Awakening a taxonomist’s third eye: exploring the utility of computer vision and deep learning in insect systematics

Miroslav Valan¹,²,³, Dominik Vondráček⁴,⁵ and Fredrik Ronquist²

¹Savantic AB, Stockholm, Sweden, ²Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden, ³Department of Zoology, Stockholm University, Stockholm, Sweden, ⁴Department of Zoology, Faculty of Science, Charles University in Prague, Prague, Czech Republic and ⁵Department of Entomology, National Museum, Prague, Czech Republic

Andrew Ng, co-founder of Coursera, founder of Google Brain and former Baidu Chief Scientist, has argued that artificial intelligence (AI) is ‘the new electricity’ (Ng, 2016). In the early 20th century, electrification transformed transportation, agriculture and manufacturing, permanently changing human societies. A century later, we can see how AI is starting to have a similar, revolutionary impact on a range of societal sectors, from finance to the automotive industry. AI is used everywhere: from product-recommendation systems to search engines and virtual voice assistants. In fact, we encounter AI on a daily basis, often without even realizing it. Everyone knows that self-driving cars are powered by AI, but it is less commonly acknowledged that manual operation of a vehicle now typically relies on AI-powered navigation, path optimization and travel time estimation. The AI revolution is generating interest in a wide range of fields, and systematic entomology is not an exception.

An important reason for the current AI hype is the recent progress in computer vision made possible by the development of convolutional neural networks (CNNs) that learn complex tasks from very large training sets, that is, through deep learning. Deep learning of complex visual tasks by CNNs would not have been possible without significant recent improvements in learning algorithms, easy access to big data (>80% of the data on the web is visual, consisting of videos or images), and the availability of cheap and fast computation (particularly the development of graphical processing power driven by the gaming industry). Given enough training data, CNNs can learn on their own which features to use to discriminate between different categories of images (often referred to as image classification). Even more astonishing, once these features are learned for one task, the CNN can use its feature-recognition skills in tackling a new but related task, thus eliminating the need to train a dedicated CNN from scratch for every new task (Azizpour et al., 2016; Donahue et al., 2014; Z. Li & Hoiem, 2018; Sharif Razavian et al., 2014; Zheng et al., 2016). For instance, a CNN that can distinguish between cats, dogs and bicycles can more easily learn to distinguish between different species of insects (Valan et al., 2019). This is particularly important for visual identification tasks where it is difficult to gather sufficiently large image sets to train dedicated CNNs from scratch, such as most insect identification tasks.

Insect systematics is to a great extent based on visual data, and it is increasingly clear that we can use state-of-the-art AI techniques like CNNs and deep learning to automate many routine identification tasks that are currently performed by humans – without compromising identification accuracy (Marques et al., 2018; Rodner et al., 2015; Valan et al., 2019; Wei et al., 2017). The aim of this article is not to show how CNNs perform compared to humans or traditional machine learning approaches on yet another image classification task. Instead, we focus on some questions that are frequently asked by members of the systematic entomology community, who wonder about the feasibility of applying these new AI techniques to their tasks of interest. We will address these questions in the form of an imagined discussion between a taxonomist and an AI expert focused on a case study, namely the discrimination of ten flower chafer species of the beetle genus Oxylabrus Mulsant. This pilot example not only allows us to illustrate several important points, but also to explore questions and ideas that are relevant to insect identification in particular. We hope that the discussion itself, as well as some of the new results we present here, will encourage more entomologists to enrich their toolbox with these easy-to-learn instant-return AI techniques.

Off-the-shelf solutions are often sufficient

Taxonomist: Developing automated systems for insect identification has proven challenging in the past. What makes these new methods different?

Correspondence: Miroslav Valan, Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Frescativagen 40, 114 18 Stockholm, Sweden. E-mail: miroslav.valan@nrm.se

© 2021 The Authors. Systematic Entomology published by John Wiley & Sons Ltd on behalf of Royal Entomological Society.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
AI expert: Previous systems had to be engineered for a special task. For instance, landmarks in the images had to be identified prior to analysis, or the systems had to be tailor-made to identify salient features for the task at hand. The new methods learn the task from scratch, including the identification of important discriminating features, entirely on their own. The only thing that is required is a large set of images with known identities, which can be used in training the systems. Automated identification systems developed for large citizen science projects — like iNaturalist, Merlin-Bird (for e-Bird) or Pl@ntNet — exemplify the new approach well. Over time, the citizen scientists engaged in these projects accumulate large sets of images, many of which are reliably identified. These huge sets of identified images provide ideal data for training AI models for automated taxon identification. The training of the CNN systems on these data sets requires considerable computational resources. However, the systems that are produced are optimized for mobile devices, with a small memory footprint and fast inference times, which comes at the expense of accuracy. This means that it would be possible to train larger, deeper and more complex CNN systems that would perform even better in terms of accuracy than these systems, which already have impressive performance.

