Semi-automatic Measure and Identification of Allergenic Airborne Pollen

Antonio García-Manso, Carlos J. García-Orellana, Rafael Tormo-Molina, Ramón Gallardo-Caballero, M. Macías-Macías, Horacio M. Gonzalez-Velasco

Semi-automatic measure and identification of allergenic airborne pollen

A. García-Manso[1], C.J. García-Orellana[1], Rafael Tormo-Molina[2], R. Gallardo-Caballero[1], M. Macías-Macías[1], and H.M. González-Velasco[1]

[1] Pattern Classification and Image Analysis Group, University of Extremadura, Avda. de la Universidad S/N, 10003 Cáceres, Extremadura (SPAIN)
[2] Department of Vegetal Biology, Ecology and Earth Sciences, University of Extremadura, Avda Elvas S/N, 06006 Badajoz (SPAIN)
Corresponding author: antonio@capi.unex.es

Abstract. Current lifestyle in developed countries makes the practice of outdoor activities to be almost mandatory. But, since these practices such as trekking, biking, horseback, or simply running or walking in urban parks, are made in nature (at least outdoors) not everyone can practice them in optimal physical conditions at any time of the year. We are referring to those who suffer from pollinosis or "*hay fever*".
This work present the first stages in the development of a semi-automatic system for counting and identifying airborne pollen, using artificial intelligence techniques for recognizing four of the most representative allergenic pollen types. The system consists of a first stage for the location of pollen grains in the slides, and a second whose goal is the identification using Independent Component Analysis (ICA) and neural nets or SVM. The overall success results achieved with our system are about 88%, averaging for all classes.
Keywords: Independent Component Analysis (ICA), airborne pollen, pollen allergy, Neural Networks, SVM.

1 Introduction

Modern cities have good urban parks or nearby natural environments where its citizens can practice outdoor activities. But, both in these urban parks and in the nearby natural environments, there may be plants producing allergenic pollen for many people. These plants can be natives or even exotic plants used for ornamental purposes, because landscape artists tend to use them in new urban parks. This can cause that many people may not practice outdoor activities, at least not in full physical conditions at any time of year. We are referring to those who suffer from pollinosis or "*hay fever*", that is, presenting allergy to certain types of airborne pollen at certain periods of the year in concrete geographical locations. In [1] it can be found a review of the main types of allergenic pollen and the rates of pollen allergy in Europe. There, authors state that the prevalence of pollen allergy is presently estimated to be up to 40%.
Count and classification of airborne pollen is a very laborious task that require a

lot of time and has to be made by skilled professionals. It is necessary a high level of training to obtain accurate classification results. The study of a preparation (Fig. 1) normally require the identification of a huge number of pollen grains. These analysis can take 2 hours or more, depending on pollen concentration in the sample. Another problem to consider is that the pollen identification can involve some error, because this task is subject to personal perceptions. Normally, the experts work with 400x optical microscopes and when they locate a pollen grain, they need to change the focal plane many times. The pollen grain is a

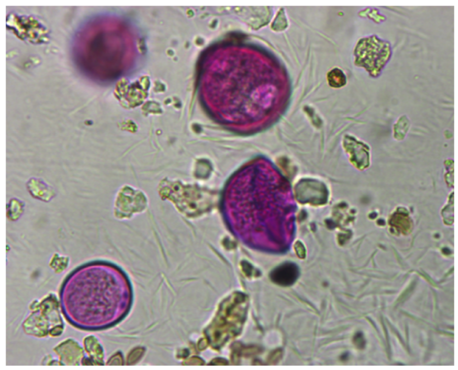


Fig. 1. The slices used to collect the pollen are subjected to a dyeing process. In that way, the pollen grains appear colored over the rest of the surface and others deposed particles in it. Here a frame extracted from a video sequence recorded with a digital camera through the microscope is shown. One can see in the picture the effect of staining the pollen grains. Since, four grains appear fully differentiated from other particles deposited on the adhesive.

