Machine vision automated species identification scaled towards production levels

Article *in* Systematic Entomology · January 2016

DOI: 10.1111/syen.12146

Colin Favret
Université de Montréal

Systematic Entomology

Systematic Entomology (2016), **41**, 133–143 DOI: 10.1111/syen.12146

Machine vision automated species identification scaled towards production levels

COLIN FAVRET[1] and JEFFREY M. SIERACKI[2]

[1]Department of Biological Sciences, University of Montreal, Montreal, Canada and [2]SR2 Group, LLC, Columbia, MD, U.S.A.

Abstract. Computer-automated identification of insect species has long been sought to support activities such as environmental monitoring, forensics, pest diagnostics, border security and vector epidemiology, to name just a few. In order to succeed, an automated identification programme capable of addressing the needs of the end user should be able to classify hundreds of taxa, if not thousands, and is expected to distinguish closely related and hence morphologically similar species. However, it remains unknown how automated identification methods might handle an increase in data quantity, be it in reference imagery or taxonomic diversity. We sought to test the scalability of an automated identification method in terms of the number of reference specimens used to train the classifier and the number of taxa into which the classifier should assign unknown specimens. Is there an optimal number of reference images, where the cost of acquiring more images becomes greater than the marginal increase in identification success? Does increasing taxonomic diversity affect identification success, whether negatively or positively? In order to test the scalability of the automated insect identification enterprise, we used a sparse processing technique and support vector machine to test the largest dataset to date: 72 species of fruit flies (Diptera: Tephritidae) and 76 species of mosquitoes (Diptera: Culicidae). We found that: (i) machine vision methods are capable of correctly classifying large numbers of closely related species; (ii) when the misclassification of a specimen occurs at the species level, it is often classified in the correct genus; (iii) classification success increases asymptotically as new training images are added to the dataset; (iv) broad taxon sampling outside a focal group can increase classification success within it.

Computer-assisted insect identification was suggested almost 50 years ago (Rohlf & Sokal, 1967). The present desire for such a system is attested to by the several attempts to identify arthropods in specific research contexts, such as spider ecology (Do *et al.*, 1999; Russell *et al.*, 2000), aquatic environmental monitoring (Larios *et al.*, 2007; Lytle *et al.*, 2010) and orchard pest monitoring (Wen *et al.*, 2009). However, the realization of an automated identification system available as an end-user application remains elusive (MacLeod, 2007; MacLeod *et al.*, 2010).

Although computing and imaging technology has advanced dramatically in recent years, we are not yet at the point of being able to render digitally and analyse comparatively thousands of biological objects in three dimensions. Thus, the images themselves have so far been two-dimensional. In the case of automated insect identification, this imaging constraint has led to a tendency to analyse images of wings (Daly *et al.*, 1982; Yu *et al.*, 1992; Vaňhara *et al.*, 2007; Bhanu *et al.*, 2008; Santana *et al.*, 2014; Li & Cao, 2015). Insect wings are relatively flat and easy to image in a standard orientation, especially in comparison with other anatomical features such as genitalia, often the structures of most interest to insect taxonomists for identifying species.

The most common approach to automated identification is to create a set of reference images representing, as best as possible, the breadth of morphological variability of each taxon. These training images are used to extract a set of machine-interpretable characters that are, in turn, used to evaluate an image of an unidentified specimen and assign it to a taxon (i.e. classify it). Most previous research efforts were proofs of the concept of automated classification and therefore employed research data of limited taxonomic scope. For example, Lytle *et al.* (2010),

Correspondence: Colin Favret, University of Montreal, Biodiversity Centre, 4101 rue Sherbrooke est, Montreal, Quebec, H1X 2B2 Canada. E-mail: ColinFavret@AphidNet.org

Wen *et al.* (2009) and Santana *et al.* (2014) used only nine stonefly, five moth, and five orchid bee taxa, respectively. To date, it has not been clear how automated identification technology might scale up when presented with much larger datasets of reference imagery, in terms of both the number of taxa and the number of specimens per taxon. Is there an optimal number of reference images, or is there a threshold of diminishing returns where the cost of acquiring more images is greater than the marginal increase in identification success? Likewise, does increasing the number of classes (e.g. species) affect identification success, whether positively or negatively, perhaps by increasing the measurable morphological overlap between taxa?

We explored machine vision automated identification methods on insect image datasets that were significantly larger than ever before tested and with sets of closely related species. We employed modern sparse signal analytics (Sieracki & Benedetto, 2005) and machine learning methods on two-dimensional images of membranous wings to automatically identify species of fruit flies (Diptera: Tephritidae) and mosquitoes (Diptera: Culicidae). Fruit flies are one of the most economically damaging insect pests of agriculture. For example, in the countries of the eastern Mediterranean, the Mediterranean fruit fly causes US$298 million in direct (e.g. yield loss) and indirect (e.g. environmental impact) damage annually (International Atomic Energy Agency, 2001); the potential establishment in the U.S. of this same species might cost as much as US$800 million annually in direct economic damage and increased management efforts (Miller *et al.*, 1992). Tephritidae have a diversity of pigment patterns on their wings, often used by taxonomists for identification purposes. Mosquitoes are the most important insect vectors of human disease, transmitting malaria, yellow and dengue fevers, and many other diseases. An estimated 7.5 million human deaths in the decade ending in 2012 have been attributed to malaria alone (World Health Organization, 2013). In contrast with the wings of fruit flies that exhibit taxon-specific variation in patterning, those of mosquitoes are covered with scales that often rub off, resulting in a great deal of pattern variation that is not at all taxon-specific.

The economic and medical importance of these two groups of Diptera means that identification services are in high demand. It also means that the museum specimens of these insects are relatively well identified in comparison to other less well-studied insect groups, an important consideration for building sets of training images.

Materials and methods

Image acquisition

Images of insect wings were acquired from identified museum specimens during the first half of 2010. Identifications were recorded from specimen or unit tray labels in collections curated by world experts. Budget constraints prevented the re-identification and re-curation of the multiple thousands of specimens. Wings were not removed; they were imaged in silhouette with a white background. We selected specimens that had their wings spread out enough to enable a relatively unobstructed view when rotated under the stereoscope lens (Fig. 1). Because the wings were imaged in silhouette, we did not prioritize dorsal or ventral views, nor did we preferentially select right or left wings. Images of pinned specimens were taken with a Leica M205C stereoscope with a 0.63× plan-apochromat objective, and a Leica DFC295 3 megapixel digital camera tied to Leica's Firecam software. In general, we kept the stereoscope at a set zoom level for each species, but aimed to fill the camera's field of view with different species. Therefore, image analyses were not able to consider overall wing size.

Fig. 1. A specimen of *Ceratitis cosyra* rotated under a stereoscope for a mostly clear view of the ventral aspect of the right wing.