Taxonomist: So, if I wanted to apply these techniques to an insect identification task I am interested in, would I need a huge training set and vast computational resources?

AI expert: No, I do not think so. These super systems are impressive, but developing them requires a fair amount of skills and resources. For most identification tasks, off-the-shelf solutions are quite sufficient. With an ‘off-the-shelf solution’, I mean a CNN that has been trained on a generic image classification task. This CNN can be used for feature extraction, and the features then used in training a simpler identification system (Azizpour et al., 2016; Donahue et al., 2014; Sharif Razavian et al., 2014; Zheng et al., 2016) for a more specialized task, such as insect identification (Valan et al., 2019).

In this context, off-the-shelf means that the AI software is not developed for a single, specialized task. The software is ready to use without modifications on a different task, bypassing the need for machine learning expertise to tune the software’s hyperparameters.

Taxonomist: What if we have some (let us say 1%) misidentified specimens in the training data? Can we build models robust enough to deal with that?

AI expert: CNNs are clearly robust enough to deal with low levels of label noise. For instance, ImageNet, a cornerstone benchmark of recent image classification systems, has 0.3% label noise (Russakovsky et al., 2015) and similar noise levels (0.2%) were shown to have no adverse effect in some experiments on insect identification (Valan et al., 2019). Of course, training sets on cryptic species may be considerably noisier. In fact, in those cases, we may not even be able to determine the noise level accurately after the training set was collected because, unlike the cases cited earlier, visually inspecting images of cryptic species would probably not suffice to say whether the specimen was correctly identified.

Learning from noisy data is an active area of research, and studies have shown that it is sometimes possible to successfully combat up to 50% noise (Azadi et al., 2015; Lee et al., 2018; Y. Li et al., 2017; Patrini et al., 2017; Tanaka et al., 2018; Veit et al., 2017; Xiao et al., 2015; Zhuang et al., 2017). However, adopting a specialized method for noisy data is probably not necessary for most taxonomic identification tasks. After all, natural history collections are assembled by experts, and we might expect the percentage of misidentifications to be low enough in most cases that they cause no problems for standard CNNs.

Taxonomist: How about cryptic species or other challenging tasks? Is it possible to train an AI algorithm to perform tasks that humans cannot do? If so, would an off-the-shelf solution be sufficient then?

AI expert: To my knowledge, nobody has tried to use CNNs to separate cryptic species. A dedicated CNN trained specifically for such a task would probably have a good chance of succeeding, but even an off-the-shelf approach might generate enough insights to decide whether automated identification is feasible. In fact, the biggest challenge here is to collect enough data, not to select the best algorithm, unlike what many people think (Fig. 1). If I understand correctly, identifying challenging taxa requires careful examination of each specimen, possibly involving dissection of genitalia or even DNA sequencing. Therefore, assembling a sufficiently large set of correctly identified images to train a dedicated CNN will be tedious and demand more resources than is probably available in most cases. Clearly, we cannot rely on standard citizen science projects to accumulate appropriate training sets for cryptic species, because the rapporteurs are unlikely to have the required skills nor the possibility of examining the specimens in sufficient detail to arrive at a correct identification. To get a better understanding, why do we not pick a task you are interested in and try an off-the-shelf solution (see Methods section, Data S1 for details) that requires no particular technical skills (machine learning or programming) and see what we can accomplish?

A case study

Taxonomist: Let us develop an identification system for the species of the flower chafer beetle genus Oxythyrea. I have worked with this genus for a long time, so I know it well. It comprises ten species (Bezděk, 2016), which are quite similar in habitus (Fig. 2), although they are well defined by a set of morphological features, including the dorsal and ventral patterns of white spots, as well as the shape of the male aedeagus and endophallic sclerites (Baraud, 1985, 1992; Mikšič, 1982; Sabatinelli, 1981, 1984). Species are also genetically distinct (Dominik Vondráček, unpublished data). To separate them based on morphology, the dorsal habitus is slightly more helpful than the ventral habitus, but both are needed for reliable identification. In some cases, it is also necessary to have access to locality, collecting date and biotope information. For instance, some Oxythyrea funesta (Poda), Oxythyrea pantherina (Gory & Percheron), and Oxythyrea subcalva Marseul specimens from North Africa can only be identified reliably if you have access to information about the biotope where they were collected,
Fig. 1. Many people think that the major challenge in developing AI systems, like systems for automated taxon identification, is coding (left). In reality, the major challenge is to assemble appropriate training data (right). (Image courtesy: Hackernoon). [Colour figure can be viewed at wileyonlinelibrary.com].

