Journal of Food Engineering 112 (2012) 50–59

Contents lists available at SciVerse ScienceDirect

Journal of Food Engineering

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

Discernment of bee pollen loads using computer vision and one-class classification techniques

Manuel Chica [a,b,*], Pascual Campoy [c]

[a] Inspiralia Tecnologías Avanzadas, Estrada 10, 28034 Madrid, Spain
[b] European Centre for Soft Computing, Gonzalo Gutiérrez Quirós, 33600 Mieres, Asturias, Spain
[c] Automatics and Robotics Center – Universidad Politécnica de Madrid, José Gutiérrez Abascal 2, 28006 Madrid, Spain

ARTICLE INFO

Article history:
Received 14 December 2011
Received in revised form 19 March 2012
Accepted 21 March 2012
Available online 29 March 2012

Keywords:
Bee pollen
Food authentication
Outliers detection
One-class classification
Computer vision

ABSTRACT

In this paper, we propose a system for authenticating local bee pollen against fraudulent samples using image processing and classification techniques. Our system is based on the colour properties of bee pollen loads and the use of one-class classifiers to reject unknown pollen samples. The latter classification techniques allow us to tackle the major difficulty of the problem, the existence of many possible fraudulent pollen types.

Also presented is a multi-classifier model with an ambiguity discovery process to fuse the output of the one-class classifiers. The method is validated by authenticating Spanish bee pollen types, the overall accuracy of the final system of being 94%. Therefore, the system is able to rapidly reject the non-local pollen samples with inexpensive hardware and without the need to send the product to the laboratory.

1. Introduction

The bee-keeping sector has a notable socio-economic relevance in Europe, according to the FAO Agricultural Statistics Division. Although honey is the most important bee product, there are other well-known products that result from bee-keeping activity, such as pollen or royal jelly. Bee pollen production, for both domestic and foreign markets, is considered by many bee-keepers a means of diversification and increasing their income. Furthermore, bee pollen products are considered an important food supplement, and can be used in medical treatments, although they are not scientifically recognised.

Bee-keepers, bee-keeping associations, and laboratories are interested in detecting fraud in pollen, and require tools to standardize and authenticate bee pollen origin in order to guarantee its nutritive and health benefits. Microscopic analysis of pollen grains, which form bee pollen loads, is a precise method of identifying origin. However, this process requires the laboratory work of melissopalynology experts, and is thus time consuming and costly. There have been many attempts to automate pollen grain identification by computer algorithms but there is no inexpensive, complete, and automated process (Allen, 2006; Boucher et al., 2002; France et al., 2000; Rodríguez-Damián et al., 2006).

Experts use macroscopic identification of bee pollen loads by means of such properties as colour. This method, although it cannot guarantee complete accuracy, can provide an initial, reliable idea of bee pollen origin (Kirk, 1994). Also, some melissopalynology experts separate pollen load samples by colour as a previous step to final microscopic identification (Campos et al., 1997; de Sá Otero et al., 2002). This process is carried out manually by experts who spend more than an hour separating each sample, an indication of the difficulty and subjectivity involved. Thus, the development of a completely automated, inexpensive system which can recognize external pollen load properties, such as colour, can bring about a twofold improvement in bee pollen origin authentication: (a) recognition of local bee pollen by non-experts (bee-keeping associations, for instance) and (b) reduction of laboratory work by experts through separating the pollen loads automatically.

Developing an automatic system to recognize and separate the pollen loads by colour is a highly complex task, requiring a specific solution. Even within a single pollen type, colour variability is high due to environmental characteristics, such as humidity during the plant growth, as well as the drying process of the pollen, or the presence of impurities. In addition, classification of known local pollen loads must be made against all other world pollen types. This is an important obstacle to the designing of an automated system, since colour data cannot be collected from all existing bee pollen types. In order to overcome these main difficulties, we propose a novel bee pollen load classification system based on image processing and one-class classification. The use of computer vision and

* Corresponding author at: Inspiralia Tecnologías Avanzadas, Estrada 10, 28034 Madrid, Spain. Tel.: +34 985456545.
E-mail address: manuel.chica@softcomputing.es (M. Chica).

0260-8774/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jfoodeng.2012.03.028

classification techniques is not new in the development of food quality control systems and has performed well in many situations (Mery et al., 2010; Kang and Sabarez, 2009; Du and Sun, 2004, 2006).

In our case, the well-known mean shift algorithm (Comaniciu and Meer, 2002) is used to filter and homogenize pollen load colour information. Moreover, one of principal novelties of our proposed method is the use of one-class classification (Moya et al., 1993; Ritter and Gallegos, 1997; Chandola et al., 2009; Tax, 2001), which was introduced as a classification paradigm for detecting anomalies or outliers in data distribution; that is, when there is enough data to model the positive classes but there is limited data on the negative classes. This characteristic is ideal for dealing with our problem, since local pollen types can be modelled, but not all possible fraudulent pollen types. The application field of one-class classification is enormous, from the fraud detection (Taniguchi et al., 1998; Phua et al., 2004) to image processing area (Pokrajac et al., 2007; Augusteijn and Folkert, 2002).