Fruit flies were imaged in the Entomology Department of the U.S. National Museum of Natural History, Washington, DC, U.S.A. We selected species for which at least 25 wings could be readily imaged from across the taxonomic diversity of Tephritidae, including three of the six recognized subfamilies, 11 of the 27 tribes, 24 of the 481 genera, and 72 of the 4352 species (Norrbom, 2010; Table 1). The current nomenclature was researched in Systema Dipterorum (Pape & Evenhuis, 2013) with the exception of *Ceratitis querita*, which was found in De Meyer & Friedberg (2006). Rarely, fruit fly species are known to be sexually dimorphic (Sivinski & Dodson, 1992; Sivinski & Pereira, 2005; Dujardin & Kitthawee, 2013), even in wing shape and venation (Aluja & Norrbom, 1999). We did not sex the flies, and therefore our image capture might have inadvertently favoured one sex over the other if one was better represented in the collection. Figure 2 presents a sampling of fruit fly training images from multiple genera.

Mosquitoes were from the U.S. National Museum of Natural History's collection located at the Walter Reed Biosystematics Unit at the Museum Support Center, Suitland, MD, U.S.A. Again, we selected species across the taxonomic diversity of Culicidae, including both subfamilies, eight of 11 tribes, 16 of 42 genera, and 79 of 3492 species (Rueda, 2008; WRBU, 2014; Table 2). The current nomenclature was researched in the Walter Reed Biosystematics Unit Systematic Catalog of Culicidae (WRBU, 2014), but we retained the classical generic combinations for *Aedes* species, as the newer ones are not well

Table 1. Classification success rates of 1800 fruit flies into 24 genera and 72 species.

Genus	Species	Classification rate (%)
Acanthiophilus[a]	*helianthi* (Rossi 1794)	92
Aciurina[b]	*bigeloviae* (Cockerell 1890)	96
Anastrepha[c]		95
	anduzei Stone 1942	84
	canalis Stone 1942	68
	coronilli Carrejo & Gonzalez 1993	60
	crebra Stone 1942	84
	debilis Stone 1942	80
	distincta Greene 1934	52
	fraterculus (Wiedemann 1830)	88
	ludens (Loew 1873)	84
	minuta Stone 1942	84
	nigrifascia Stone 1942	92
	obliqua (Macquart 1835)	48
	panamensis Greene 1934	84
	pickeli Lima 1934	64
	robusta Greene 1934	88
	serpentina (Wiedemann 1830)	88
	spatulata Stone 1942	64
	striata Schiner 1868	76
	superflua Stone 1942	96
	suspensa (Loew 1862)	88
	turpiniae Stoen 1942	72
	zeteki Greene 1934	60
	zuelaniae Stone 1942	36
Bactrocera[d]		93
	cucurbitae (Coquillett 1899)	88
	frauenfeldi (Schiner 1868)	96
	umbrosa (F. 1805)	92
Ceratitis[e]		97
	anonae Graham 1908	76
	capitata (Wiedemann 1824)	88
	colae Silvestri 1913	76
	cosyra (Walker 1849)	72
	ditissima (Munro 1938)	88
	fasciventris (Bezzi 1920)	84
	flexuosa (Walker 1853)	80
	hamata Meyer 1996	72
	marriotti Munro 1933	84
	podocarpi (Bezzi 1924)	88
	querita (Munro 1937)	100
	rosa Karsch 1887	76
	rubivora Coquillett 1901	84
	simi Munro 1933	96
Dacus[d]		96
	bivittatus (Bigot 1858)	96
	ciliatus (Loew 1862)	92
Dioxyna[a]	*picciola* (Bigot 1857)	92
Euaresta[a]		98
	aequalis (Loew 1862)	96
	bella (Loew 1862)	100
Euarestoides[a]	*acutangulus* (Thomason 1869)	88
Eurosta[b]	*floridensis* Footte 1977	100
Eutreta[f]	*diana* (Osten Sacken 1877)	92
Neaspilota[g]		96
	achilleae Johnson 1900	100
	albidipennis (Loew 1861)	84
Paracantha[f]	*gentilis* Herin 1940	92
Rhagoletis[h]		97
	cingulata Wilson & Lovett 1913	80
	indifferens Curran 1932	96
	pomonella (Walsh 1867)	100
Strauzia[i]	*longipennis* (Wiedemann 1830)	92
Tephritis[a]		96
	araneosa (Coquillett 1894)	92
	signatipennis Foote 1960	100
	stigmatica (Coquillett 1899)	96
Terellia[g]	*occidentalis* (Snow 1894)	92
Tomoplagia[j]	*quinquefasciata* (Macquart 1835)	80
Toxotrypana[c]	*curvicauda* Gerstaecker 1860	92
Trupanea[a]		99
	actinobola (Loew 1873)	88
	jonesi (Curran 1932)	92
	nigricornis (Coquillett 1899)	92
	wheeleri (Curran 1832)	88
Trypanaresta[a]	*delicatella* (Blanchard 1852)	88
Urophora[k]		69
	pauperata (Zaitzev 1945)	60
	sirunaseva (Hering 1938)	72
	solstitialis (L. 1758)	64
Xanthaciura[a]	*insecta* (Loew 1862)	100
Zonosemata[h]	*electa* (Say 1830)	88

[a]Tephritinae–Tephritini; [b]Tephritinae–Dithrycini; [c]Trypetinae–Toxotrypanini; [d]Dacinae–Dacini; [e]Dacinae–Ceratidini; [f]Tephritinae–Eutretini; [g]Tephritinae–Terelliini; [h]Trypetinae–Carpomyini; [i]Trypetinae–Trypetini; [j]Tephritinae–Acrotaeniini; [k]Tephritinae–Myopitini.

defined and have not been universally adopted (Rueda, 2008). We photographed 100 wings per species of *Anopheles* and 25 for all others. Only female mosquitoes were photographed. Mosquito wings are covered in scales that are often lost or rubbed off during the insect's life or during museum preparation. We prioritized imaging wings that were in relatively good condition, although it was extremely rare to find any mosquito with fully intact wing scales. Figure 3 shows training images of wings of three important disease vectors.

The image capture rate for fruit flies averaged 29.5 h^{-1} and that of mosquitoes was 22.9 h^{-1}. The fruit flies were easier to image because the specimens themselves were usually larger than mosquitoes and their wings were more likely to be spread to the side and thus more readily viewed: less time was spent rotating the specimen so as to get an optimal view of the wing. The fruit fly collection was physically located immediately adjacent to the imaging station, whereas the mosquitoes were some distance away, also contributing to the difference in image capture rates.