Fig. 2. Datasets of ten Oxythyrea species. In the first row, we show example images of dataset B collected in a ‘quick and dirty’ way using a smartphone and a cheap $2.00 attachable zoom lens. The remaining four rows are example images from dataset A – standardized taxonomic imaging setting, where the first two rows are dorsal habitus and the last two are ventral habitus. Note that Oxythyrea beetles show sexual dimorphism only on their abdomen. More about specimen acquisition, identification and imaging can be found in Materials and Methods, Data S1. [Colour figure can be viewed at wileyonlinelibrary.com].
especially if they are females so that you cannot examine the shape of the aedeagus. *Oxythyrea funesta* is tied to humid, *O. pantherina* to semi-arid and *O. subcalva* to arid biotopes. Specimens from Europe are easier to identify because *O. funesta* is the only species in this group that occurs there, as far as we know. Another challenging set of species is *Oxythyrea dulcis* Reitter and *Oxythyrea abigail* Reiche & Saulcy. They are very similar but geographically well separated, which facilitates identification. The morphology of the male genitalia is also quite different, especially the shape of the endophallic sclerites.

**AI expert:** As a baseline, let us try the dorsal view only without location information or any other information that could facilitate identification. These constraints should make the task more challenging for the AI system. We can add information from the ventral habitus later and see if it improves the performance of the algorithm.

**Taxonomist:** How many images do we need for the training set? And how about image quality? Do we need high-resolution images?

**AI expert:** A few dozen images per taxon is usually enough for a proof-of-concept system. Let us aim for that. High-resolution images are not required, and although it could help, my guess is that it would not have a significant impact if we chose low-resolution images instead. Due to the heavy computation involved in training, almost all state-of-the-art CNNs by default resize input images to 299 × 299 or 224 × 224 pixels (1000 × 1000 + pixels is considered high resolution here). Regarding the number of images, you could in theory start with as little as one training example per category (one-shot learning) and still get a reasonably high-performing CNN model. With only one image per category, however, the model might struggle when test images deviate from the training images in ways that are not relevant for identification. It would also be difficult to gain insights into why, and how the model fails with such a small training and testing set.

When building a CNN model for taxon identification, the performance will depend to a large extent on how the categories and the training examples are distributed in morphological space. Individuals within the same species can vary due to sex, colour morph, life stage, and so on. A small training set may not contain all the major variants, in which case performance will suffer. With more images you increase the likelihood that the relevant morphs are seen in the training phase. By increasing the sample size, it is also easier for the CNN model to form high-level abstractions, ‘concepts’ if you wish, which define the type of variation one might expect to see among individuals belonging to the same species, and between individuals belonging to different species. For instance, specimens preserved or photographed with a certain technique, say a blue or white background, may stand out from all other specimens, but this pattern may be irrelevant for separating the species. Certain morphs may also be more challenging to identify than others, and reliable discrimination may require special features specific for those morphs. An example would be males and females; it may well be that males have to be separated based on other features than females.

In *Oxythyrea*, can you distinguish the sexes easily? If so, which sex is more challenging to determine to species, males or females?

**Taxonomist:** There are sex-specific differences in the white pattern on the ventral side, which makes it easy to identify the sex, but not in all ten species. Only in six of them, this works. For the other four species, you need to look for the specific tiny groove on the abdomen, which is present only in males, but very difficult to find. Usually, it is necessary to examine the abdomen from different angles using a stereomicroscope to be sure. The dorsal habitus in these ten species does not exhibit any sexual dimorphism; at least, sexual dimorphism has not been observed nor hypothesized in the literature.

**AI expert:** Let us use this then to test whether an off-the-shelf system can do something that is impossible or at least very difficult for humans. It can serve as a model for the potential to identify cryptic species with off-the-shelf solutions.

