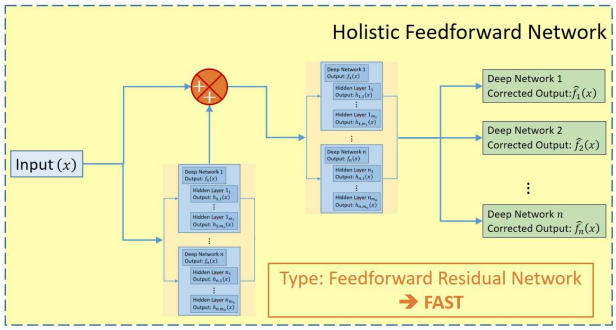
Figure 1 show three architectures. Figure 1(a) represents a number of concurrent feedforward deep networks giving inputs to the cognitive system that corrects the outputs to make them consistent. Hand classifier (left/right) will be the example. On Fig. 1(b) the cognitive system modifies the input in order to make the representation consistent. The driver in the car will demonstrate this case. One can collect training examples in both cases. Training, eventually, the reentrant network of Fig. 1(b) to become a feedforward holistic recognition system comprised of the initial and a correcting network as shown in Fig.1(c). Residual networks [14] and the Pose Machine [46], in particular, are the examples.



(a)



(b)



(c)

Figure 1. Deep architectures. For details, see text.

3.3. Deep architectures exploited

We are to list the deep architectures used in our demonstrations.

Pixels are the outputs of light sensors. Higher order detectors can be trained on pixel regions for signaling for hands, see e.g., [40] and the references therein, and can also work as a class-generic 'object sensor' [1] approximating the probability that an image window contains an object plus a possibly minimal portion of background pixels. Object detectors and object sensors implicitly exploit a number of Gestalt Laws that can be illustrated by investigating different levels of the representations of the deep networks. The list of relevant properties follows:

1. Figure-Ground and Proximity Laws are implicit in the data. We illustrate in Fig. 2 that the ground – in the absence of segmentation – influences the best estimated category of the object embedded in the bounding box (the red rectangle in the figure) showing the great strength of the Proximity Law. Results depicted on Fig. 2 were computed by the Faster R-CNN network trained on the Pascal VOC database [36].
2. These networks are convolutional networks and thus exploit translation invariance justifying the pooling operation and, in turn, combining the Common Fate Law and the Proximity Law.
3. Edges emerge as important feature and they invoke the Good Continuation Law being supported by features of the neuronal substrate, too [3].
4. Closure as well as Past Experience Laws are invoked implicitly by the applied bounding boxes that alleviate closure operation and by the training procedures, respectively.

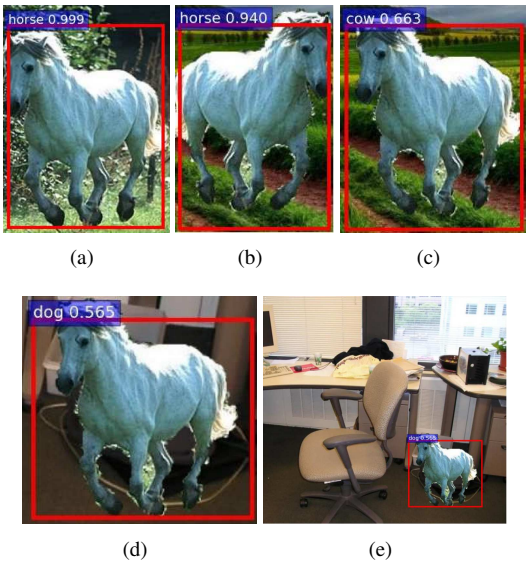
3.3.1 Deep hand detector and handedness classifier

We used Mittal's Hand Dataset[3] [30] for hand detection and augmented it with the VIVA Hand Detection Dataset[4] for left and right hand classification. We employed the Region-based Fully Convolutional Network (RFCN) [8] for the localization of the hands and a vanilla Convolutional Network for classifying if a hand is left or right. The procedure is as follows: (i) find the hand, (ii) classify them.