Image analysis

Wing images were rotated and flipped as necessary so that the base of the wing was near the right margin of the image and

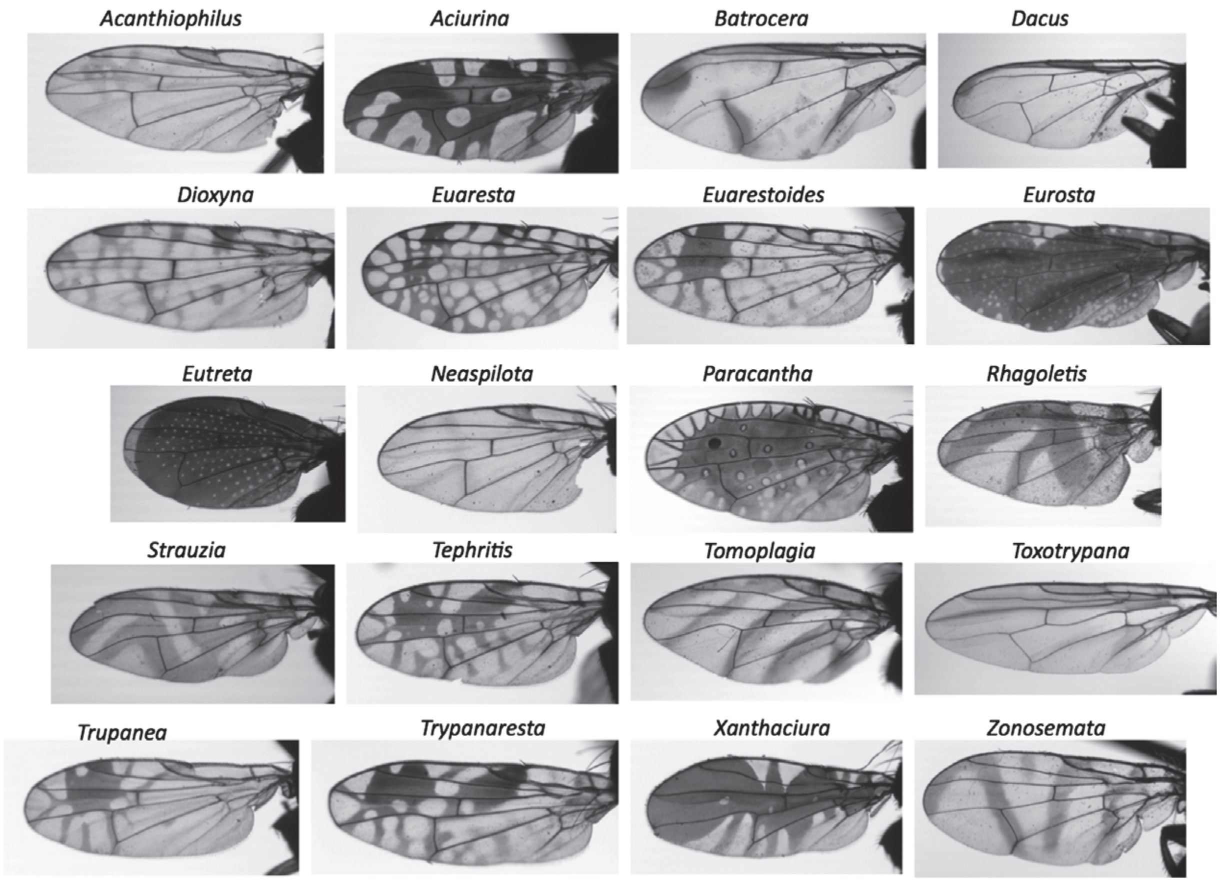

Fig. 2. Twenty photographs of a variety of fruit fly wings showing the diversity in shape and patterning at the generic level.

the posterior margin of the wing was near the bottom margin of the image. As such, each image appeared to be that of the dorsal aspect of the left wing with the specimen facing forward, even if the actual image was of the right wing or the ventral side. In order to achieve reasonable consistency among such a large set of images, we developed a preprocessing script in MATLAB (MathWorks Inc., Natick, MA, USA) that employed standard techniques to find, align and crop the wing region of each photograph. Within the MATLAB image processing toolbox, we applied thresholding and filling to produce a bitmap mask of the wing, the leading edge was identified using a Hough transform, and the remaining boundaries were determined by comparison with an expected shape template so that portions of the insect body or stray limbs that intersected the wing did not bias the cropping window. The image was then aligned so that the leading edge was approximately level and cropped to the boundary of the mask. Output images were reviewed by eye to catch and correct rare errors before further processing. Within each working group, these cropped wing images were then further processed to resize them to a common resolution and greyscale range so that they could be compared against their peers without regard to photographic variations in lighting or magnification. Lighting correction was limited to a simple automatic contrast adjustment in Adobe Photoshop, followed by re-ranging the greyscale levels to span the available bit depth. Cropping and resizing the images to a common resolution meant that the absolute size of wings was not considered in this analysis, leaving only detail patterns within the wings and, to a lesser extent, their shape as possible distinguishing features.

With the images coarsely co-aligned, we used a sparse processing technique called greedy adaptive discrimination (GAD) (Sieracki & Benedetto, 2005; Sieracki *et al.*, 2008) to find and extract commonly occurring signature characteristics within the image groups. To summarize, GAD works by simultaneously considering an ensemble of data and seeking a common, joint representation by which to compactly describe the members of the ensemble. The representation is selected according to a mathematical cost function and can be thought of as a form of nonlinear minimization problem. It is thus related to other sparse analytics approaches such as the compressive sensing methods introduced by Candès *et al.* (2006), Donoho (2006), and others. All of these methods focus on the recovery of information with a relatively small number (i.e. a sparse set) of coefficients and features. GAD uses joint information from multiple samples to recover signals significantly below the noise floor that would otherwise limit recovery from any one sample alone; moreover, GAD is largely unaffected by positional jitter between these multiple samples. In the context of the present study, this results

Table 2. Classification success rates of 1975 mosquitoes into 16 genera and 79 species.

