Computers and Electronics in Agriculture 114 (2015) 68–77

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Evaluating classification and feature selection techniques for honeybee subspecies identification using wing images

Felipe Leno da Silva [a,*], Marina Lopes Grassi Sella [c], Tiago Mauricio Francoy [b], Anna Helena Reali Costa [a]

[a] Escola Politécnica da Universidade de São Paulo, Av. Prof. Luciano Gualberto, travessa 3, 158, São Paulo, SP CEP: 05508-970, Brazil
[b] Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, Rua Arlindo Béttio, 1000, São Paulo, SP CEP: 03828-000, Brazil
[c] Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes, 3900, Ribeirão Preto, SP CEP: 14049-900, Brazil

ARTICLE INFO

Article history:
Received 9 August 2014
Received in revised form 12 March 2015
Accepted 17 March 2015

Keywords:
Geometric morphometrics
Machine learning
Apis mellifera
Feature selection

ABSTRACT

The main pollinator commercially available, i.e. *Apis mellifera*, is now facing a severe population decrease worldwide due to the so-called Colony Collapse Disorder. Measures to preserve this species are urgent. Honeybees inhabit several different environments, from swamps to deserts, from high mountains to the African savannah. They are classified into several different subspecies, each one adapted to a particular set of environmental characteristics. The identification of subspecies is based on morphometric features from the entire bee body, but in the last years features from the fore wings have proven to be very efficient for classification. Several methods have been developed to perform the automatic classification through images of bee wings, and geometric morphometrics has been reported to achieve good results in terms of consumed time and reliability of the results. However, there has been no study evaluating the impact of feature selection and new classification methods on the identification performance. We here evaluate seven combinations of feature selectors and classifiers by their hit ratio with real bee wing images. Feature selection proved to be beneficial to all the evaluated combinations and the Naïve Bayes classifier combined with a correlation-based feature selector achieved the best results. These conclusions can benefit researches that rely on classification by geometric morphometrics features, both for bees and for other animal species.

1. Introduction

In its original range of distribution, the honeybee (*Apis mellifera*) originally occupied large portions of Africa, Europe and the Middle East; later, it was introduced in the Americas and Australia (Ruttner, 1988). The honeybee occupies several different environments, from mountains to swamps, from deserts to wet lands, and is currently classified into about 30 subspecies, each one adapted to a particular set of environmental characteristics (Ruttner, 1988). The most widely accepted classification system is based on morphometric features, including several measures of the body, wing venation angles and pigmentation (Ruttner et al., 1978). Albeit efficient, classification with such features is very time-consuming. In recent years, wing venations have produced good classification of *Apis mellifera* subspecies (Francoy et al., 2008, 2009; Tofilski, 2008; Oleksa and Tofilski, 2014), and software packages have been developed for automated identification. For instance, the ABIS (Automated Bee Identification System) system (Steinhage et al., 2006) obtained good results in bee identification, even though its feature extraction method prevents the identification of stingless species. The DrawWing system (Tofilski, 2004) allows the automatic identification of the landmarks for some genera; however, it is only possible to perform identification in three species previously defined by the author, and this software cannot be trained to identify new species.

Automated bee identification through wing images can be divided into four sub-processes: Image Acquisition, Digital Image Processing, Classification, and Validation (Santana et al., 2014). In the Digital Image Processing sub-process, usually the vein junctions in the wing are marked (called *landmarks*, shown in Fig. 1) and the features extracted for classification are based on these landmarks (Koca and Kandemir, 2013; Miguel et al., 2011; Steinhage et al., 2006; Santana et al., 2014). In the Classification sub-process, the choice of suitable features and classifier is fundamental to achieve a good classification performance (Witten et al., 2011). Linear Discriminant Techniques are used in most of the works in which bee species identification is performed (Francoy et al., 2008, 2012a; Koca and Kandemir, 2013;

* Corresponding author. Tel.: +55 11 3091 5397.
E-mail addresses: f.leno@usp.br (F.L.d. Silva), marinalg@usp.br (M.L. Grassi Sella), tfrancoy@usp.br (T.M. Francoy), anna.reali@usp.br (A.H.R. Costa).

http://dx.doi.org/10.1016/j.compag.2015.03.012
0168-1699/

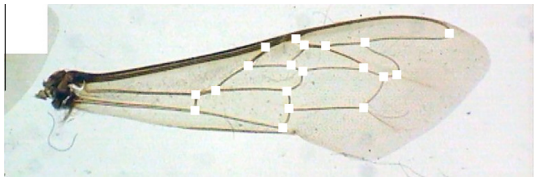

Fig. 1. *Apis mellifera adami* forewing with landmarks (squares on the vein junctions).

Meulemeester et al., 2012; Michez et al., 2009); however, a few studies showed that non-linear classifiers can improve the classification rate in some situations (Santana et al., 2014; Roth et al., 1999).

No systematic study has assessed the impact of feature selection techniques prior to classification in this domain. Feature selection removes irrelevant or noisy features, and tends to improve classification accuracy (Dash et al., 2002). In practice, feature selection has been able to improve the performance of classifiers in others domains (Silva et al., 2013; Chen et al., 2011). Even though Santana et al. (2014) showed a ranking of classifiers by classification rate, this ranking could be changed with the introduction of feature selection techniques, as the classifiers do not usually deal with redundant or noisy features in the same way (Witten et al., 2011).

This article investigates whether feature selection can improve the classification rate, and seeks to determine the best combination of feature selection techniques and classifiers in the context of honeybee subspecies identification. All the conclusions of this article can benefit researches on both bee and other animal species identification. The article is organized as follows: Section 2 describes which features were used and how they were extracted from bee wing images; Section 3 describes the classifiers evaluated; while Section 4 presents the feature selection methods evaluated; in Section 5, we explain how we chose the combinations of classifiers and feature selectors for the experiments, while describing how the experiments were performed and analyzed in Section 6; Section 7 shows the experimental results and their discussion; and, finally, Section 8 concludes the article, outlining the results and open questions to be analyzed in further works.

