Schroder S et al. 2002. The New Key to Bees: Automated Identification by Image Analysis of Wings. IN: Kevan P & Imperatriz Fonseca VL (eds) - Pollinating Bees - The Conservation Link Between Agriculture and Nature - Ministry of Environment / Brasília. p.209-216.

THE NEW KEY TO BEES: AUTOMATED IDENTIFICATION BY IMAGE ANALYSIS OF WINGS

Stefan Schroder, Dieter Wittmann, Wilhelm Drescher, Volker Roth, Volker Steinhage and Armin B. Cremers

ABSTRACT

World-wide studies on bee diversity, conservation and on pollination ecology are hampered by the difficult taxonomy of bees, the lack of suitable literature and bee taxonomists.

The automated identification system consists of an electronic notebook connected with a CCD camera mounted on a stereomicroscope. The identification of bees is based exclusively on characters of the fore-wing venation: The fore-wing is video-recorded and the image of the wing is transferred to the notebook. With a mouse-click the user marks defined vein junctions. The system then connects the junctions by automatic line-following and thus digitises the whole venation. The system has to be trained with a minimum of 30 well defined specimens of each sex per species. With data of each bee it learns and gets better. Species identification is achieved by automatic comparison of incoming data with already memorised data. Currently the system employs linear and non-linear discriminant analysis methods. We tested the system with very difficult cases like closely related species of *Andrena, Bombus* and *Colletes* which are a real problem for traditional taxonomy. In all cases the system identified the species with a confidence between 98 and 99,8%.

This system can be applied by museum taxonomists as well as by field workers with no special training in bee taxonomy. Dry specimens as well as live bees can be identified. Identification of a bee takes no more than 5 minutes. Wing images or readymade data can also be sent on disc or via internet to institutions which offer this automatic identification service.

INTRODUCTION

We can only monitor and conserve those animals that we know. For a long time all of us have been aware that studies on bee diversity, on conservation of bees and on pollination ecology are severely hampered by to

- the difficult taxonomy of bees,
- the lack of bee taxonomists
- and as a consequence the lack of classification literature such as modern identification keys and actual revisions of many taxa (O´Toole 1996).

These were the main reasons we developed a computer-based system for the automated identification of bees. Such a system should use the informations hidden in the wing venation. In the early stages of the development of the system (Schröder *et al*. 1994) many experts in bee taxonomy had severe doubts about this approach, as in classical bee taxonomy wings are rarely used for identification to the species level. Their general assumption was that there is too little discriminative information in the wing venation. Already in the1950's, when numerical taxonomy was developed as a helpful tool in taxonomy, other experts envisioned at least semi-automatic identification machines (Michener by personal

communication). However, at those times image processing and other computer tools were not available to realise such visionary plans.

Our automated identification system has the following advantages:

- it is small, mobile and handy so that it can be used in the field
- it works with live bees as well as with mounted collection specimens without removal of any body parts
- it works with a minimum of interaction by the user
- to operate the system, no knowledge of taxonomy is required
- only wing venation is used to identify bees to the species level.

The system should enable any person or working group which studies bees

- to either install and operate its own identification system or
- to send photographs of wings via mail or via internet to an institution which provides access to its identification system.

MATERIAL AND METHODS

Hardware

The identification system consists of an electronic notebook connected with a CCD camera which is mounted on a stereomicroscope (fig 1). The notebook is equipped with a standard video port in the PCMCIA-slot. The image of the wing is transferred from the camera to the notebook.