Genus	Species	Classification rate (%)
Aedes[a]		90
	aegypti (L. 1762)	68
	albolineatus (Theobald 1904)	92
	albopictus (Skuse 1894)	72
	angustivitttatus Dyar & Knab 1907	68
	canadensis (Theobald 1901)	68
	cataphylla Dyar 1916	80
	communis (De Geer 1776)	84
	dorsalis (Meigen 1830)	84
	excrucians (Walker 1856)	76
	fitchii (Felt & Young 1904)	72
	intrudens Dyar 1919	88
	pullatus (Coquillett 1904)	92
	punctor (Kirby 1837)	48
	scapularis (Rondani 1848)	80
	serratus (Theobald 1801)	84
	sollicitans (Walker 1856)	60
	sticticus (Meigen 1838)	80
	taeniorhynchus (Wiedeman 1821)	80
	togoi (Theobald 1907)	52
	triseriatus (Say 1823)	92
	vexans (Meigen 1830)	84
Anopheles[b]		96
	aconitus Doenitz 1902	92
	albimanus Wiedemann 1820	92
	dirus Peyton & Harrison 1979	88
	marajoara Galvao & Damasceno 1942	84
	minimus Theobald 1901	92
	oswaldoi (Peryassu 1922)	96
	pseudopunctipennis Theobald 1901	76
	punctulatus Donitz 1901	80
	quadrimaculatus Say 1824	96
	triannulatus (Neiva & Pinto 1922)	92
Coquillettidia[c]		88
	fasciolata (Lynch Arribalzaga 1891)	68
	nigricans (Coquillett 1904)	80
	perturbans (Walker 1856)	88
Culex[d]		91
	annulirostris Skuse 1889	64
	bitaeniorhynchus Giles 1901	80
	coronator Dyar & Knab 1906	76
	erraticus (Dyar & Knab 1906)	92
	fuscocephala Theobald 1907	84
	mollis Dyar & Knab 1906	76
	nigripalpus Theobald 1901	76
	pipiens L. 1758	92
	quinquefasciatus Say 1823	80
	restuans Theobald 1901	80
	salinarius Coquillett 1904	56
	sitiens Wiedemann 1828	84
	tarsalis Coquillett 1896	68
	tritaeniorhynchus Giles 1901	88
	vishnui Theobald 1901	84
Culiseta[e]		86
	incidens (Thomson 1869)	76
	inornata (Williston 1893)	88
Deinocerites[d]		87
	cancer Theobald 1901	60
	magnus (Theobald 1901)	92
	pseudes Dyar & Knab 1909	72
Haemagogus[a]	*argyromeris* Dyar & Ludlow 1921	88
Limatus[f]		78
	asulleptus (Theobald 1903)	72
	durhamii Theobald 1901	84
Lutzia[d]		88
	fuscana (Wiedemann 1820)	88
	halifaxii (Theobald 1903)	84
Mansonia[c]		84
	titillans (Walker 1848)	80
	uniformis (Theobald 1901)	76
Orthopodomyia[g]	*signifera* (Coquillett 1896)	92
Psorophora[a]		81
	albipes (Theobald 1907)	68
	ciliata (F. 1794)	92
	cingulata (F. 1805)	72
	confinnis (Lynch Arribalzaga 1891)	64
	ferox (von Humboldt 1819)	84
	pygmaea (Theobald 1903)	88
Toxorhynchites[h]		93
	moctezuma (Dyar & Knab 1906)	60
	septentrionalis (Dyar & Knab 1906)	100
	theobaldi (Dyar & Knab 1906)	60
Tripteroides[f]	*aranoides* (Theobald 1901)	96
Uranotaenia[i]		96
	anhydor Dyar 1907	100
	bicolor Leicester 1908	84
	geometrica Theobald 1901	88
	lowii Theobald 1901	88
	lutescens Leicester 1908	88
	obscura Edwards 1915	56
Wyeomyia[f]	*felicia* (Dyar & Nunez-Tovar 1927)	88

[a]Aedini; [b]Anophelinae; [c]Mansoniini; [d]Culicini; [e]Culisetini; [f]Sabethini; [g]Orthopodomyiini; [h]Toxorhynchitini; [i]Uranotaeniini.

in robustness to image noise and variations in alignment. This is born out in the results shown later in the paper, which are achieved with only course-grained wing position registration between the hand-acquired photographs, with no steps needed to suppress noise, speckle, shadows, bright spots or other photographic imperfections.

The GAD feature vectors were used as input for support vector machine (SVM) learning and classification (cf. Burges, 1998). SVM is a machine learning tool that is largely agnostic to the statistical structure of data other than at the boundary between classes (Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2002). SVM attempts to separate classes of data by plotting feature vector points in an N-dimensional space and drawing a boundary between the classes. The results in our case are a set of emergent feature characteristics that can thereafter be treated analogously to principal components results. These characteristics are then exploited to distinguish

Fig. 3. Photographs of a wing of each of the mosquitoes *Culex pipiens*, *Anopheles quadrimaculatus* and *Aedes albopictus*, representing the three major genera of disease vector species.

each target group of insects from respective confounding or challenge groups. The degree to which each set of signature characteristics occurs in any particular image generates a feature vector. While our methods were implemented using a library of in-house software, SVM tools with similar function are widely available (e.g. Pelckmans *et al.*, 2002, MATLAB SVM toolbox). It should be noted that the SVM classifiers were adjusted to minimize the overall interspecific error rates, without regard to whether the errors were false positives or false negatives and without preference to any classification category. While it is possible to improve success in some categories at the expense of others, this trade-off was not explored in the present study.

The signature features employed by the system were adaptively discovered from each dataset and, as such, do not generally correspond to any morphological or morphometric character ordinarily employed by taxonomists. We have made a tentative investigation of the nature of the discovered characteristics, confirming, for example, that spatial patterns in certain areas of the wings are exploited among the discriminatory features between some species; however, we report here only on our success in discrimination and a detailed analysis of those emergent features' characteristics remains for future work.

Classification 'success', defined as a machine-delivered classification of a test that conforms to the taxonomist-delivered identification, was first quantified using a leave-one-out method. The specific analyses we ran with the leave-one-out method included: (i) 25 images of each of 72 species of Tephritidae; (ii) 25 images each of 79 species of Culicidae; (iii) and 100 images each of ten species of *Anopheles*. Within each working group of insects, the entire dataset of identified images was used to train the program in pattern recognition, with the exception of a single individual, which was then classified based on the prior training. The procedure was repeated as many times as there were images, each repetition leaving one individual out of the training set to be subsequently classified. Results are summarized in confusion matrices; these indicate the frequency with which identified individuals of each species were assigned to each of the possible species within the classification set. The sum of the counts on the diagonal divided by the total number of specimens in the experiment provides a measure of overall classification success.

The ability to classify using sets of training data of different sizes was further quantified using the method of k-fold cross-validation (Kohavi, 1995) on the 1000 images of ten species of *Anopheles*. In this approach, $1/k\%$ of the data is used for testing and the remaining (100 to $1/k\%$) for training. At each repetition, different training and testing subset permutations are selected at random from the sample space of available images. The proportion of data used for training is varied in size (k), with training and testing repeated multiple times (> 30) at each size, each time with different data permutations, to generate parametric performance statistics. In this instance we varied the training set size from a single individual up to 90% of the available data, repeating each one 40 times. Statistics produced by this process give an indication of how well our classifier results will generalize to new, independent sets of data.