**Taxonomist:** I do not know if this is even feasible but let us try and see what happens. Imagine how helpful it could be if we succeeded. You have hundreds, maybe thousands of these beetles glued on cards across many collections in European museums and you are not able to unglue them because it is typically impractical or forbidden. You have no idea whether knowing the sex may be crucial for many types of analyses, such as standard morphometric analyses. Furthermore, say you are allowed to unglue a few select specimens to examine the male genitalia, which includes preparation of the aedeagus. If your analyses could separate the sexes based on dorsal habitus alone, then only male specimens would have to be unglued and the females could be left untouched, instead of randomly ungluing specimens until you find a male. So you will not potentially damage the females during the process of ungluing, which can happen sometimes.

This all sounds very interesting. I will start collecting the images. I will go for high-resolution photos using my standard imaging setup.

**AI expert:** Have you assembled the images yet?

**Taxonomist:** No. I can only do around 50 specimens per day for both dorsal and ventral views, if I focus entirely on this task. Unfortunately, I have other things to do as well, so I will need another week or so. My top pace may not be sustainable in the long term anyway if I considered a more ambitious project comprising more species. And the cost is substantial. I saw a study of a mass digitization project indicating that an average staff member can generate 40 images per day, resulting in a labour cost of €3.99 per image (Tegelberg et al., 2012).

**AI expert:** With that approach, it will take ages and cost a fortune to digitize all the collections across the globe! I need some sample images to start experimenting, so I will just photograph some *Oxythyrea* specimens in a ‘quick and dirty’ way with a smartphone and an attachable $2.00 zoom lens (Fig. 2B). We can compare the results obtained with these images to the results obtained with your high-resolution images.
Awakening taxonomist’s third eye

Fig. 3. Confusion matrices for different imaging techniques and image combinations. We show results when high-resolution images of dorsal (top left) and ventral habitus (bottom left) are used alone; when features from dorsal and ventral habitus are concatenated (top right) and when features are obtained from smartphone images of the dorsal habitus (bottom right). Rows show true categories and the number of images per species (N) and columns show predicted categories. Colours and numbers in the matrix indicate the proportion of predictions normalized so that the sum in each row is equal to 1. (abi = O. abigail, alb = O. albopicta, cin = O. cinctella, dul = O. dulcis, fun = O. funesta, gro = O. gronbechi, noe = O. noemi, pan = O. pantherina, sub = O. subcalva, tri = O. tripolitana). [Colour figure can be viewed at wileyonlinelibrary.com].

Baseline

AI expert: [another week later …]. Now I have the results from the baseline study based on your high-resolution images (see Methods for details, Data S1). I will summarize the performance of the AI system in the form of a confusion matrix, which specifies how often different categories are confused by the identification system (Fig. 3).

Taxonomist: Wow, these results already surpass all my expectations! Human experts (there are only a handful of us) would never have been able to match those numbers if the identification had to be based on dorsal habitus only (Fig. 3, top left). Also, the identification accuracy based on ventral habitus would be very difficult to beat for human experts.

It is interesting that the AI system finds the dorsal view (95%) better than the ventral (88%) for species identification; it is the same for human experts. However, it is surprising that the combination of both views does not boost the performance of the AI system. Actually, the accuracy got slightly worse for some species (e.g. O. pantherina and O. subcalva). Could this be some sort of artefact?

AI expert: It is true that the accuracy decreases for some species but note that it increases for others. In total, two more images were correctly identified compared to the best of the single-view datasets, that is, a slight improvement. However, this improvement is in the range of ‘being lucky’. One possible reason for the lack of a significant boost in identification accuracy is that we achieved high accuracy already with dorsal habitus alone (95%). For further improvement, we would have needed a bit better features or performance from the other habitus. However, the ventral habitus had 3x higher error rate than the dorsal habitus. Combining images of different views would likely bring more gain when the difference in error rate is not as high as in this case, when you have similarly performing views, or when you have more limited training sets (see Fig. 4).

Taxonomist: I noticed that O. abigail, Oxythyrea tripolitana and O. funesta seem to be the most challenging for the AI system to identify in both views. The first two are likely challenging because we photographed so few specimens; these are rare species and there are not many specimens available in collections. The third, O. funesta, is indeed hard to separate from O. pantherina and O. subcalva also for human experts. It is interesting that the model also often confuses these three species in both views. Four out of six erroneous predictions based on dorsal habitus mistake one of these species for one of the other two. The challenge of separating these species is even more apparent for the ventral habitus, where there are 15 such cases out of 17 total misidentifications.