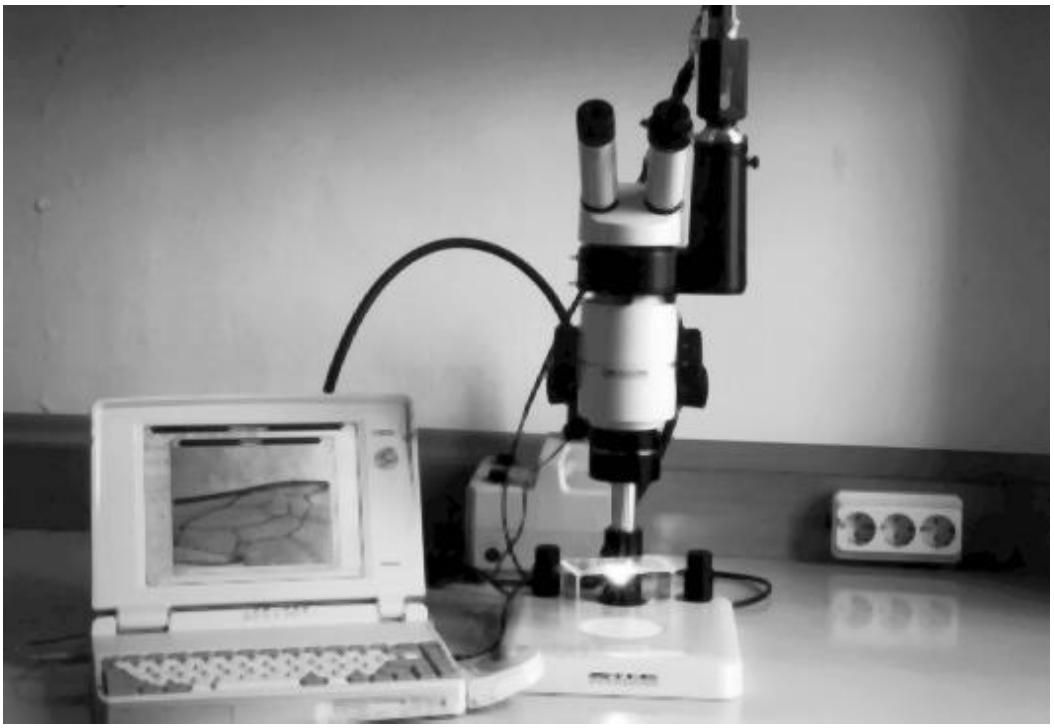


FIGURE 1: The portable identification system.

Software

Image processing: The identification of the bee is based on characters of the venation of the fore-wing like vein length, width, curvature, angles and descriptions of the cell area. To extract these parameters from the wing image we equipped the system with a modified automated line following program (Steinhage *et al*.1997).

Identification: For the identification the system employs the discriminant analysis. For statistical analysis we currently employ linear discriminant analysis (Hastie *et al.* 1994) but also non linear methods (Schölkopf *et al*. 1998). All identification processes are conducted by our newly developed and adapted programs.

RESULTS

A) Identification process

The result of our study is an identification process which consists of simple manipulations carried out by the user and the following 2 processes - image analysis and identification - which are conducted automatically by the system. However, it should be made clear that before any bee can be identified, the system has to be trained.

1. Training

For this it is necessary to have at least 30 specimens of each species. These bees must have been well identified by experienced bee taxonomists. In the training phase one fore-wing of each specimen has to be processed in the below described steps. The system memorises all data of each bee that participated in the training. With each further bee which is grouped into the training set the confidence of the identification will increase.

2. Image analysis

The first action of the user is to clip the fore-wing of the bee under a microscope slide and to video-record it. This procedure is done in a few seconds. The image of the wing appears on the screen and will be stored in the database. If a alive bee has been used, it can now be set free again.

Now follows the image analysis that consists of two steps.

a) Only in the first step the user interacts with the system. Supported by the program he marks defined vein junctions with a mouse-click. The system then connects the junctions by automatic line-following.

For the final version of the system we are actually working on the implementation of a completely automatic image analysis with which the system itself detects and marks the vein junctions.