Results

Fruit flies (Diptera: Tephritidae)

The full 72-species confusion matrix of fruit flies, with 25 images per species, yielded an overall classification success rate of 86.2% to the species level (Fig. 4a) and 94.4% to genus (Fig. 4b). Classification success ranged from 36 to 100% for species, and from 69 to 100% for genera (Table 1). The most frequent misclassifications were for species within a genus. For example, if a specimen of an *Anastrepha* species was misclassified, it was most likely to be classified as another species of *Anastrapha* rather than as a species of another genus (Fig. 4a). This phenomenon was evident for both of the genera with the most sampled species, *Anastrepha* and *Ceratitis*, with 22 and 14 species, respectively. The misclassification rate for species of *Anastrepha* averaged 25.5%, but the likelihood of being correctly identified to genus was 92% (22 species possible out of 72 in all). Likewise, for *Ceratitis*, the rate of misclassification to the species level was 17.2% (Table 1), but the likelihood of those being classed correctly within the genus was 97% (13 species out of 72). The lowest classification success was seen in individual species of *Anastrepha* (*A. zuelaniae* and *A. obliqua*) and at both within-genus species and genus level in *Urophora*. We achieved 100% classification success in distinguishing the species *Ceratitis querita*, *Euaresta bella*, *Eurosta floridensis*, *Neaspilota achilleae*, *Rhagoletis pomonella*, *Tephritis signatipennis* and *Xanthaciura insecta*, and the genera *Eurosta* and *Xanthaciura* (Table 1).

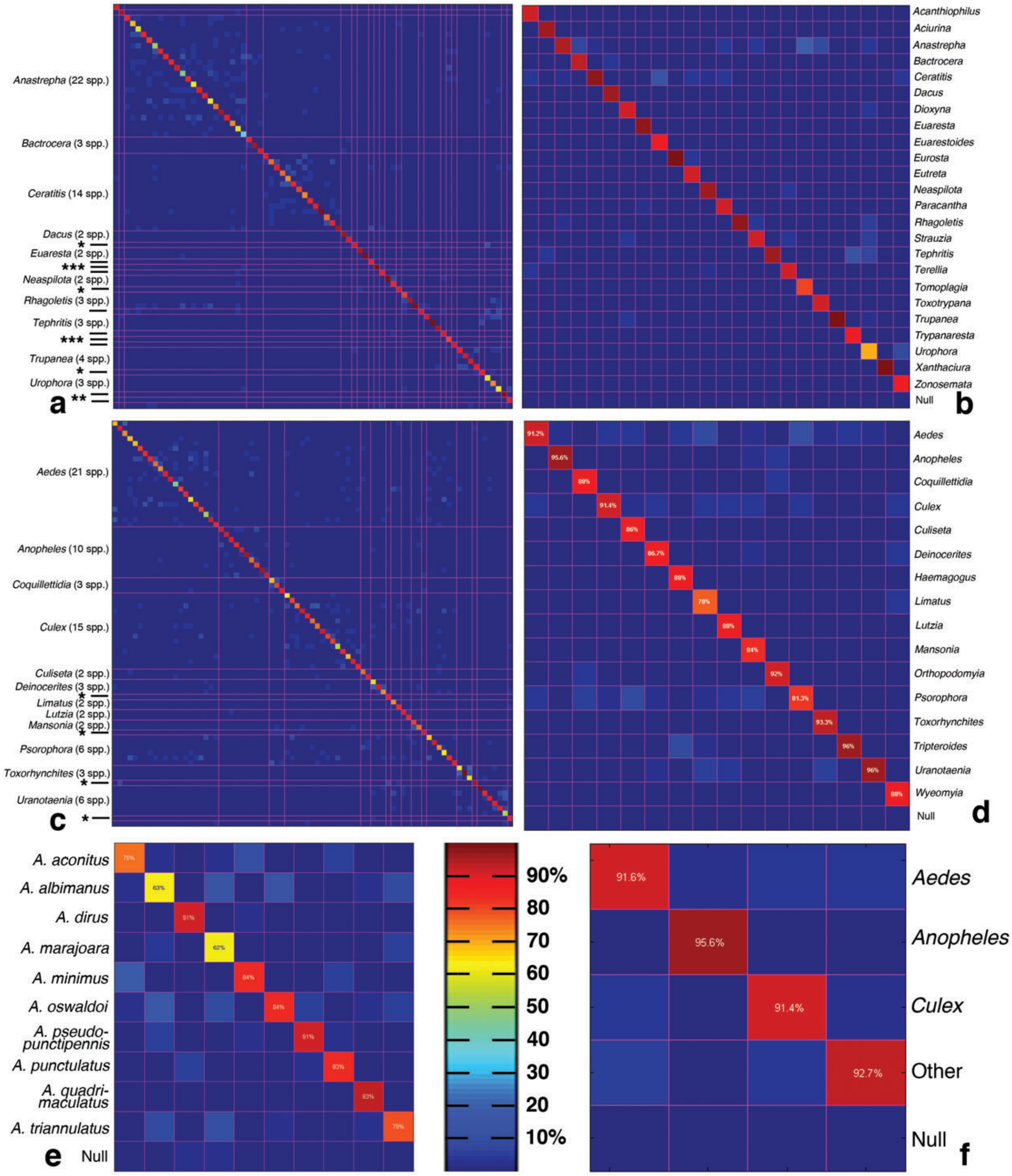

Fig. 4. (a) Confusion matrix for the automated classification of 1800 fruit flies into 72 species, shown grouped in blocks by genus. The values on the diagonal indicate the frequency of successful classification of individuals within each species and values off the diagonal indicate errors. The last row, labelled 'null', contains counts of instances in which a specimen could not be classified. Asterisks denote genera for which a single species was included [refer to (b) and Table 1 for genus and species names]; (b) confusion matrix for automated classification of 1800 fruit flies into 24 genera; (c) confusion matrix for automated classification of 1975 mosquitoes into 79 species. Asterisks denote genera for which a single species was included [refer to (d) and Table 2 for genus and species names]; (d) confusion matrix for automated classification of 1975 mosquitoes into 17 genera; (e) confusion matrix for automated classification of 1000 *Anopheles* specimens into ten species; (f) confusion matrix for automated classification of 1975 mosquitoes into four classes, the three disease-vector genera and a fourth class including all others.

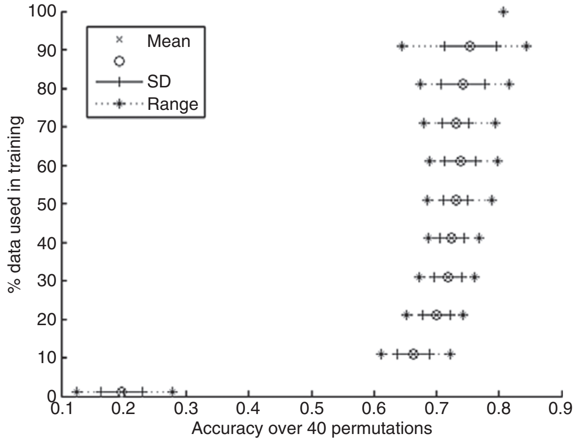

Fig. 5. *K*-fold testing of the classification of 1000 specimens into ten species of *Anopheles*. The first datum (1 on the *y*-axis) represents a test involving a single training image and an average of 20% classification success. The last datum (99 on the *y*-axis) represents a test involving 99 training images (i.e. a leave-one-out analysis) and a classification success rate of 80.6%.