In addition to one-class classification, a multi-classifier algorithm was designed to aggregate one-class classifier outputs, given a unique response with a confidence measure. Unlike existing schemes, our fusion algorithm incorporates an ambiguity discovery mechanism to find pollen types with identical colour properties, in the case of which the system must behave with sufficient robustness.

The proposed method has been validated for the authentication of the most common Spanish pollen types, *Cistus ladanifer*, *Rubus*, *Echium*, and *Quercus ilex*, with respect to non-Spanish pollen types (de Sá Otero et al., 2002). In total, a dataset of around 2000 instances has been used to validate the trained system. In addition, a comparison of one-class classifier approaches has been done with four different algorithms: a Gaussian estimator, a Parzen classifier, a support vector data description (SVDD), and a k-nearest neighbours (kNN) technique. The classifiers were validated using ROC analysis and classification accuracy indicators.

This paper is structured as follows. In Section 2, the bee pollen images and the system used to acquire the images are described. Also, the proposed method, formed by image processing algorithms and one-class and multi classifiers is given in the same section. The developed experiments are presented and analyzed in Section 3. Finally, in Section 4, some concluding remarks and proposals for future work are made.

2. Materials and methods

2.1. Bee pollen problem description

Different samples of Spanish bee pollen loads were obtained from bee-keepers to build the authentication models and validate them against non-local samples. Samples belonging to the four local pollen types (*Rubus*, *Echium*, *Cistus ladanifer*, and *Quercus ilex*) and non-local samples were identified and grouped by experts. Images where these pollen loads samples appear are in Fig. 1. In this figure it can be observed how, even for experts, colour separation is difficult and subjective, and how non-local samples can be misleading (image on the right of the figure).

A computer vision system was used to take images from each of the pollen load samples, as shown in Fig. 2. It is composed of an inexpensive camera, an LED illumination system, and a tray for sample placement. This set of images is the input of the system and the pollen loads appearing in those images must be identified as local or unknown pollen types.

2.2. Outline of the proposed method

An overview of the proposed system can be observed in the diagram of Fig. 3. In this diagram, the initial acquired image of the pollen loads by computer vision hardware as explained in Section 2.1 can be seen. This image is segmented and processed by the mean shift algorithm (see Section 2.3).

Then, the colour instances of the pollen loads are used to train a multi-classifier model based on one-class classifiers (one for each local pollen type). See Sections 2.4, 2.5, and 2.6 for the methods used in this part of the system.

Finally, the multi-classifier outputs the authentication of each colour instance, classifying them as a known local or non-local (outlier) pollen type.

2.3. Image processing

In the following sections we will describe the segmentation and selection of the colour space (Section 2.3.1) as well as the used image filtering algorithm (Section 2.3.2).

2.3.1. Segmentation algorithm and colour space

First, the well-known Otsu segmentation algorithm (Otsu, 1979) is applied to the gray-scale image to extract the pollen loads from the background. Then, a morphological opening operation is applied to the thresholded binary image (Gonzalez and Woods, 2008). This operation removes those small objects having less than 50 connected pixels in a 8-connected neighbourhood.

Pixels extracted in the latter phase are to be analysed by the remaining processing algorithms. Their colour information can be represented in several ways. The most common is the RGB space where colours are represented by their red, green, and blue components in an orthogonal Cartesian space. However, the RGB space does not lend itself to mimicking the higher level processes which allow human colour perception. Colour is better represented in terms of hue, saturation, and intensity, as HSI or HSV spaces do (Lucchese and Mitra, 2001).

However, the latter colour spaces are not perceptually uniform. The CIE $L^*u^*v^*$ and $L^*a^*b^*$ are ideal for colour recognition because of the following three properties: (a) separation of achromatic information from chromatic information, (b) uniform colour space, and (c) similarity to human visual perception (Wyszecki and Stiles, 1982). In these colour spaces, for instance, the Euclidean distance between two colour points can be easily calculated, as in Eq. (1). This property will ease the work of the classification algorithms

$$D_{12} = \sqrt{(L_2^* - L_1^*)^2 + (u_2^* - u_1^*)^2 + (v_2^* - v_1^*)^2} \quad (1)$$

Although there are many other possible colour spaces, e.g. CIE *YUV* or CIECAM02, we have used the CIE $L^*u^*v^*$ colour space because of its good results in different colour computerised applications (Chen et al., 2004; Gökmen et al., 2007).

2.3.2. Mean shift filtering

Each extracted pollen load from the image has many different $L^*u^*v^*$ colour values, possibly as many as pixels contained in the load. This has a negative impact on pollen load colour authentication, since human experts customarily identify each pollen load as a unique colour. Therefore, there is a need for a image processing algorithm before applying the classification methods. The goal is to homogenize and smoothen the large quantity of different colour points of a pollen load in just a few unique and representative colour instances.

One of the best methods for discontinuity-preserving smoothening in image processing is the mean shift algorithm, proposed in Comaniciu and Meer (2002), and in line with the feature space analysis. The main strengths of the mean shift algorithm are: (a) it is an application independent tool, (b) it is suitable for real data analysis, (c) it does not assume any prior shape on data clusters, (d) it can handle arbitrary feature spaces, and (e) it has only one parameter, the bandwidth selection window size, to be chosen.