© 2021 The Authors. Systematic Entomology published by John Wiley & Sons Ltd on behalf of Royal Entomological Society. doi: 10.1111/syen.12492
O. pantherina and abigail reached based on high-resolution images still apply. Moreover, the conclusions we obtained for dorsal habitus images (94.6 vs 95.2% identification accuracy for a CNN model to match the performance we obtained for experts but surprisingly they seem to be sufficient enough for a smartphone approach. AI expert: Now, let us look at the results obtained with my smartphone images. 

**Lab-based versus smartphone photos in the field**

*AI expert:* Now, let us look at the results obtained with my smartphone images.

*Taxonomist:* The phone images are blurry and hardly usable for experts but surprisingly they seem to be sufficient enough for a CNN model to match the performance we obtained for experts but surprisingly they seem to be sufficient enough for a smartphone approach.

**Male versus female – a task unsolvable by humans**

*AI expert:* Let us now look at the task of separating males from females.

*Taxonomist:* I assure you, it is impossible for experts to identify the sex from dorsal view, even when using geometric morphometrics (Dominik Vondráček, unpublished data using O. dulcis as a model organism). There is no evidence in the literature that anyone noted or hypothesized sexual dimorphism in the dorsal habitus of any of the Oxythyrea species. Yet, according to your results (Fig. 5), such differences do exist at least in some species (Oxythyrea cinctella (Schaum), O. dulcis and O. tripolitana). An expert can identify the sex from the ventral habitus mainly because males and females have different patterns of white marks in many species (Fig. 2). For the species without colour dimorphism, it is necessary to identify the presence or absence of a very thin groove in the medial part of the abdomen, going from the first segment to the fourth (similar in position to the four white spots in O. funesta; Fig. 2). The groove is clearly seen if you look at the abdomen from the side and with appropriate illumination, but it is hardly visible in the ventral habitus images we collected. This is because we used a strict ventral view, and the lighting we used made the groove difficult to distinguish in the shiny midpart of the abdomen. Therefore, it makes sense that the model had difficulties in some cases; for instance, the identification of sex from ventral habitus images of Oxythyrea gronbechi Petrovitz was close to a random guess.

Another interesting pattern that I noted given these results is the case of Oxythyrea albobracta (Motschulsky) and Oxythyrea noemi Reiche & Saulcy, where the model seems to be able to tell whether a specimen is a male but for female specimens it predicts sex close to the rate of a random guess. This suggests to me some form of sex-limited polymorphism, that is, the fact that females exhibit polymorphism but males do not. It is quite remarkable that in some cases, the machine was able to distinguish both sexes only from the dorsal view, but in some similar cases, it was not able to do so, for example, O. cinctella versus O. gronbechi, which are very closely related species with similar type of sexual dimorphism, at least as far as a human expert can tell. Overall, it is still very difficult for the machine to sort these specimens into males and females from only the dorsal view (e.g. O. pantherina, O. subcalva).

*AI expert:* ‘Absence of evidence is not evidence of absence’. If all it takes to tell the sex of a specimen is to turn it upside down, and the sex dimorphism is not obvious from the dorsal side, then that is what most people will do – turn the specimen around. So maybe it is not that surprising that the sex dimorphism in dorsal habitus has been overlooked in this group. However, your...
Fig. 5. Confusion matrix for identification of species and their sex. Experiments are based on either dorsal (left) or ventral habitus (right) alone. [Colour figure can be viewed at wileyonlinelibrary.com].

A morphometric study indicates that it is actually quite challenging to find what the dimorphic features are. I guess one of the most frustrating aspects of difficult discrimination tasks is that it can take an expert many, many hours to find pertinent differences, and there is no guarantee that they exist. Now we have a tool that could give us some insight as to whether it is worth looking for those differences, and from which specimens we should start. We can assume that adding data and/or training a dedicated model would provide more information on whether discriminating features exist and where they can be found.

Beyond off-the-shelf solutions

**Taxonomist:** What sort of improvements can we expect with systems that go beyond off-the-shelf approaches?

**AI expert:** First of all, it is highly likely that we could improve the performance in terms of accuracy. In fact, I ran some experiments to verify that this is the case for the problems we looked at (see Figs 6, S1 and S2 in the supplementary material). However, due to the improved accuracy, some of the insights we gained from our simpler experiments might be hidden. For instance, if only one to two specimens are misidentified, the confusion matrix no longer contains any useful information about which species are more challenging for the system to identify, as the misidentified species could be coincidental.