2. Feature extraction

Geometric Morphometrics has achieved excellent results in feature extraction for bee species identification (Tofilski, 2008; Koca and Kandemir, 2013; Miguel et al., 2011). However, there is no clear consensus about which and how many features should be used to maximize the classification rate, as the number and type of features vary in different articles. The goal of feature selection is to automatically choose the relevant features (Witten et al., 2011). We used feature selection techniques to select among many features and to find the optimum set of features that maximize the classification rate.

We extracted all landmark-based features from Geometric Morphometrics that had already been used successfully for bee species identification in the literature (Francoy et al., 2008, 2011, 2012a,b; Tofilski, 2008; Santana et al., 2014; Koca and Kandemir, 2013; Meulemeester et al., 2012; Kandemir et al., 2011). Since all features rely on the landmarks position, we started by choosing the 19 landmarks shown in Fig. 1 using the tpsDig software (Rohlf, 2010). Then, we extracted the Centroid Size (*CS*), a widely used feature from Geometric Morphometrics (Tofilski, 2008; Francoy et al., 2012a; Meulemeester et al., 2012), as follows (Bookstein, 1991):

$$CS = \sqrt{\sum_{i=1}^{|\boldsymbol{L}|}(x_i - \overline{x})^2 + (y_i - \overline{y})^2}, \tag{1}$$

where the set of landmarks $\boldsymbol{L}$ is composed of $|\boldsymbol{L}| = 19$ landmark coordinates $(x_i, y_i), i = 1, 2, \ldots, |\boldsymbol{L}|$, and $(\overline{x}, \overline{y})$ is the centroid of $\boldsymbol{L}$. The centroid can be computed as follows:

$$\overline{x} = \frac{1}{|\boldsymbol{L}|}\sum_{i=1}^{|\boldsymbol{L}|} x_i \quad \text{and} \quad \overline{y} = \frac{1}{|\boldsymbol{L}|}\sum_{i=1}^{|\boldsymbol{L}|} y_i. \tag{2}$$

where (x_i, y_i) are the coordinates of landmark i. As can be noted by Eq. (1), the calculation of the Centroid Size is only dependent on the landmark positions (defined individually wing by wing), thus, when calculating the features values for unlabeled wings, this procedure remains the same.

After the Centroid Size computation, the next extracted features are the Aligned Coordinates. These features are used in most of the literature (Francoy et al., 2008; Tofilski, 2008; Koca and Kandemir, 2013; Francoy et al., 2011, 2012a,b; Meulemeester et al., 2012; Kandemir et al., 2011) and consist in performing affine transformations on the configurations of landmarks, aiming at configurations of coordinates invariant to translation, scale and rotation (called Aligned Coordinates). This procedure can be performed in a number of ways, e.g. calculating the Bookstein coordinates (Bookstein, 1991) or performing an Orthogonal Procrustes Analysis (Rohlf and Slice, 1990).

In order to provide translation, scale and rotation invariance, we performed an Orthogonal Procrustes Analysis (Rohlf and Slice, 1990) in all the landmark configurations, as follows:

$$\boldsymbol{L}' = \frac{(\boldsymbol{I} - \boldsymbol{N})\boldsymbol{L}}{s}, \tag{3}$$

where

$$s = \sqrt{tr((\boldsymbol{I} - \boldsymbol{N})\boldsymbol{L}\boldsymbol{L}^{\boldsymbol{t}}(\boldsymbol{I} - \boldsymbol{N}))} \tag{4}$$

and $\boldsymbol{L}$ is the matrix with the (x, y) coordinates of all the landmarks, $\boldsymbol{I}$ is an $|L| \times |L|$ identity matrix, $\boldsymbol{N}$ is an $|L| \times |L|$ matrix with all elements equal to $\frac{1}{|L|}$ and $tr(A)$ refers to the sum of all the principal diagonal elements of A.This step will provide translation and scale invariance for all landmark configurations. In order to achieve the rotation invariance, rotation matrix $\boldsymbol{H}$ must be defined, and the Aligned Coordinates are calculated in Eq. (5). The rotation matrix gives the best approximation, $\boldsymbol{L}^*$, comparing to a reference landmark configuration (Rohlf and Slice, 1990).

$$\boldsymbol{L}^* = \boldsymbol{L}'\boldsymbol{H}. \tag{5}$$

where $\boldsymbol{L}^*$ are the final Aligned Coordinates, $\boldsymbol{L}'$ is defined in Eq. (3) and $\boldsymbol{H}$ is the rotation matrix. Finally, after this procedure is applied to all wings, the translation, scale and rotation invariance is achieved.

Principal Warps (Bookstein, 1989) is a tool to shape variation analysis, and has successfully been used to extract features for bee species recognition (Meulemeester et al., 2012; Francoy et al., 2012b). The first step is to define a reference object (called *consensus* object) to be used in the ensuing computations. In this article, the consensus object was defined by calculating the mean value of the (x, y) coordinates for each landmark of all the samples of our entire dataset, resulting in $2|\boldsymbol{L}|$ coordinates. Note that the coordinates must be aligned (Eq. (3)) prior to this computation.

The next step is to compute the bending energy matrix $\left(\boldsymbol{B}_{|L|}^{-1}\right)$ for the consensus object. This can be done by assembling the partitioned matrix (Rohlf, 1993):

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{Q}^{\boldsymbol{t}} & \boldsymbol{O} \end{bmatrix}, \tag{6}$$