Improving Spider Recognition Based on Biometric Web Analysis

Carlos M. Travieso Gonzalez[1], Jaime Roberto Ticay-Rivas[1], Marcos del Pozo-Baños[1], William G. Eberhard[2], and Jesús B. Alonso-Hernández[1]

[1] Signals and Communications Department
Institute for Technological Development and Innovation in Communications
University of Las Palmas de Gran Canaria
Campus University of Tafira, 35017, Las Palmas de Gran Canaria, Las Palmas, Spain
[2] Smithsonian Tropical Research Institute and Escuela de Biologia
Universidad de Costa Rica, Ciudad Universitaria, Costa Rica
{ctravieso,jalonso}@dsc.ulpgc.es, {jrticay,mpozo}@idetic.eu,
william.eberhard@ucr.ac.cr

Abstract. This work presents an improvement of the automatic and supervised spider identification approach based on biometric spider web analysis. We have used as feature extractor, a Joint Approximate Diagonalization of Eigen-matrixes Independent Component Analysis applying to a binary image with a reduced size (20×20 pixels) from the colour original image. Finally, we have applied a least square support vector machine as classifier, reaching over 98.15% in our hold-50%-out validation. This system is making easier Biologists' tasks in this field, because they can have a second opinion or have a tool for this work.

Keywords: Spider webs, spider classification, independent component analysis, support vector machine.

1 Introduction

The pollution and socioeconomic growth is generating serious problems in our actual world and one of big handicap is the loss diversity in natural environments. Therefore, biodiversity conservation has become a priority for researchers [1]. Knowledge about species is critical to understand and protect the biodiversity of life on Earth. Sadly, spiders have been one of most unattended groups in conservation biology [2]. These arachnids are plentiful and ecologically crucial in almost every terrestrial and semi-terrestrial habitat [3-5]. Moreover, they present a series of extraordinary qualities, such as the ability to react to environmental changes and anthropogenic impacts [5-6].

Several works have studied the spider behaviour. Some of them analyse the use of the way spiders build their webs as a source of information for species identification [7-8]. Artificial intelligent systems have been proven to be of use for the study of the spider nature. In [9], Authors proposed a system for spider behaviour modelling, which provides simulations of how specific spider specie builds its web. In [10], it is

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 438–446, 2012.

recorded how spiders build their webs in a controlled scenario for further spatial-temporal analysis.

Due to spider webs carry an incredibly lot of information, this work proposed the used of them as a source of information for spider specie identification. From our point of view, this work has improved the first version [11], increasing the success on 4%. Independent Component Analysis has been used as biometric features for this purpose. This feature extraction, added to image processing tools for preparing images, and Least Square Support Vector Machines for classification, has reached an improvement vs. our previous work [11].

The remainder of this paper is organized as follow. First, our pre-processing system is presented in order to detect the spider webs from the background. Section 3 explains how feature extraction images were applied. Least Square Support Vector Machine is introduced in section 4. Next, experimental settings are shown by the dataset, the experimental methodology, results and the discussion. Finally in section 6, conclusions derived from the results are presented.

2 Pre-processing System

Spider web images were taken in both controlled and uncontrolled environments. Thus, the pre-processing step was vital in order to isolate the spider webs and remove possible effects of background in the system's results.

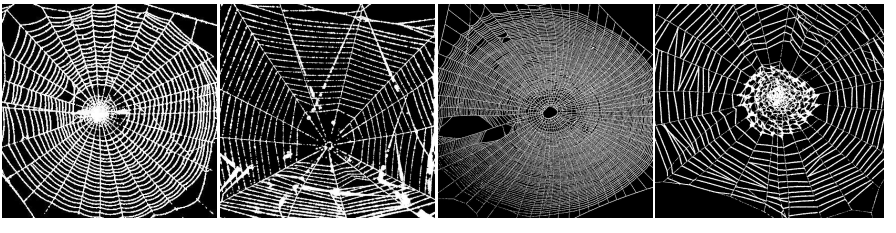

Fig. 1. Examples of full spider web images after preprocessing corresponding to *Allocyclosa*, *Anapisona Simoni*, and *Micrathena Duodecimspinosa, Zosis Geniculata,* respectively