3.3.2 Deep convolutional pose machine

The pose machine is a great example for Gestalt Laws based training procedures. We note that the pose machine is optimized for outputting the best representation, but the design of the architecture is not made for self-training from the data

[3] http://goo.gl/s2vKoE
[4] http://goo.gl/Lf6mRD



(a) (b) (c)

(d) (e)

Figure 2. Objects and contexts. (a): PascalVOC training sample, (b)-(d): segmented horse placed into different environments. Best guesses are 'horse' when it is placed along a road (b), 'cow' when it is placed across a road (c), and 'dog' when it is in the office at the leg of the table (d), respectively. (e) is the same as (d), but zoomed out. Red: bounding box, white characters in blue background: best network proposals with the corresponding scores. Note that recognition of components such as the mane and the hoof could easily disambiguate classification.

it is collecting. This distinction in the context of neural networks concerns short term memory (the result of the algorithm that produces the representation) and long term memory (the weight tuning procedure). We shortly describe the architecture below. For more details and the software itself[5], please consult the original work [46].

The architecture is a straightforward input-output system. Using our terms, in spite of information fusion about the different components of the body and in spite of resolution seeking the architecture is not a cognitive component, but works as a sensor.

3.3.3 Deep optical flow

Optical flow is a great tool for detecting Common Fate. It estimates the movement vector of each pixel in the full image or in the bounding box. The interested reader is referred to the paper [16] that describes the method of the FlowNet 2.0 software[6], a dense optical flow estimator adopted in our studies.

We are to exploit optical flow to uncover spatio-temporal context that has been shown to play a role in both ambiguous and rivalrous figures [34] to be considered in Sect. 5.

[5] https://goo.gl/gMkoU3

[6] https://github.com/lmb-freiburg/flownet2

3.3.4 The video set

We used a set of 300 videos 'in the wild'[7] to help to demonstrate our thoughts. We integrated information from body pose and from hands. We also used one frame from the State Farm Distracted Driver Scenario[8]

3.4. Algorithmic procedures

In the algorithmic approach, we start by checking for complete consistency in space and time. We run the convolutional pose machine, the hand detector, the left or right hand classifier. The number of hands may differ from frame to frame. Slight changes in the images may change the output of the left or right hand classifier from left to right or from right to left.

We say that a pose is fully consistent in one frame if

1. the pose machine detects the left and the right elbows, the left and right shoulders, the middle shoulder point and the left and right wrists,
2. the hand detectors detect two bounding boxes and the scores of the boxes are above threshold,
3. the bounding boxes have proper sizes according to the elbow wrist distance,
4. the hand detector centers are within proper range to the wrists,
5. the wrists and the nearby hands match in their handedness, and
6. the same conditions hold for the frames before and after the actual frame and this consistency is also supported by the optical flow within the bounding boxes of the hands and the neighborhoods of the points of the poses.

We use optical flow to fix the errors of the left and right hand classifier and to eliminate erroneously detected hands and poses. We manually checked the improvements of this consistence seeking procedure. Note that all elements of the pose, the hand, including handedness that were fixed by consistence seeking can serve further self-training and improve the outputs of the individual units. In turn, consistence seeking improves the representation, i.e., short term memory, whereas such improvements can be used for network training, i.e., for the tuning of the weights of the deep network, i.e., the long-term memory.

4. Demonstrative results

Figure 3 illustrates the first step of the procedure. There are four suspected hands in Fig. 3(a). Two of them are far from the wrists and are removed in Fig. 3(b). Classification

[7] https://ibug.doc.ic.ac.uk/resources

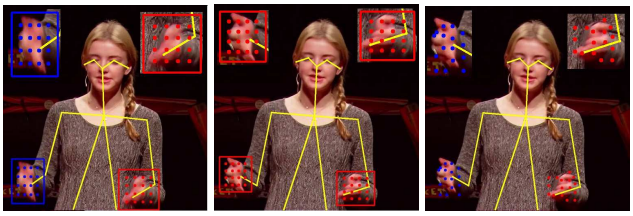
[8] https://goo.gl/Tf6m8A

says that both of them are left hands. One of them is consistent with the left wrist. This one is kept and consistency is shown by the thick magenta line between the wrist and the center of the bounding box of the hand, whereas the other bounding box is removed in Fig. 3(c). Given the *Proximity Law*, it could be relabeled as right hand.

The Common Fate Law is applied in Fig. 4. Figure 4(a) shows a frame where both hands are properly classified. Dots withing the bounding boxes will be followed by the optical flow. In the next frame shown in Fig. 4(b), both hands are classified as left hands. However, optical flow shows that all dots moved only slightly, i.e., they have Common Fate and the classification of the right hand can be fixed (Fig. 4(c)).