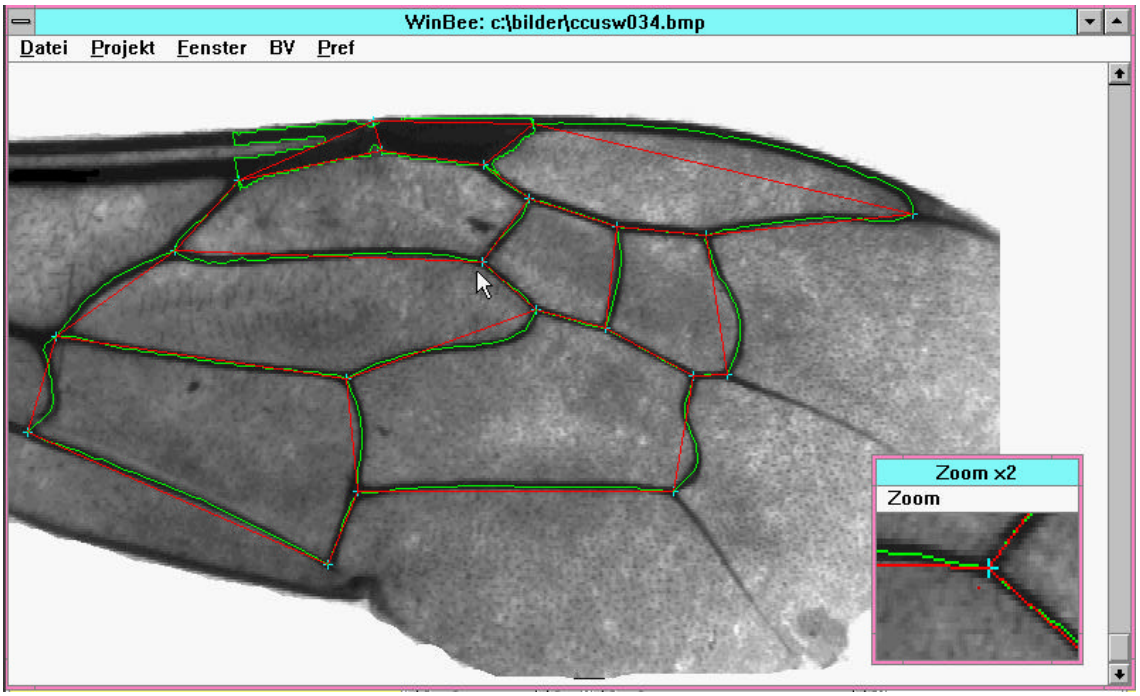


FIGURE 2: The vein junctions were zoomed and marked with a mouse click (insert lower right). The line following program has then digitised the wing venation.