Mosquitoes (Diptera: Culicidae)

Overall classification success for 79 mosquito species was 80.3% (Fig. 4c). Classification success ranged from 48 to 100% for species and from 78 to 96% for genera (Fig. 4d; Table 2). Of the three large genera of important disease vectors, *Aedes*, *Culex* and *Anopheles* had 90.4, 91.2 and 95.6% genus classification success, respectively (Table 2). *Aedes* and *Culex* had lower species classification success (80.3 and 78.7%, respectively), whereas *Anopheles* species had an overall classification rate of 88.8%. However, when we analysed the ten species of *Anopheles* alone, that is without species of other genera, and with the maximum available number of training images (100 per species), we saw the classification success rate drop to 80.6% (Fig. 4e). Using more training images and restricting the analysis to ten species of a single genus did not in itself yield better performance. Most notably, *Anopheles aconitus*, *Anopheles albimanus* and *Anopheles marajoara* were less frequently correctly classified in the *Anopheles*-only analysis with 100 images per species (Fig. 4e) than in the combined analysis with 25 images per species (Fig. 4c; Table 2).

Increasing the size of the *Anopheles*-only training set yields better results, following a roughly asymptotic curve in our *k*-fold validation test (Fig. 5). For each proportionate quantity of training data, the training testing cycle was repeated 40 times using different random permutations of the training and testing groups. Using one wing image as a training example, the classification success rate is approximately twice that of chance alone, approximately 20%. Using 10 of the 100 wing images as training, classification success increases to 67%. We see asymptotic behaviour and diminishing returns in adding more training data. The last point in Fig. 4f reflects the leave-one-out testing overall success rate, that is, a single test image classified against 99 training images.

Finally, we considered the three disease vector genera and a fourth category that included all 14 others. When we constrained the system to distinguish only these four classes within the 79 species × 25 image training set, we saw overall genus-level classification success rates of 92.9%: 91.4% for *Culex*, 92.0% for *Aedes*, 92.7% for the 'others' category, and 95.6% for *Anopheles* (Fig. 4f).

Discussion

Pattern recognition algorithms often employ a training dataset meant to represent the variability inherent in the pattern: the larger the training set of images, the greater the range of variability. This range increases as training images are added to individual taxa, as well as when the number of taxa is increased. Intuitively, one might imagine that larger training datasets within classes (taxa) would increase classification success, as the size of the virtual classification space increases. Conversely, one might expect that the addition of classes would decrease the rate of successful classification as the number of choices increases.

Generally, the performance of our analyses with significantly larger datasets was consistent with other automated arthropod identification studies with fewer taxa, which had classification success rates ranging from 81 to 96% (Table 3). Each of those studies used different discriminant methods and taxa, and each had their own particularities with regard to specimen preparation and condition, so one-to-one comparisons are not justified. However, it is important to note that previous studies did not explicitly target closely related species that are generally harder to identify, whether by person or by machine.

Examining the pattern of classification, when a specimen is misclassified, it is more likely to be misclassified as a different species of the same genus than that of a different genus. This phenomenon, also seen by Do *et al.* (1999), is graphically demonstrated in *Anastrepha* and *Ceratitis* (Fig. 4a) and *Aedes* and *Culex* (Fig. 4c). Because insect classification other than at the species level is not part of the character extraction and training, the machine is independently recognizing patterns correlated with taxonomists' classifications. However, taxonomist misidentifications are more likely in groups of closely related species, so specimens misidentified to species but attributed correctly to genus in our training data could bias our results (see later).

Anastrepha specimens were likely to be classified as another species in the same genus, but they were also the most likely of tephritid species to be misclassified: *A. distincta*, *A. obliqua* and *A. zuelaniae* only had classification rates of 52, 48 and 36%, respectively (Fig. 4a; Table 1). The difficulties with *Anastrepha* may be related to the taxonomic complexity of the genus: it has over 184 species, has not been thoroughly revised in its entirety for a long time (Aluja, 1994) and is replete with species complexes (Norrbom, 1988, 1998, 2002, 2009). *Aedes punctor* and *Aedes togoi* had the lowest classification rate, by a margin, among the mosquitoes (48 and 52%; Table 2). *Aedes punctor* was most frequently mistaken for *Aedes cataphylla* (Fig. 4c); these two species can be confused even using DNA barcodes

Table 3. Summary of select automated arthropod identification studies employing two-dimensional imagery.

Study	Subject	No. of species	Average no. of images per species	Classification success (%)
Weeks *et al.* (1997)	Ichneumonid wings	5	35	94
Do *et al.* (1999)	Spider genitalia	6	9	81
Watson *et al.* (2003)	Macrolepidoptera	35	20	83
Wen *et al.* (2009)	Orchard moths	5	93	88
Lytle *et al.* (2010)	Stonefly naiads	4	310	82
Kang *et al.* (2012)	Butterflies	7	38	86
Joutsijoki *et al.* (2014)	Benthic macroinvertebrates	8	169	96
Santana *et al.* (2014)	Orchid bees	5	28	88
This study	Fruit fly wings	72	25	86
This study	Mosquito wings	79	25	80

(Zhang *et al.*, 2012). Reinert *et al.* (2004) split the *Aedes* species into multiple other genera and subgenera; although they had placed *A. togoi* in a separate genus (*Tanakaius*), most misclassifications of that species were for other species of *Aedes* (Fig. 4c).

Unsurprisingly, training datasets of increasing size increase the rate of classification success. A single training specimen yields predictably low classification success (20%), but the success rate increases rapidly with the first added specimens (67% with ten training images). This rate then extends gradually as the training set size increases (Fig. 5). Each set of ten additional training images between 20 and 90 adds an average of 1.0% to the classification success rate; this gain is asymptotic and cannot continue indefinitely. Watson *et al.* (2003) analysed only 20 training images of each of their Lepidoptera species, but they extrapolated their classification success curves out to 50 images and found similar results to our own. Towards the development of an end-user-ready system, a calculation of the ideal number of training images to acquire could thus be made for any number of images beyond a minimum of 30 or so. Assuming the specimens are available, this cost–benefit analysis would take into account the number of taxa to be included, the financial cost of imaging each additional specimen, and the economic value of the actual classifications to be made.