(a) (b) (c)

Figure 3. In the preprocessing step, consistency is checked. (a) Some of the hand bounding boxes are close to, others are far from the wrists. (b) Bounding boxes of hand detectors are kept if they are close to the wrist. (c) Handedness for hands and wrists are checked. If they match, then wrist points and the center of the bounding boxes are connected – see inset – otherwise they are dropped. The left hand is '*consistent in space*'.



(a) (b) (c)

Figure 4. Optical flow based extension. A frame is '*fully consistent in space and time*', if the previous frame and the next frame are both consistent in space and if the optical flow between the frames supports the results. (a) Frame is not fully consistent, since hand classification on next frame – i.e., on (b) – shows two left hands. (c) Optical flow starting from the left (red) and from the right (blue) rectangles moves the blue and red points of frame (a) to the similarly colored points of frame (c) respectively. Optical flow is highly trusted and the category (the color of the points) are changed to blue.

The statistics of corrected classifications are shown in number of videos is shown in Table 1. We chose videos of the database having larger number of visible hands. All corrected hands can be used for training the classifier.

ID	Frames	Totally consistent Frames		Ratio	
		before	after	before	after
001	1574	524	1237	0.332	0.785
004	899	233	571	0.259	0.635
009	1822	98	759	0.053	0.416
011	1400	358	914	0.255	0.652
017	1770	76	262	0.042	0.148
019	1800	805	1288	0.447	0.715
031	2050	902	1818	0.440	0.886
041	1800	85	536	0.047	0.297
410	1315	196	701	0.149	0.533
411	1454	246	773	0.169	0.531
511	2324	33	58	0.014	0.024
516	1290	116	665	0.089	0.515
518	2218	157	498	0.070	0.224
538	2163	927	1355	0.428	0.626
553	1624	506	1202	0.311	0.740
Mean	**1700.2**	**351**	**842**	**0.207**	**0.515**

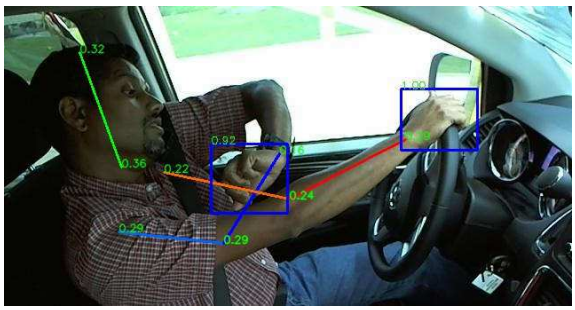
Table 1. Number of totally consistent frames before and after corrections and the corresponding ratios to the total number of frames. We note that not all frames have individuals and not all poses can be made fully consistent.

In the previous cases, we used the pose machine to fix the result of the classification. The case[9] is different for Fig. 5. Now, one of the hand classifiers must be wrong, similarly to the previous example. However, in this scenario, we trust one of our classifiers, since the hand on the steering wheel of a car has a high score and this high score is supported by a large number of samples in the database, due to the high abundances of such situations in driving scenarios. If we take the classification of the right hand as our starting point, then we have to modify the class of the other hand to left. Better consistency can't be reached at the level of the representation. However, the input of the pose machine can be modified. We constrained the left and right wrists to the *proximity* of the left and right hands, respectively and rerun the feedforward pose machine under such constraints. This is a feedback action from the point of view of the ongoing evaluation. Result is shown in Fig. 5(b). Note that we collected two new samples for training, one for the left hand, and one for the pose machine.

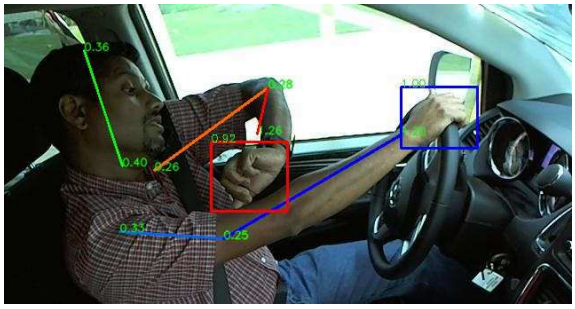
5. Discussion

We started this paper by mentioning consciousness and cognition and suggested to separate inherited or trained feedforward 'sensors' from consistence seeking. Consider Fig. 6. The picture shows some chairs surrounding a table. However, components of texture of the chairs suggest that

[9] A former version of the pose machine was used for demonstration purposes. goo.gl/dQ4Vi6



(a)



(b)

Figure 5. Red (blue) color denotes left (right) components. (a): Pose Machine breaks down for both arms. Both hands are classified as right hands. (b): One hand is on the wheel, its classification is trained and thus trusted in the scenario, the label of other hand is modified to left and these pieces of information are injected into the pose machine giving rise to an improved pose estimation. This way, training examples are also generated for the hand detectors and the pose estimator.

those are soccer balls. Consistence is achieved if we ignore the details of the textures of the chairs and make sense by means of the configuration of the components.