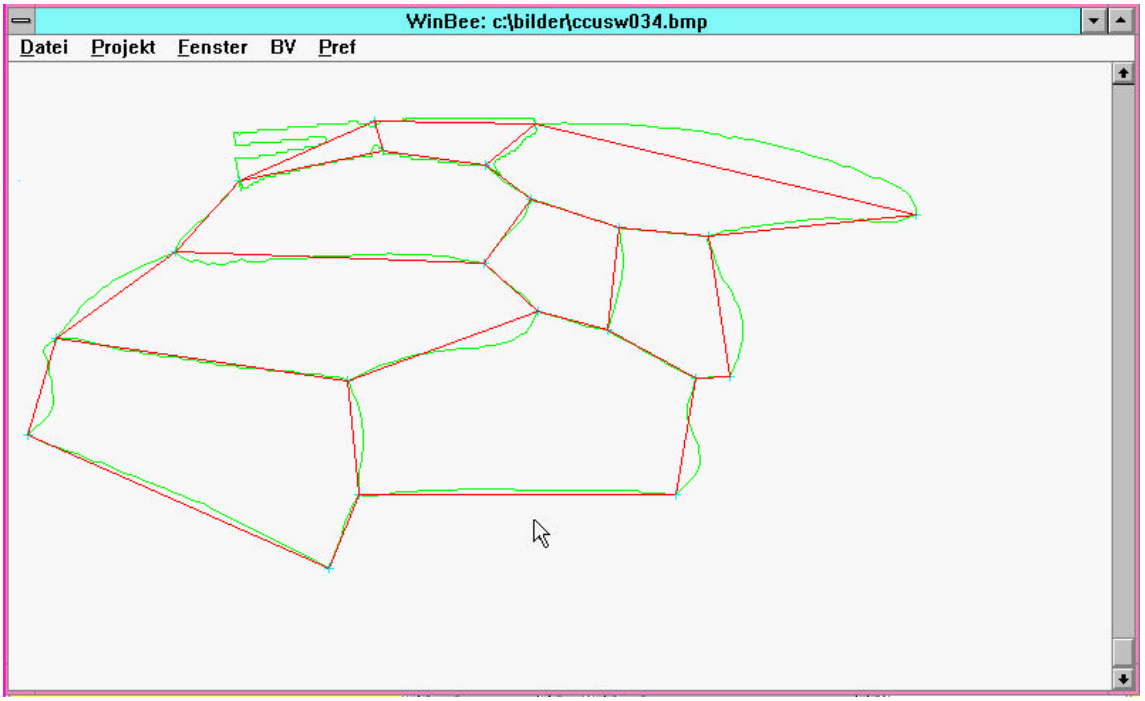


FIGURE 3: From the wing image (curved lines) the venation graph is extracted. This is the graph in which all vein junctions are connected by straight lines

As a prerequisite for any identification process, characters have to be named and measured. However, while the system follows single venation lines and measures all elements like the length of the veins, the angles between them and the area of all cells it has no information about the surroundings, for example, whether it is measuring the first or second cubital cell.

a) Therefore, in the second step of image processing the system has to name veins, angles and cells. This task is resolved by automatic comparison of the graph with model-venation-graphs from the database. Only when all features are named then all data can be stored in a data file together with the correct name of the measured item.

To represent all European bee genera the system needs only 9 model vein graphs. If we would add a South American bee that does not fit these graphs, the system would ask whether we want to incorporate this new wing graph as a model graph.

3. Identification

The image analysis results in a data file with about 200 measured features of the venation. All attributes and relations of the veins, junctions and cells of an extracted vein graph can be used as quantified characters for a statistical identification process. The system currently employs multivariate discriminant analyses that are implemented in a newly developed classifier for the automatic processing of the identification. In the first phase of the discriminant analysis the classifier uses the data of the training specimens to calculate the discriminant functions. In the second phase these functions are used with the data of the unknown bee to calculate its position in the multi-dimensional classification space.