Perhaps the most surprising result of our analyses was the higher classification success of *Anopheles* in the context of the 79-species analysis compared with the analyses of *Anopheles* species alone. Although the classification success of *Anopheles* species alone topped out at 80.6% when training with 100 images per species (Fig. 4e), when only 25 training images were used along with the other 69 mosquito species, specimens of *Anopheles* were correctly classified to species 88.8% of the time (Fig. 4c). It seems that the accurate identification of species may benefit more from having a larger variety of comparison points outside the genus than from increasing the training set within the genus. This benefit occurred even though the potential number of competing categories into which an individual *Anopheles* specimen could be misclassified increased from nine to 78, and thus the chance rate of correct classification decreased from 1 in 10 to 1 in 79. This phenomenon may be due to improved learning by the machine classifier: a larger number of independent comparison points provides an increased opportunity to distinguish differences in noisy data. The system may likewise benefit from improved feature discovery in a larger dataset. Interestingly, a similar phenomenon has been documented in phylogenetic studies where an increase in taxon sampling often leads to better phylogenetic resolution (Agnarsson & May-Collado, 2008; Heath *et al.*, 2008; Nabhan & Sarkar, 2012).

An important but unquantifiable variable in any test of insect machine vision classification is the accuracy of the training data; the starting assumption is that the initial taxonomist-rendered specimen identifications are correct. Unless the identifications of the source specimens are independently ground-truthed, it is impossible to know what the actual taxonomist-rendered identification success rate is, which will then directly affect the measured machine vision classification success. In fact, a misidentified training specimen would corrupt each step of the analysis: character extraction, training and classification. These kinds of confounding input data are much more likely with taxa that are hard to identify, the very taxa for which fully developed automated identification methods would be most valuable. Misidentifications in training data have a confounding effect on automated identification methods: first by artificially increasing the variability of a species' training set, and secondly by attributing variability to one species that rightly belongs associated with another. Ensuring the accuracy of the basic training datasets is paramount to a successful machine vision automated identification system.

Experts do, of course, make mistakes, but it is hard to know how often they do so because a large number of variables are at play: these include the complexity of the taxon in question and the taxonomist's individual expertise, workload, and even mood or time of day. In two studies, the top experts correctly identified dinoflagellate specimens 84–95% of the time (Culverhouse *et al.*, 2003); Epler (2001) documents a range of misidentification of larval Chironomidae from 6 to 60%, with fully 25% of 713 specimens misidentified among ten taxonomists. The problem of misidentification is exacerbated when nonexperts are the identifiers (Krell, 2004), prevalent in practical day-to-day field identification situations. Quantification of the accuracy of taxonomists and parataxonomists remains an important and understudied problem.

Given these aforementioned inherent challenges with data quality, our results are heartening. The machine's ability to correctly classify four out of five individuals to one of over

70 species, many of which are closely related, is on par with and may even surpass that of some professionals. Additionally, once trained, per-specimen machine identification costs are orders of magnitude less than those of salaried personnel. A rapid screening system for vector mosquito genera, for example (Fig. 4f), might greatly aid health workers in the field by bringing new response capability to large numbers of nonexpert technicians, increasing the number of insects that can be examined while at the same time freeing the expert professionals to focus more closely on high threat risk specimens. One can envision a scenario where routine identifications are automated, experts being called upon to intervene in day-to-day field identifications only in the cases of greatest importance, be they legal, security or economic.

It is important to underscore that expert taxonomists are indispensable. Indeed, in order to correctly train the classifier and keep it up to date as taxonomic science necessarily evolves, taxonomic research and expertise will be increasingly valuable.

Acknowledgements

We thank Allen Norrbom (Systematic Entomology Laboratory, US Department of Agriculture), Richard Wilkerson, Leopoldo M. Rueda, Desmond Foley, and James Pecor (Walter Reed Biosystematics Unit, Walter Reed Army Institute of Research), and David Furth and F. Christian Thompson (Department of Entomology, US National Museum of Natural History) for help in accessing specimens and equipment for imaging. Leopoldo M. Rueda also provided helpful comments on an earlier version of the manuscript. The authors each run their own Maryland-based limited liability companies, SR2 Group, LLC (Sieracki) and AphidNet, LLC (Favret); they declare no conflicts of interest.

References

Agnarsson, I. & May-Collado, L.J. (2008) The phylogeny of Cetartiodactyla: the importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. *Molecular Phylogenetics and Evolution*, **48**, 964–985.

Aluja, M. (1994) Bionomics and management of *Anastrepha*. *Annual Review of Entomology*, **39**, 155–178.

Aluja, M. & Norrbom, A.L. (1999) *Fruit Flies (Tephritidae) Phylogeny and Evolution of Behavior*. CRC Press, Boca Raton, Florida.

Bhanu, B., Li, R., Heraty, J. & Murray, E. (2008) Automated classification of skippers based on parts representation. *American Entomologist*, **54**, 228–231.

Burges, C.J. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Candès, E.J., Romberg, J.K. & Tao, T. (2006) Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, **59**, 1207–1223.

Cristianini, N. & Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, U.K.

Culverhouse, P.F., Williams, R., Reguera, B., Herry, V. & González-Gil, S. (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, **247**, 5.

Daly, H.V., Hoelmer, K., Norman, P. & Allen, T. (1982) Computer-assisted measurement and identification of honey bees (Hymenoptera: Apidae). *Annals of the Entomological Society of America*, **75**, 591–594.

De Meyer, M. & Friedberg, A. (2006) Revision of the subgenus *Ceratitis* (*Pterandrus*) Bezzi (Diptera: Tephritidae). *Israel Journal of Entomology*, **35-36**, 197–315.

Do, M.T., Harp, J.M. & Norris, K.C. (1999) A test of a pattern recognition system for identification of spiders. *Bulletin of Entomological Research*, **89**, 217–224.

Donoho, D.L. (2006) Compressed sensing. *IEEE Transactions on Information Theory*, **52**, 1289–1306.

Dujardin, J.-P. & Kitthawee, S. (2013) Phenetic structure of two *Bactrocera tau* cryptic species (Diptera: Tephritidae) infesting *Momordica cochinchinensis* (Cucurbitaceae) in Thailand and Laos. *Zoology*, **116**, 129–138.

Epler, J.H. (2001) *Identification Manual for the Larval Chironomidae (Diptera) of North and South Carolina. A Guide to the Taxonomy of the Midges of the Southeastern United States, including Florida*. North Carolina Department of Environment and Natural Resources, Raleigh, North Carolina, and St. Johns River Water Management District, Palatka, Florida.

Heath, T.A., Hedtke, S.M. & Hillis, D.M. (2008) Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution*, **46**, 239–257.

International Atomic Energy Agency (2001) *Economic Evaluation of Three Alternative Methods for Control of the Mediterranean Fruit Fly (Diptera: Tephritidae) in Israel, Jordan, Lebanon, Syrian Arab Republic and Territories under the Jurisdiction of the Palestinian Authority*. IAEA-TECDOC 1265. International Atomic Energy Agency, Vienna, Austria.

Joutsijoki, H., Meissner, K., Gabbouj, M. *et al.* (2014) Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. *Ecological Informatics*, **20**, 1–12.