I want to stress that off-the-shelf systems are often sufficient to solve even challenging tasks that appear unsolvable to humans, as we have seen above. Few entomologists will have access to an AI expert; even machine learning labs struggle currently to attract relevant AI expertise. Fortunately, however, it is perfectly feasible for entomologists to gain the necessary expertise to run off-the-shelf systems on their own. Even if your goal is to develop a species identification tool that is as accurate as possible, first exploring an off-the-shelf approach might generate enough insight to help you decide whether the task is feasible and to what extent further investment of time and resources is justifiable.

Having said this, there are cases when a dedicated CNN is clearly preferable. For instance, with such an approach you are more likely to compute accurate activation maps, that is, maps that pinpoint the image areas that are used by the system to discriminate among categories.

**Taxonomist:** So this technique can be used to identify which parts of *Oxythyrea* specimens are important in separating species? It would be interesting to see whether the machine is
Fig. 7. Class activation maps for ten *Oxythyrea* species. This technique highlights category-specific regions of images by producing heat maps. The select specimens are exemplary representatives of respective species. [Colour figure can be viewed at wileyonlinelibrary.com].

using the same or at least similar morphological features to those that taxonomists find useful.

*AI expert:* Yes, we can prepare ‘heat maps’, which show the areas in the image on which the machine is focusing. I generated such heat maps for the dorsal and ventral views from the high-resolution images, as well as the dorsal views from the cell phone pictures (Fig. 7).

*Taxonomist:* It seems that the machine is focusing on the same morphological details as expert taxonomists. The most informative part in the dorsal habitus view is the pattern of white marks on the pronotum, and we can see that the machine is focusing there for almost all species. For the easiest species, it is the only part (e.g. *O. albopicta*, *O. cinctella*); apparently, this region contains enough information to distinguish them. When it gets more difficult, the machine starts using a wider region, including the distal part of the elytra (e.g. *O. dulcis*, *O. noemi*), to separate the species. For the most difficult species (e.g. *O. abigail*, *O. funesta*, *O. subcalva*), the machine looks at the whole elytra region. Interestingly, for *O. tripolitana* the machine mostly focuses on the part between the pronotum and scutellum, where there is an accumulation of shorter setae in this species. This has not been an important character for taxonomists, at least not until now. For *O. gronbechi*, the machine was checking not only the pronotum but the pin as well. This appears to be a case where the machine is misled by biases in the training set. All specimens of *O. gronbechi* were collected a long time ago by an entomologist who pinned them from a different angle than is usual. Of course, this means that the model we trained for this task might have problems with specimens of this species that are pinned in the standard way.

In the smartphone pictures, the machine is using much wider regions for species discrimination. Although these pictures are enough for the machine to provide accurate identifications, as we have seen, it apparently needs to check wider regions of the images to be sure of the correct identification.

For the ventral habitus, it seems that the machine is using the amount of setation, the presence and size of the white spots on the sides of the abdomen and probably also the shape of the mesometaventral processus. The latter is typical for the subfamily Cetoniinae, and it is known that the shape can be useful to identify the specimens of some groups in the subfamily to species. However, for *Oxythyrea*, no significant differences in the shape of the processus had been noted previously.

Let us now look at the case where the machine is sorting specimens to sex as well as species (Fig. 8). For the dorsal view, the heat maps are very similar to those for the first task, where the machine was separating the specimens only to species, but it seems that sometimes the region is a little bit wider and more complicated. Usually, the machine is focusing on several smaller parts in species that are more challenging to identify. Maybe to check the overall shape of the body and some specific structures?

If we check the results based on the ventral side, it is getting more interesting. We can see that the machine is merely focusing on the patterns of white marks, much more so than in the simple ‘species task’. We can see these results for the species with easily visible dimorphism, where it is checking exactly the median regions with the spots that are only present in males (e.g. *O. funesta*, *O. tripolitana*). For those species with less distinct dimorphism, it is anyway checking the median region of the abdomen, probably in an attempt to identify the presence or absence of the median groove mentioned earlier. Although the groove is often difficult to detect to the naked eye in these images, it is quite visible in some cases, for instance in the male specimen of *O. cinctella* (Fig. 8).
Awakening taxonomist’s third eye

Fig. 8. Class activation maps for the specimens from the species/sex task (only select species are depicted). [Colour figure can be viewed at wileyonlinelibrary.com].

Conclusions – AI and the future of insect taxonomy

Taxonomist: Are there no limits to what these AI systems can do? Will taxonomists be jobless in the future?