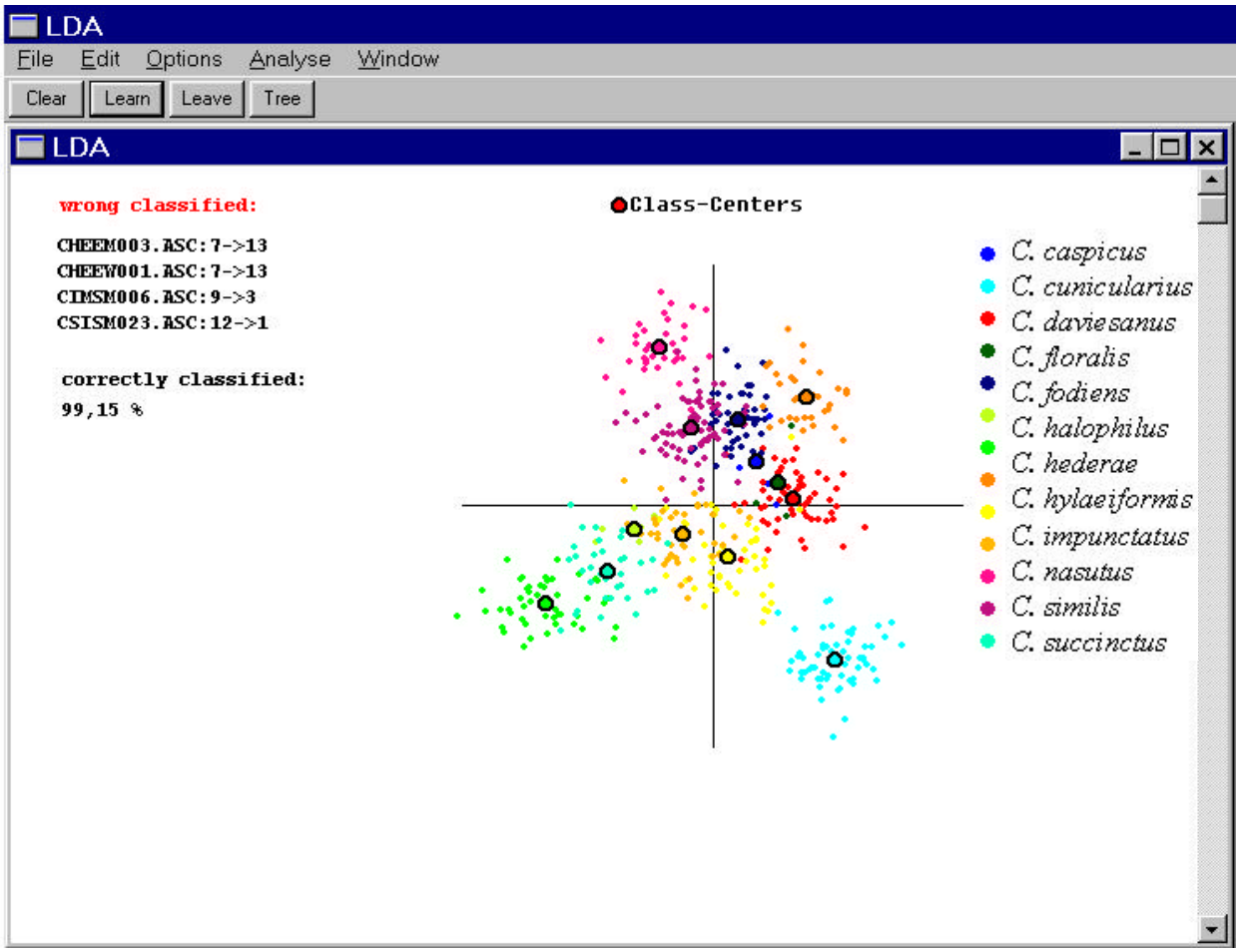


FIGURE 4: A screen snap that shows the status of the trained system ready for the identification of an unknown bee. Coloured dots (here grey scaled) represent the 469 specimens of 13 species. Names of the species are given on the right. Left: Code names of 4 training specimens that could not be grouped to the right clusters. Lower left: classification rate = 99.15%.

B) Application tests and confidence control

We tested the automatic identification steps with difficult cases of closely related species of European bees (see below). Here we present the test with 13 *Colletes* species, which is a real problem for traditional taxonomy.

1. Training of the classifier

For this test and for the confidence control the system has been trained with 469 specimens of the 13 *Colletes* species. In Fig.4 these specimens are represented by dots which form 13 clusters around their class centres. One should be aware that in the reality of the system these clusters are distributed in 12 dimensions. This means the clouds are much more distant from each other with little intermingling. After the training the system gave the list of those 4 bees which could not be attached to their species cluster. Reasons for this may be that these training bees were not correctly labelled or that they have aberrant venation. In this test 99,15% of the training bees were attached to the correct cluster (classification rate).

2. Identification of an unknown bee and confidence test

For the identification of an unknown bee the data of its image analysis are loaded and automatically processed in the trained classification program. The result of the identification is shown as a screen snap in Fig.5.