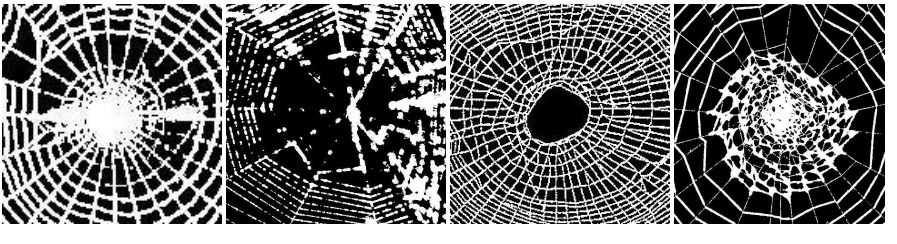

Fig. 2. Examples of centre spider web images after preprocessing corresponding to *Allocyclosa*, *Anapisona Simoni*, and *Micrathena Duodecimspinosa, Zosis Geniculata,* respectively

To enhance the contour of cobweb's threats an increase of colour contrast was first applied. Then, images were multiplied by two to further intensify the spider webs in relation to the background. Once images have been enhanced, they were binarized by

Otsu's Method [12] and cleaned up by morphological transformations [13]. Finally, the spider webs were cropped following two criterions. As results, two full set of spider web images were obtained. One of both contains the full web and another one shows only the central area. Examples of these sets can be seen in Fig. 1 and 2 respectively. Finally, all images were normalized to dimensions 20 x 20 pixels.

3 Feature Extraction: Joint Approximate Diagonalization of Eigen-Matrixes Independent Component Analysis

Independent Component Analysis (ICA) is a particularization of Principal Component Analysis (PCA) to extract components that are, at the same time, non-gaussian and statistically independent [14]. When used on images, ICA obtains independent base images which are not necessarily orthogonal. Application of these base images extracts between pixels information related to high order statistics.

In this study, an approach based on Joint Approximate Diagonalization of Eigen-matrixes Independent Component Analysis (JADE-ICA) has been used to implement this tool. JADE-ICA is based on joint diagonalization of cumulant matrixes. For simplicity, the case of symmetric distributions is considered, where the odd-order cumulants vanish. Let $X_1, \ldots, X_4$ be random variables, and defined $X_i^* = X_i + E(X_i)$. The second order cumulants can be written as:

$$C(X_1, X_2) = E(X_1^*, X_2^*) \tag{1}$$

And the fourth-order cumulants as:

$$\begin{aligned} C(X_1, X_2, X_3, X_4) = {} & E(X_1^*, X_2^*, X_3^*, X_4^*) - E(X_1^*, X_2^*)E(X_3^*, X_4^*) - \\ & E(X_1^*, X_3^*)E(X_2^*, X_4^*) - E(X_1^*, X_4^*)E(X_2^*, X_3^*) \end{aligned} \tag{2}$$

In addition, the definitions of variance and kurtosis of a random variable X are:

$$\begin{aligned} \sigma^2 &= C(X, X) = E(X^{*2}) \\ kurt(X) &= C(X, X, X, X) = E(X^{*4}) - 3E^2(X^{*4}) \end{aligned} \tag{3}$$

Now, under a linear transformation $Y=AX$, the cumulants of fourth-order transformation became:

$$C(Y_i, Y_j, Y_k, Y_l) = \sum_{p,q,r,s} a_{ip} a_{jq} a_{kr} a_{ls} C(X_p, X_q, X_r, X_s) \tag{4}$$

with a_{ij} the i-th row and j-th column entry of matrix A. Since the ICA model ($X=AS$) is linear, using the assumption of independence by $C(S_p,S_q,S_r,S_s)=kurt(S_q)\delta_{pqrs}$ where:

$$\delta_{pqrs} = \begin{cases} 1 & if \quad p = q = r = s \\ 0 & otherwise \end{cases} \tag{5}$$

and S has independent entries:

$$C(Y_i, Y_j, Y_k, Y_l) = \sum_{m=1}^{n} kurt(S_m) a_{im} a_{jm} a_{km} a_{lm} \tag{6}$$

the cumulants of the ICA model are obtained.

Given any *n x n* matrix *M* and a random *n x 1* vector *X*, we consider a cumulant matrix $Q_x(M)$ defined by:

$$Q_X(M) = \sum_{m=1}^{n} C(X_i, X_j, X_k, X_i) M_{ki} \tag{7}$$

If *X* is centered, the definition of (4) shows that:

$$\begin{aligned} Q_X(M) = E\left\{(X^T M X^T) X X^T\right\} - R^X tr(M R^X) - \\ R^X M R^X - R^X M^T R^X \end{aligned} \tag{8}$$

where *tr(B)* denotes the trace of matrix B and $[R_X]_{ij} = C(X_i, X_j)$.