AI expert: AI will certainly replace or at least support humans in many tasks. At the same time, AI will create a plethora of new opportunities. There is a common belief in the AI community that today, or in the near future, AI can replace humans on tasks that take less than a second to observe, understand and execute. The same applies to all tasks that can be partitioned into several one-second tasks. This has been shown to be true already for many well-defined tasks of this type. An example would be the routine identification of common taxa, a task that taxonomists often find time consuming and that takes time from more challenging problems, such as circumscribing and describing new taxa, or inferring phylogeny and studying evolution.

Today’s AI does not think, it usually cannot explain its decisions and above all, it is not creative. These are qualities that every scientist should have and taxonomists are no exception. AI should be considered just a tool, which could help taxonomists to be faster and more efficient in gathering knowledge. It is up to scientists to postulate hypotheses, test them, draw conclusions and share what has been discovered with the rest of society.

As an illustration, consider heat maps that highlight regions of images that a model finds informative. Now, imagine a situation where the highlighted image region depicts specimen parts with a lot of hairs. What feature of the hairs is important for discrimination? Is it their number, distribution, thickness, length or something else? How about some underlying feature not even related to hairs such as the shape or size of the hairy body part? While off-the-shelf solutions can tell us whether a machine identification task is feasible, and heat maps give us a hint of where to look for interesting features, it is up to the expert to hypothesize what the feature of interest is.

Back to the original question, I think we should turn it around, and ask instead how taxonomists can benefit from AI. I hope the experiments that we described here show that modern AI tools are not all about developing automated identification systems. These tools can be quite helpful for insect taxonomists in many other ways. As we have seen, they can be used to help identify the features that best distinguish closely related species. They can also be used to find out whether it is possible to distinguish sibling species morphologically. Hopefully, our discussion will inspire many insect systematists to explore the full range of opportunities offered by modern AI tools.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Confusion matrices for the identification of sex from different views and imaging settings when a dedicated CNN is trained. We show results when high-resolution images of dorsal (top left) and ventral habitus (bottom left) are used alone; when dorsal and ventral habitus images are combined (top right); and when features are obtained from smartphone images of the dorsal habitus (bottom right). Rows show true categories and number of images per species (N) and columns show predicted categories. Colors and numbers in the matrix indicate the proportion of predictions normalized so that the sum in each row is equal to 1. (abi = O. abigail, alb = O. albopicta, cin = O. cinctella, dul = O. dulcis, fun = O. funesta, gro = O. gronbechi, noe = O. noemi, pan = O. pantherina, sub = O. subcalva, tri = O. tripolitana)

Figure S2. Confusion matrices for the identification of species and sex when a dedicated CNN is trained. Experiments are based on either dorsal (left) or ventral habitus (right) images alone.

Data S1. Supporting information

© 2021 The Authors. Systematic Entomology published by John Wiley & Sons Ltd on behalf of Royal Entomological Society. doi: 10.1111/syen.12492
Acknowledgements

This work was supported by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 642241 to Miroslav Valan, by the Ministry of Culture of the Czech Republic (D KRVO 2019-2023/5.1.b, National Museum, 00023272), Charles University in Prague (SVV 260 571/2020) and by the European Community Research Infrastructure Action under the Synthesys Project [FR-TAF-5869 (http://www.synthesys.info/)] to Dominik Vondráček. The authors are obliged to Dirk Ahrens (ZFMK) and Olivier Montreuil (MNHN) for their help to gather more specimens of species used in this study. The authors declare that they have no competing interests.

Author Contributions

Miroslav Valan, Dominik Vondráček and Fredrik Ronquist conceived the experiments, Miroslav Valan carried out the experiments and analysed the results, Dominik Vondráček identified the specimens. Dominik Vondráček photographed the specimens for Dataset A and Miroslav Valan for Dataset B. Miroslav Valan wrote the first draft. All authors reviewed the manuscript and contributed to the final version.