Any time after the training the user can check the system with the so called ”leave one out test” (Fig.5). This test measures the identification confidence with is achieved with the actual training set of bees. In this test the system calculates the position of each but one *Colletes* bee species in the database. It then incorporates the data of the one bee that was left out and then leaves out the data of another bee. It repeats this for each of the 469 specimens and then gives the confidence of the identification, in this case 98.3%. In general the identification will became even more confidential with an increasing number of training specimens.

A further ability of the program is to calculate the real distances between the clusters in the multi-dimensional classification space. These distances are represented in a dendrogram that gives a preliminary view of species similarity (insert in fig.5). Such dendrograms may be useful as a first indication of phylogenetic relations between species.

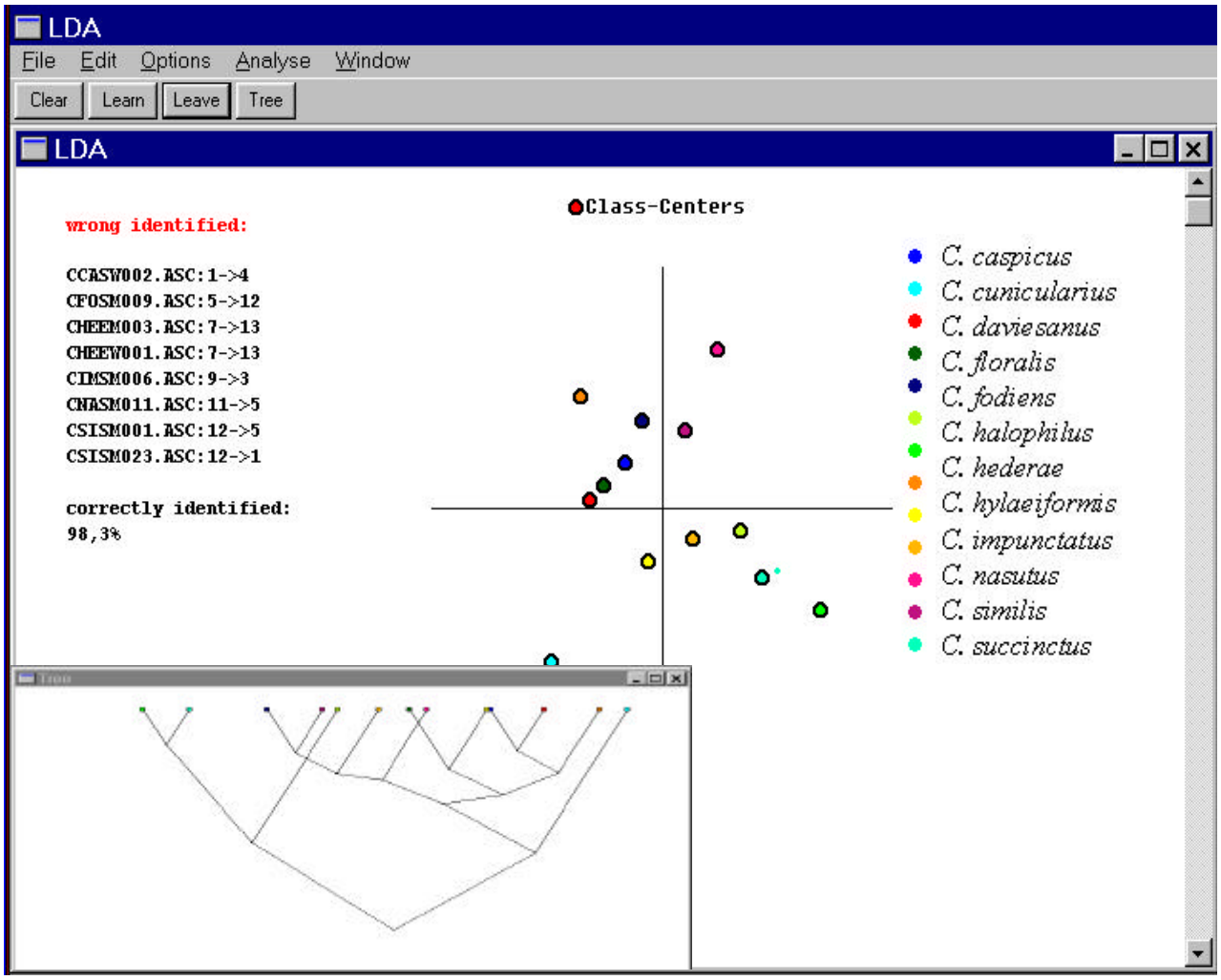
3. Further applications

The application of the system was also successfully tested with closely related species of the genera *Andrena*, *Osmia* and *Bombus (*Schröder *et al*. 1995, 1998*).*

Also, in all cases bees of different sexes could be distinguished by the system. In the case of social bees it succeeded in separating casts and different *Bombus* populations and even colonies. We also have adapted the system to cope with reduced wing

A further ability of the program is to calculate the real distances between the clusters in the multi-dimensional classification space. These distances are represented in a dendrogram that gives a preliminary view of species similarity (insert in fig.5). Such dendrograms may be useful as a first indication of phylogenetic relations between species.

FIGURE 5: Screen snap showing the result of the identification: The single dot (marked by an arrow) represents the unknown specimen. It was identified to belong to the cluster of *Colletes succinctus*. Upper left: list of those bees that could not be identified during the “omit one” test. Insert: Similarity dendrogram calculated from the distances between the clusters.



4. Further applications

The application of the system was also successfully tested with closely related species of the genera *Andrena*, *Osmia* and *Bombus (*Schröder *et al*. 1995, 1998*).*

Also, in all cases bees of different sexes could be distinguished by the system. In the case of social bees it succeeded in separating casts and different *Bombus* populations and even colonies. We also have adapted the system to cope with reduced wing venation as in stingless bees. Analyses of the ”similarity” of populations could give new and important informations for species conservation.