(a) (b)

Figure 6. (a): segmented object, (b): same object in its actual environment

Similar influence may appear for deep neural networks, too as we demonstrated in Fig. 2, where the best scores did depend on small pieces of the environment covered by the bounding boxes.

In turn, we have the following distinct processing steps:

Feedforward processing is the computation accomplished by the 'sensors' (Fig. 1(c))

Label correction in classification is part of the making sense procedure. It is driven by higher order rules experienced or prewired, e.g., that people have two hands and they can't be both left or both right hands (Fig. 1(a)).

Input modulation is the result of unresolvable conflicts in the representation that can be fixed by deciding which part taken for granted and to modify input-output processing by means of this decision. The example is the constraint or modulation on the feedforward processing of the pose machine (Figs. 1(b) and 5).

Feedforward processing is fast, especially if no label correction is needed. *Label correction* can be both feedback and feedforward. It is feedback if units can overwrite each others' output, like in competitive neural networks. For an early work, see [13] on this subject. *Label correction* is feedforward if different information fragments are fused at a next (feedforward) stage. Fig. 5 is an example.

Having this in mind, we shortly explain the essence of our computer studies. Then, we turn to experimental findings related to consciousness and cognition.

5.1. Computer studies

In our computer studies, we focused on the certainty and the consistency of the observations and propagated the information forward and backward in time. This is an oversimplified model of the integration time of conscious interpretation that can (should) cover both the past and the observation based prediction in the future.

We did not include alternating rivalrous interpretations. In our framework such alternations call for approximate input generation from the representation that autoencoding deep networks are capable of [45] that we didn't model here, but refer to the literature [25].

We illustrated how to resolve the discrepancy at the level of the representation by relying on one part of the output and constraining the input of feedforward units accordingly.

As a result, input and representation make sense *together* and the infinite regress of the homunculus fallacy is turned to a self-correcting feedback based loop structure made of feedforward processing units.

The deterministic world assumption motivates one to use the most reliable information for both input and output modification. Determinism is also exploited via optical flow.

5.1.1 Deep networks

We used a number of trained deep networks. Such networks are feedforward algorithms and they perform holistic recognition of those objects or episodes that they were trained for.

For example, components of the hand are evaluated implicitly by the hierarchy of filters of the convolutional layers of the hand detector. The case of the left and right hand classifier is similar. However, the pose machine is somewhat different, since different components of the body are processed differently and this additional information is available during the training process. Processing is still feedforward and thus – from our point of view – the full network is a (sophisticated) sensor.

Consistence seeking procedure could generate a number of new training samples for the 300 VW database by relabeling the outputs of the classifiers. Relabeling is a competitive mechanism at the level of the representation, similar to some extent to the genuine interactive activation model [26]. Such *collaboration* between deep network components has additional advantages, since 'deep neural networks are easily fooled' [31]. For example, face recognition can be fooled by adding special glasses designed for change the class label [38]. However, if the presence of glasses is detected then they can be removed and inpainting can fill in the missing pixels [32].

In the car example, training samples were generated, too: the representation modified the label of one of the hands and then it constrained the input of one of the pose machine 'sensor'. This computational procedure acts top-down and resembles to goal-driven orienting of attention[10]. However, the algorithmic method does not require conscious observation.

We distinguish the following processes:

1. 'Sensory' feedforward processing
2. Checking for spatio-temporal consistency
3. Correcting the outputs in case of inconsistencies
4. 'Making sense' of the representation by constraining the inputs.

5.2. Making sense and controlling

We targeted prediction in a limited way, we used optical flow alone. Prediction can be learned and it means spatio-temporal model construction. One may say the following: the input and the representation make sense if the model (i.e., the representation) can predict the input.

We propose a four step consistency seeking procedure for sensing, processing, modulating both outputs and inputs, and thus making sense of information arriving from the environment. Processing delays should be compensated in order to launch appropriate control instructions to muscles that will be observed and perceived later. Furthermore, conscious observation in a deterministic world prescribes a single interpretation.