The structure of a cumulant $Q_x(M)$ in ICA model is easily deduced from (9) as:

$$Q_X(M) = A \Delta(M) A^T \tag{9}$$

with:

$$\Delta(M) = diag(kurt(S_1) a_1^T M a_1, ..., kurt(S_n) a_n^T M a_n \tag{10}$$

where a_i is the i-th column of *A*.

Now, let *W* be a whitening matrix and *Z=WX*. Let us assume that the independent sources matrix S has unit variance, so that *S* is white. Thus *Z=WX=WAS* is also white, and the matrix *U=WA* is orthonormal. Similarly, the previous techniques can be applied into (13) for any *n x n* matrix *M*.

First, the whitening matrix *W* and the cumulant matrix *Z* are estimated. Then, the estimation of an orthonormal matrix denoted by *U*, is calculated. Therefore, an estimated matrix *A* denoted by A is obtained from $W^{-1}U$, and the sources matrix S is calculated by $A^{-1}X$.

To measure non-diagonality of a matrix *B*, *off(B)* is defined as the sum of the squares of the non-diagonal elements:

$$off(B) = \sum_{i \neq j} (b_{ij})^2 \tag{11}$$

where b_{ii} are elements of the matrix *B*. In particular $off(U^T Q_Z(M_i) U) = U \Delta_i U^T) = off(\Delta_i) = 0$ since $Q_Z(M_i) = U \Delta_i U^T$ and *U* is orthogonal. For any matrix set *M* and orthonormal matrix *V*, the joint diagonality criterion is defined as:

$$D_M(V) = \sum_{M_i \in M} off(V^T Q_Z(M_i) V) \tag{12}$$

which measures diagonality far from the matrix *V* and bring the cumulants matrixes from the set *M*.

4 Classification System

At this section, the system has got the useful information from the input images by JADE-ICA. Now, the classification component uses this information to take a decision on behalf the spider specie from the spider web biometric as input. To do so, this work uses the well known Support Vector Machine (SVM) [15].

The SVM is a structural risk minimization learning method of separating functions for pattern classification, which was derived from the statistical learning theory elaborated by [16]. In other words, SVM is a tool able to differ between classes characterized by parameters, after a training process.

What makes this tool powerful is the way it handles non-linearly separable problems. In these cases, the SVM transforms the problem into a linearly separable one by projecting samples into a higher dimensional space. This is done using an operator called kernel, which in this study is set to be a Radial Basis Function (RBF). Then, efficient and fast linear techniques can be applied in the transformed space. This technique is usually known as the kernel trick, and was first introduced by [17].

For simplicity, we configure the SVM to work as a verification system. In this particular case, the positive class (1) corresponds to spider specie to verify and the negative class (-1) to the rest of spider species. As a result, the classifier answers the "is the actual spider specie to verify?" question. The output of the SVM is a numeric value between -1 and 1 named score. A threshold has to be set to define a border between actual spider specie (1) and other different spider species (-1) responses.

However, if all samples are used for training, there are no new samples for setting the threshold, and using the training samples for this purpose will lead to bad adjustments. Therefore, a 30 iterations hold-50%-out validation procedure is used over the training samples to obtain scores. These scores are then used to set error rate. The system's margin, defined as the distance of the closest point to the threshold line, is also measured. All these measures are referred to as validation measures.

When the threshold is finally set, the SVM is available to work in test mode. Because no big differences exist in the number of training samples used for this final training and the validation, we can expect the system to have a very similar threshold than that computed before.

In particular, the Least Squares Support Vector Machines (LS-SVM) implementation is used [18]. Given a training set of N data points $\{y_i, x_i\}_{K=1}^{N}$, where x_i is the k-th input sample and y_i its corresponding produced output, we can assume that:

$$\begin{cases} w^T \phi(x_i) + b \geq 1 & if \quad y_i = +1 \\ w^T \phi(x_i) + b \leq 1 & if \quad y_i = -1 \end{cases} \tag{13}$$

where ϕ is the kernel function that maps samples into the higher dimensional space. The LS-SVM solves the classification problem:

$$\min L_2(w,b,e) = \frac{\mu}{2} w^T w + \frac{\zeta}{2} \sum_{i=1}^{N} e_{c,i}^2 \tag{14}$$

where μ and ξ are hyper-parameters related to the amount of regularization versus the sum square error. Moreover, the solution of this problem is subject to the constraints:

$$y_i \left[w^T \phi(x_i) + b \right] = 1 - e_{c,i}, \quad i = 1, ..., N \tag{15}$$

5 Experimental Settings

5.1 Database

The database contains spider web images of four different species named *Allocyclosa*, *Anapisona Simoni*, *Micrathena Duodecimspinosa and Zosis Geniculata.* Each class has 28, 41, 39 and 42 images, respectively, in total, 150 images (see Fig. 3).