The system was also successfully applied to identify other Hymenoptera like wasp species of the genera *Ceramius* (Masarinae) (Mauss 1998).

DISCUSSION

This identification system is by no means a substitute for well-trained taxonomists. Rather, it requires them to establish training a set of well-identified specimens. Once trained and installed, the system unburdens taxonomists from routine identification jobs giving them more

time for the scientific work on species descriptions and revisions etc. As the system employs statistical identification methods the results can be checked by a confidence test. This is a great advantage in comparison with the use of conventional keys.

The data in the training set can easily be copied and made available for other researchers who can work with it and eventually add further data to it. Thus the data, which can be viewed as a digitised reference collection, can rapidly be increased and multiplied. They can swiftly be given from one working group to another: Wing images or readymade data can be sent on disc or via internet to institutions which offer this automatic identification service.

Important for the effective operation of the system is a large pool of training data. To build up this database, we suggest that those institutions which want to employ the system should form a network. Museums, universities, private institutions and any other students of bees should co-operate in the exchange of bee data in order to create the basis for an automated identification of bees on the local, regional or countrywide level.

REFERENCES

Hastie T, Tibshirhani R, Buja A. Flexible discriminant analysis by optimal scoring. JASA 1994; 89: 1255–70.

Mauss V. Taxonomie und Biographie nordafrikanischer Pollenwespen der Gattung C*eramius*: einsatz morphometrischer methoden bei der taxonomischen entscheidungs-findung (Hymenoptera, Vespidae). Beitr. d. Hymenopt.-Tag. 1998; 18.

O´Toole C. Bee systematics in Europe: the continuing crisis and some possible cures. In: Matheson A, Buchmann SL, O'Toole C, Westrich P, Williams IH, editors. The conservation of bees [Based on Symposium organized jointly by the International Bee Research Association and the Linnean Society of London, held in April, 1995]. London: Academic Press; 1996. p.227-32. (Linnean Society Symposium Series, 18)

Schölkopf B, Smola A, Müller KR. Non-linear component analysis as a kernel eigenvalue problem. Neural Computation 1998; 10(5): 1299–398.