We suggest that component(s) with highest confidence(s) might influence the conscious making sense process more than it is justified by the unconscious Bayesian optimization of the internal model of the environment [33]. Considering the driving case, the hand class estimation followed by top-down constraints on pose estimation combines rules with uncertain observations and unfolds the interpretational puzzle. This procedure generates training samples in a straightforward manner in a deterministic world provided that consistency can be achieved. The global-to-local constraints on awareness seem to optimize the context based consistence seeking procedure [7] and thus self training, whereas Bayesian procedure can be at work at early processing levels having high noise content [33].

The requirement of a single interpretation is a challenge since there are delays both in sensory information processing, motor command launching, but they should be matched in time. The full delay in this loop launching and sensing the motor command can be about 500 ms or more. Improper matching may give rise to subtle experiences, an intriguing example being that schizophrenic patients can tickle themselves [23], i.e., the tactile input is unexpected, or as one may say, 'it doesn't make sense'. Also, the requirement of single interpretation together with the approximately 500 ms processing delays seem to be in agreement with the switching time of perception for the case of rivalrous figures.

6. Conclusions

We proposed an algorithmic procedure that (i) combines deep learning methods by means of Gestalt principles, (ii) emphasizes the deterministic nature of the world, (iii) searches for consistent interpretation at all times (iv) using both spatial and temporal contexts.

The algorithmic procedure is a model for perception and cognition, it offers a resolution for the homunculus fallacy, emphasizes model based prediction to compensate for delays of the processing of sensory information, the delays between the launching of motor commands and the start of the movements.

The combination of component based and holistic recognition may alleviate the problem of deep neural networks known to be vulnerable to specially designed spurious inputs.

The procedure may fail in each proposed component giving rise to problems similar to those found in human behavior, including the selection of a single and consistent representation, the putative problem in schizophrenia [23].

Acknowledgment: This work was partially supported by the EIT Digital Grant (No. 14386) on Cyber-Physical Systems for Smart Factories.

[10] http://goo.gl/NJ2Gts

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE PAMI*, 34:2189, 2012. 4

[2] J. Aru and T. Bachmann. In and out of consciousness. *Frontiers in Psych.*, 8, 2017. 2, 3

[3] O. Ben-Shahar and S. Zucker. Geometrical computations explain projection patterns of long-range horizontal connections in visual cortex. *Neural Comp.*, 16:445–476, 2004. 4

[4] I. Biederman. Recognition-by-components. *Psych. Rev.*, 94:115, 1987. 1

[5] A. Bognár, G. Csete, et al. Transcranial stimulation of the orbitofrontal cortex affects decisions about magnocellular optimized stimuli. *Frontiers in Neurosci.*, 11, 2017. 2

[6] L. Boutsen, G. W. Humphreys, et al. Comparing neural correlates of configural processing in faces and objects. *NeuroImage*, 32:352–367, 2006. 3

[7] F. Campana, I. Rebollo, et al. Conscious vision proceeds from global to local content in goal-directed tasks and spontaneous vision. *J. Neurosci.*, 36:5200–5213, 2016. 2, 8

[8] J. Dai, Y. L. Li, et al. R-FCN. *arXiv:1605.06409*, 2016. 4

[9] P. A. Davis. Effects of acoustic stimuli on the waking human brain. *J. Neurophys.*, 2:494–499, 1939. 3

[10] D. Dennett. *Consciousness Explained*. Little Brown and Co, New York, 1991. 1

[11] V. Di Lollo, J. T. Enns, et al. Competition for consciousness among visual events. *J. Exp. Psych.*, 129:481, 2000. 2

[12] T. W. Drew, A. W. McCollough, and E. K. Vogel. Event-related potential measures of visual working memory. *Clinical EEG and Neurosci.*, 37:286–291, 2006. 3

[13] P. Földiák. Forming sparse representations by local anti-Hebbian learning. *Biol. Cyb.*, 64:165–170, 1990. 7

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[15] M. H. Herzog, T. Kammer, and F. Scharnowski. Time slices. *PLoS Biology*, 14:e1002433, 2016. 2

[16] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0. In *CVPR*, 2017. 5

[17] W. Köhler. Gestalt Psychology. *Horace Liveright, N.Y.*, 1929. 3

[18] M. Koivisto. Is reentry critical for visual awareness of object presence? *Vis. Res.*, 63:43–49, 2012. 2