Fig. 3. Examples of spider web images from the data base corresponding to *Allocyclosa*, *Anapisona Simoni*, *Micrathena Duodecimspinosa and Zosis Geniculata,* respectively

5.2 Experimental Methodology

Our proposal used the first M components obtained from the JADE-ICA of the spider webs images as inputs for a RBF-kernel LS-SVM with specific regularization and kernel parameters. These two parameters (the number of components and the kernel parameters) were automatically optimized by iteration using validation results. To obtain more reliable results the available samples were divided into training and test sets, so that the system is trained and tested with totally independent samples, according to supervised classification. We have done two experiments, in order to find our best approach.

Our first experiment will be to determine where the most discriminate information is, comparing the whole the spider web or only the central part. We have used 20 components from the pre-processing image to 20 x 20 pixels. This experiment will be done under hold-50%-out validation techniques.

Our second experiment will be to adjust these two parameters (the number of components and the kernel parameters) for the most discriminate information, obtained from previous experiment; under the well known K-Folds cross-validation techniques, which have been used to obtain the final results. In particular, experiments with K equal 3, 5, 7, and 10 were run. Too, multiple hold-out validation has been used in this experiment. In Table 2, accuracy rates can be observed.

All the experiments have been repeated 30 times, showing our accuracy rates in averages and standard deviation.

5.3 Results and Discussions

From our first experiment (see Table 1), we can see the most of the information are located on the central part of the spider web, with over 5% of accuracy rate. This goal agrees with the Biological point of view, because the Biologists use this central part in order to try to do a manual identification. Besides, this goal will design the following experiment, because the tests with whole spider web will be removed.

Table 1. Accuracy Rates in processing images of 20 x 20 pixels for 20 independent components using LS-SVM under 50% hold out cross-validation techniques

Type of image	Whole spider web	Central part of spider web
Accuracy Rate	91.85% ± 4.92	97.16% ± 2.47

The second experiment has been done for the central part of spider web. The Fig. 4 shows the evolution of the number of independent components for 50% hold-out validation techniques. This evolution is done from 1 to 50 independent components, with a step of 5. It can be seen in Fig. 4 that the best accuracy rate versus the number of independent components is 10. Therefore, the rest of experiment will be done with that condition, and the kernel parameter of LS-SVM will be automatically searched (see Section 4) when the mayor accuracy rate is found (see Table 2).

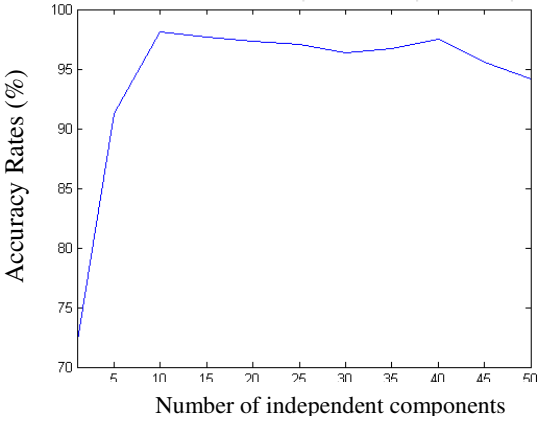

Fig. 4. Evolution of the number of independent components vs. Accurate Rates

Table 2. Accuracy Rates for different validation methodologies

K-Fold	Accuracy Rate	Hold-out	Accuracy Rate
10	96.16% ± 0.89	50	98.15% ± 3.96
7	95.17% ± 1.03	40	98.06% ± 4.09
5	94.60% ± 1.52	30	96.59% ± 5.22
3	94.18% ± 2.06	20	94.40% ± 6.74

From Table 2, we can observe our approach is working over 98%, in order to identify the spider specie using the central part of the spider web as biometric information. Therefore, the error between species is low, when that central part of the spider web is used.