Kang, S.-H., Jeon, W. & Lee, S.-H. (2012) Butterfly species identification by branch length similarity entropy. *Journal of Asia-Pacific Entomology*, **15**, 437–441.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Vol. 2 (12). San Mateo, California, pp. 1137–1143.

Krell, F.-T. (2004) Parataxonomy vs. taxonomy in biodiversity studies–pitfalls and applicability of "morphospecies" sorting. *Biodiversity & Conservation*, **13**, 795–812.

Larios, N., Deng, H., Zhang, W. *et al.* (2007) Automated insect identification through concatenated histograms of local appearance features: feature vector generation and region detection for deformable objects. *Machine Vision and Applications*, **19**, 105–123.

Li, F. & Cao, W. (2015) Automated identification of butterfly species. *Journal of Computational Information Systems*, **11**, 2529–2538.

Lytle, D.A., Martínez-Muñoz, G., Zhang, W. *et al.* (2010) Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society*, **29**, 867–874.

MacLeod, N. (2007) *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, *The Systematics Association Special Volume Series*. CRC Press, Boca Raton, Florida.

MacLeod, N., Benfield, M. & Culverhouse, P. (2010) Time to automate identification. *Nature*, **467**, 154–155.

Miller, C.E., Chang, L., Beal, V., McDowell, R., Ortman, K. & LaCovey, T. (1992) *Risk Assessment of Mediterranean Fruit Fly*. USDA-APHIS, Washington, District of Columbia.

Nabhan, A.R. & Sarkar, I.N. (2012) The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics*, **13**, 122–134.

Norrbom, A.L. (1988) Revision of the *schaussi* group of *Anastrepha* Schiner (Diptera, Tephritidae), with a discussion of the terminology of the female terminalia in the Tephritoidea. *Annals of the Entomological Society of America*, **81**, 167–173.

Norrbom, A.L. (1998) A revision of the *Anastrepha daciformis* species group (Diptera: Tephritidae). *Proceedings of the Entomological Society of Washington*, **100**, 160–192.

Norrbom, A.L. (2002) A revision of the *Anastrepha serpentina* species group (Diptera: Tephritedae). *Proceedings of the Entomological Society of Washington*, **104**, 390–436.

Norrbom, A.L. (2009) A revision of the *Anastrepha robusta* species group (Diptera: Tephritidae). *Zootaxa*, **2182**, 1–91.

Norrbom, A.L. (2010) *The Diptera Site: Fruit Fly (Diptera: Tephritidae) Taxonomy Pages* [WWW document]. URL http://www.sel.barc.usda.gov/Diptera/tephriti/tephriti.htm [accessed on 19 June 2015].

Pape, T. & Evenhuis, N.L. (2013) *Systema Dipterorum*. Version 1.5 [WWW document]. URL http://www.Diptera.org [accessed on 19 June 2015].

Pelckmans, K., Suykens, J.A. & Gestel, T. (2002) *LS-SVMlab: A Matlab/C Toolbox for Least Squares Support Vector Machines*. Internal Report 02-44, ESAT-SISTA, K.U. Leuven Leuven.

Reinert, J.F., Harbach, R.E. & Kitching, I.J. (2004) Phylogeny and classification of Aedini (Diptera: Culicidae), based on morphological characters of all life stages. *Zoological Journal of the Linnean Society*, **142**, 289–368.

Rohlf, F.J. & Sokal, R.R. (1967) Taxonomic structure from randomly and systematically scanned biological images. *Systematic Biology*, **16**, 246–260.

Rueda, L.M. (2008) Global diversity of mosquitoes (Insecta: Diptera: Culicidae) in freshwater. *Hydrobiologia*, **595**, 477–487.

Russell, K.J., Do, M.T., Huff, J.C. & Platnick, N.I. (2000) Introducing SPIDA-Web: wavelets, neural networks and internet accessibility in an image-based automated identification system. *Automated Taxon Identification in Systematics: Theory, Approaches, and Applications*, pp. 131–152. CRC Press, Taylor & Francis Group, Boca Raton, Florida.

Santana, F.S., Costa, A.H.R., Truzzi, F.S., Silva, F.L., Santos, S.L., Francoy, T.M. & Saraiva, A.M. (2014) A reference process for automating bee species identification based on wing images and digital image processing. *Ecological Informatics*, **24**, 248–260.

Scholkopf, B. & Smola, A. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts.

Sieracki, J.M. & Benedetto, J.J. (2005) Greedy adaptive discrimination: component analysis by simultaneous sparse approximation. *Wavelets XI-59141R. Presented at the SPIE*. San Diego, California.

Sieracki, J.M., Crone, N.E. & Benedetto, J.J. (2008) Human electrocortigraphic signature determination by EGAD sparse approximation. *Proceedings of the Sensor*, *Signal and Information Processing (SenSIP) Workshop*. Sedona, Arizona.

Sivinski, J.M. & Dodson, G. (1992) Sexual dimorphism in *Anastrepha suspensa* (Loew) and other tephritid fruit flies (Diptera: Tephritidae): possible roles of developmental rate, fecundity, and dispersal. *Journal of Insect Behavior*, **5**, 491–506.

Sivinski, J. & Pereira, R. (2005) Do wing markings in fruit flies (Diptera: Tephritidae) have sexual significance? *Florida Entomologist*, **88**, 321–324.

Vaňhara, J., Muráriková, N., Malenovský, I. & Havel, J. (2007) Artificial neural networks for fly identification: a case study from the genera *Tachina* and *Ectophasia* (Diptera, Tachinidae). *Biologia*, **62**, 462–469.

Walter Reed Biosystematics Unit (WRBU) (2014) *Catalog of Culicidae* [WWW document]. URL http://www.mosquitocatalog.org/ [accessed on 10 November 2014].

Watson, A.T., O'Neill, M.A. & Kitching, I.J. (2003) Automated identification of live moths (Macrolepidoptera) using Digital Automated Identification SYstem (DAISY). *Systematics and Biodiversity*, **1**, 287–300.

Weeks, P.J.D., Gauld, I.D., Gaston, K.J. & O'Neill, M.A. (1997) Automating the identification of insects: a new solution to an old problem. *Bulletin of Entomological Research*, **87**, 203–212.

Wen, C., Guyer, D.E. & Li, W. (2009) Local feature-based identification and classification for orchard insects. *Biosystems Engineering*, **104**, 299–307.

World Health Organization (2013) *World Malaria Report 2013*. WHO Press, Geneva.

Yu, D.S., Kokko, E.G., Barron, J.R., Schaalje, G.B. & Gowen, B.E. (1992) Identification of ichneumonid wasps using image analysis of wings. *Systematic Entomology*, **17**, 389–395.

Zhang, A.B., Muster, C., Liang, H.B. *et al.* (2012) A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Molecular Ecology*, **21**, 1848–1863.

Accepted 16 August 2015
First published online 2 October 2015