[19] M. Koivisto, I. Harjuniemi, et al. Transcranial magnetic stimulation of early visual cortex. *NeuroImage*, 2017. 2

[20] M. Koivisto and J. Silvanto. Visual feature binding. *NeuroImage*, 59:1608–1614, 2012. 2

[21] P. S. Laplace. *A Philosophical Essay on Probabilities*. Dover Publications, N.Y., 1951. 1

[22] K. Lee, P. C. Quinn, et al. Development of face-processing ability in childhood. In *Dev. Psych.*, volume 1, chapter 12, pages 338–370. Oxford Univ. Press, 2013. 3

[23] A.-L. Lemaitre, M. Luyat, et al. Individuals with pronounced schizotypal traits are particularly successful in tickling themselves. *Consc. Cogn.*, 41:64–71, 2016. 8

[24] D. A. Leopold and N. K. Logothetis. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature*, 379:549, 1996. 2

[25] A. Lőrincz, B. Szatmáry, et al. The mystery of structure and function of sensory processing areas of the neocortex: a resolution. *J. Com. Neurosci.*, 13:187–205, 2002. 1, 2, 7

[26] J. L. McClelland and D. E. Rumelhart. An interactive activation model. *Psych. Rev.*, 88:375, 1981. 8

[27] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976. 2

[28] E. Meaux, N. Hernandez, et al. ERP and eye tracking evidence of the developmental dynamics of face processing. *Eur. J. Neurosci.*, 39:1349–1362, 2014. 3

[29] B. A. Metzger, K. E. Mathewson, et al. Regulating the access to awareness. *J. Cogn. Neurosci.*, 2017. 3

[30] A. Mittal, A. Zisserman, et al. Hand detection using multiple proposals. In *BMVC*, pages 1–11, 2011. 4

[31] A. Nguyen, J. Yosinski, et al. Deep neural networks are easily fooled. In *CVPR*, pages 427–436, 2015. 8

[32] A. v. d. Oord, N. Kalchbrenner, et al. Conditional image generation with PixelCNN. *arXiv:1606.05328*, 2016. 8

[33] G. Orbán, P. Berkes, et al. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92:530–543, 2016. 1, 2, 8

[34] M. Ouhnana, B. J. Jennings, et al. Common contextual influences in ambiguous and rivalrous figures. *PloS One*, 12:e0176842, 2017. 2, 5

[35] D. H. Rakison and L. M. Oakes. *Early category and concept development*. Oxford Univ. Press, 2003. 1

[36] S. Ren, K. He, et al. Faster R-CNN. In *NIPS*, pages 91–99, 2015. 4

[37] M. Salti, Y. Bar-Haim, et al. The P3 component of the ERP reflects conscious perception, not confidence. *Consci. Cogn.*, 21:961–968, 2012. 3

[38] M. Sharif, S. Bhagavatula, et al. Accessorize to a crime. In *CCS*, pages 1528–1540, 2016. 8

[39] J. Silvanto, N. Lavie, et al. Double dissociation of V1 and V5/MT activity in visual awareness. *Cereb. Ctx.*, 15:1736–1741, 2005. 2

[40] A. Sinha, C. Choi, et al. Deephand. In *CVPR*, pages 4150–4158, 2016. 4

[41] D. Soto and J. Silvanto. Reappraising the relationship between working memory and conscious awareness. *TICS*, 18:520–525, 2014. 3

[42] M. Steyvers, M. D. Lee, et al. A Bayesian analysis of human decision-making on bandit problems. *J. Math. Psych.*, 53:168–179, 2009. 1

[43] D. Todorovic. Gestalt principles. *Scholarpedia*, 3:5345, 2008. revision #91314. 3

[44] P. Vakli, K. Németh, et al. The electrophysiological correlates of integrated face and body-part perception. *Quarterly J. Exp. Psych.*, 70:142–153, 2017. 3

[45] P. Vincent, H. Larochelle, et al. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008. 7

[46] S.-E. Wei, V. Ramakrishna, et al. Convolutional pose machines. *arXiv:1602.00134*, 2016. 4, 5

[47] H. Wiese, J. M. Kaufmann, et al. The neural signature of the own-race bias. *Cereb. Ctx.*, 24:826–835, 2012. 3

[48] Y. Zhang, X. Zhang, et al. Misbinding of color and motion in human early visual cortex. *Vis. Res.*, 122:51–59, 2016. 2