About the computational time, our approach uses 435 milliseconds in order to do the pre-processing and the extraction of 10 independent components using JADE-ICA. Besides, it uses 0.25 milliseconds in order to test one sample. Therefore, our algorithm runs need 435.25 milliseconds in order to evaluate on test sample. This was implemented using AMD Phenom^{TM} II X6 1090T 3.2GHz Processor with 6GB RAM memory, programmed on MATLAB.

6 Conclusions

An automatic identification approach is implemented in this work for the spider classification from its spider web, reaching an accuracy rate of 98.15%. Our computational time is minor to 500 milliseconds, given good efficiency between accuracy rate and this time.

The future lines will be to increase the dataset and to search a new parameterization system in order to improve our present system. These advances give very important biological information, because it validates the Biologists' work, showing the objective way, as it is possible to have an automatic system for identifying spider species, as it is done by biologists.

Acknowledgement. This work has been supported by Spanish Government, in particular by "*Agencia Española de Cooperación Internacional para el Desarrollo*" under funds from D/027406/09, D/033858/10 and A1/039089/11 for 2012.

References

1. Sytnik, K.M.: Preservation of biological diversity: Top-priority tasks of society and state. Ukrainian Journal of Physical Optics 11(suppl. 1), 2–10 (2010)
2. Carvalho, J.C., Cardoso, P., Crespo, L.C., Henriques, S., Carvalho, R., Gomes, P.: Biogeographic patterns of spiders in coastal dunes along a gradient of mediterraneity. Biodiversity and Conservation, 1–22 (2011)
3. Johnston, J.M.: The contribution of microarthropods to aboveground food webs: A review and model of belowground transfer in a coniferous forest. American Midland Naturalist 143, 226–238 (2000)
4. Peterson, A.T., Osborne, D.R., Taylor, D.H.: Tree trunk arthropod faunas as food resources for birds. Ohio Journal of Science 89(1), 23–25 (1989)
5. Cardoso, P., Arnedo, M.A., Triantis, K.A., Borges, P.A.V.: Drivers of diversity in Macaronesian spiders and the role of species extinctions. J. Biogeogr. 37, 1034–1046 (2010)
6. Finch, O.D., Blick, T., Schuldt, A.: Macroecological patterns of spider species richness across Europe. Biodivers. Conserv. 17, 2849–2868 (2008)
7. Eberhard, W.G.: Behavioral Characters for the Higher Classification of Orb-Weaving Spiders. Evolution, Society for the Study of Evolution 36(5), 1067–1095 (1982)

8. Eberhard, W.G.: Early Stages of Orb Construction by Philoponella Vicina, Leucauge Mariana, and Nephila Clavipes (Araneae, Uloboridae and Tetragnathidae), and Their Phylogenetic Implications. Journal of Arachnology, American Arachnological Society 18(2), 205–234 (1990)
9. Eberhard, W.G.: Computer Simulation of Orb-Web Construction. J. American Zoologist, 229–238 (1969)
10. Suresh, P.B., Zschokke, S.: A computerised method to observe spider web building behaviour in a semi-natural light environment. In: 19th European Colloquium of Arachnology, Denmark (2000)
11. Ticay-Rivas, J.R., del Pozo-Baños, M., Eberhard, W.G., Alonso, J.B., Travieso, C.M.: Spider Recognition by Biometric Web Analysis. In: Ferrández, J.M., Álvarez Sánchez, J.R., de la Paz, F., Toledo, F.J. (eds.) IWINAC 2011, Part II. LNCS, vol. 6687, pp. 409–417. Springer, Heidelberg (2011)
12. Otsu, N.: A thresholding selection method from gray-level histogram. IEEE Transactions on Systems, Man, and Cybernetics 9(1), 62–66 (1979)
13. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing. Pearson Prentice Hall (2003)
14. Hyvärinen, A.: Independent Component Analysis: Algorithms and Applications. Neural Networks 13(4-5), 411–430 (2000)
15. Schölkopf, B., Smola, A.J.: Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press (2002)
16. Vapnik, V.: The Nature of Statistical learning Theory. Springer, New York (1995)
17. Yan, F., Qiang, Y., Ruixiang, S., Dequan, L., Rong, Z., Ling, C.X., Wen, G.: Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. Bioinformatics 20(12), 1948–1954 (2004)
18. Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)