Data availability statement

The data that support the findings of this study are openly available on github at https://github.com/valanm/awakening-taxonomists-third-eye

References

Azadi, S., Feng, J., Jegelka, S., & Darrell, T. (2015). Auxiliary Image Regularization for Deep cnns With Noisy Labels. arXiv Preprint arXiv:1511.07069.
Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A. & Carlsson, S. (2016) Factors of transferability for a generic ConvNet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, 1790–1802.
Baraud, J. (1985) Coléoptères Scarabaeoidea: faune du Nord de l’Afrique, du Maroc au Saini. Encyclopédie Entomologique, p. 650. Éditions Lechevalier, Paris.
Baraud, J. (1992) Coléoptères Scarabaeoidea d’Europe. Fédération Française des Sociétés de Sciences Naturelles & Société Linnéenne de Lyon, Faune de France, 78, 1–856.
Bezděk, A. (2016) Subfamily Cetoninae. Catalogue of Palearctic Coleoptera. Vol. 3. Scarabaeoidea – Scirtoidea – Dascilioidea – Buprestoidea – Byrrhoidea. Revised and Updated Edition (ed. by I. Löbl and D. Löbl), pp. 367–412. E. J. Brill, Leiden.
Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. & Darrell, T. (2014) DeCAF: a deep convolutional activation feature for generic visual recognition. International Conference on Machine Learning, pp. 647–655. ICML, Beijing.
Lee, K.-H., He, X., Zhang, L., & Yang, L. (2018). Cleannet: Transfer learning for scalable image classifier training with label noise. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5447–5456.
Li, Y., Yang, J., Song, Y., Cao, L., Luo, J. & Li, L.-J. (2017) Learning from noisy labels with distillation. Proceedings of the IEEE International Conference on Computer Vision, pp. 1910–1918. IEEE, USA, Li, Z. & Hoiem, D. (2018) Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 2935–2947.
Marques, A.C.R., Raimundo, M.M., Cavalheiro, E.M.B., Salles, L.F.P., Lyra, C. & Von Zuben, F.J. (2018) Ant genera identification using an ensemble of convolutional neural networks. PLoS One, 13, e0192111.
Mikšík, R. (1982) Monographie der Cetoninae der Palaärtischen und Orientalischen region Coleoptera. Lamellicornia. Band 3. Systematischer Teil. Cetonini I Teil, p. 530. Forstinstitut in Sarajevo, Sarajevo. 14 pls.
Ng, A. (2016). Why AI Is the New Electricity. Nikkei Asian Review Online. Nikkei Inc., Tokyo.
Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1944–1952.
Rodner, E., Simon, M., Brehm, G., Pietsch, S., Wägele, J. W., & Denzler, J. (2015). Fine-Grained Recognition Datasets for Biodiversity Analysis. arXiv Preprint arXiv:1507.00913. Cornell University, New York.
Russakovsky, O., Deng, J., Su, H. et al. (2015) ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115, 211–252.
Sabatinelli, G. (1981) Le Oxythyrea Muls. del Mediterraneo: studio morfologico sistematico (Coleoptera, Scarabaeoidea). Fragmenta Entomologica, 16, 45–60.
Sabatinelli, G. (1984) Studi sul genere Oxythyrea Muls: note sulle specie del gruppo cinctella (Schaum) (Scarabaeidae Cetoniinae). Bollettino Della Societa Entomologica Italiana, 116, 102–104.
Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 806–813.
Tanaka, D., Ikami, D., Yamasaki, T., & Aizawa, K. (2018). Joint optimization framework for learning with noisy labels. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5552–5560.
Tegelberg, R., Haapala, J., Mononen, T., Pajari, M. & Saarenmaa, H. (2012) The development of a digitising service Centre for natural history collections. ZooKeys, 209, 75–86.
Valan, M., Makonyi, K., Maki, A., Vondráček, D. & Ronquist, F. (2019) Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. Systematic Biology, 68, 876–895.
Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., & Belongie, S. (2017). Learning from noisy large-scale datasets with minimal supervision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 839–847.
Wei, X.-S., Luo, J.-H., Wu, J. & Zhou, Z.-H. (2017) Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing, 26(6), 2688–26811. https://doi.org/10.1109/TIP.2017.2688133.
Xia, T., Xia, T., Yang, Y., Huang, C., & Wang, X. (2015). Learning from massive noisy labeled data for image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2691–2699.
Zheng, L., Zhao, Y., Wang, S., Wang, J., & Tian, Q. (2016). Good Practice in CNN Feature Transfer. arXiv. http://arxiv.org/abs/1604.00133
Zhuang, B., Liu, L., Li, Y., Shen, C., & Reid, I. (2017). Attend in groups: a weakly-supervised deep learning framework for learning from web data. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1878–1887.

Accepted 5 March 2021

© 2021 The Authors. Systematic Entomology published by John Wiley & Sons Ltd on behalf of Royal Entomological Society. doi: 10.1111/syen.12492