tri-dimensional particle which have an aerodynamic size of $15 - 40\ \mu m$. Therefore, its appearance in the microscope image changes dramatically depending on the chosen focal plane. Furthermore, the pollen grains can present some degree of translucence, depending on the type of treatment they have undergone. And, in this case, it is possible to see its apparent inner structure. In Fig. 1 (image extracted from a video sequence) there are four pollen grains of different sizes and, therefore, differently focused. It can be appreciated that the one located at the upper left corner is fully unfocused, whereas the other three can be seen clearly. In subsequent sequences of the video, this pollen grain will appear well focused and completely defined while the other three will be unfocused. This makes pollen identification a very complicated and challenging task, where it is necessary to consider a number of features such as the grain shape, the polarity, the number of openings and their arrangement and shape, texture features,..., and finally, this is translated to a number of perceptions that each expert can have [2].

In recent decades many works have been proposed aimed at locating and iden-

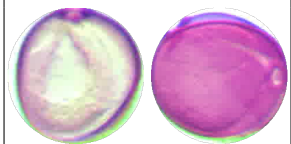
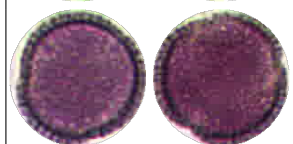
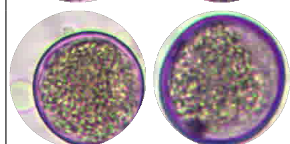
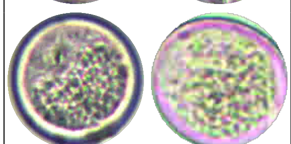
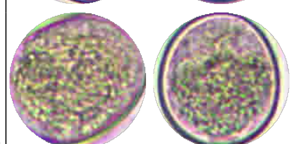
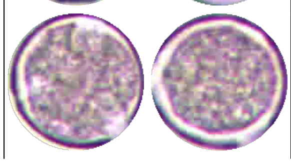
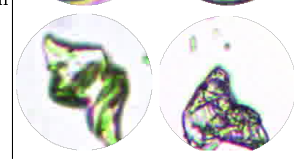
tifying pollen using computer vision techniques. In [3], the authors assess the potential of IR spectroscopy of the Fourier transform as an alternative method for fast and realizable identification of some of the pollen grains types that cause a higher rate of allergy. In [4], a prototype of a full self-station for counting and identifying of pollen on standard glass slides is presented. In [5] a system for automatic detection and classification of pollen on standard slides is also proposed. And, as in the previous case, it also uses parameters and texture analysis to obtain the feature vectors. As can be seen, several approaches have already been proposed to get an automatic or semi-automatic counting station and classification of airborne pollen. It is relatively easy to locate the pollen grains on the slide once the sample has been prepared, that is, it has been subjected to a dyeing process so that the pollen grains are colored (Fig. 1).
Currently, we are developing a self-station where a video camera linked to a microscope is controlled by a computer to get the sample images. The microscope digital camera set can be moved in the direction of Z axis to fit the focus, while the sample to be analyzed is placed on a standard microscope slide and it can be moved in the XY plane, so that the system can move along the entire surface of the slide capturing conventional light microscope video images. Then, each one of the captured videos is analyzed to select the frames with better definition, which will provide the main information. The selected frames are subjected to a process of feature extraction using Independent Component Analysis (ICA), resulting in a feature vector that it will be the input to a classifier. ICA has been widely used in the last decade as a feature extraction technique for its ability to get a base of functions adapted to every problem [6], especially for natural images [6,7]. Different approaches for the analysis of frames are given in [4] and [5], where dark field images (photograph of the silhouettes) are captured and then analyzed together with shape and texture features.

2 Dataset

Our aim is to develop a self-station for counting and identifying airborne potentially allergenic pollen. For this reason, we have opted for capturing prototypes from pollen images taken under the same conditions in which palynologists, responsible for this task, carry it out, and not to take the prototypes from a standard database as the European Pollen Database [8]. The samples were collected by an aerobiological sensor (7 days recording volumetric spore trap), Hirst type, Burkard 7-day. Once they were treated, we used those samples to record the videos. And from these videos, the pollen prototype images used in this work (Table 1) were captured by means of an algorithmic method. As will be explained in detail in the next section, the algorithm, looks for the frame with higher contrast (larger number of edges), for each pollen grain present in the sample. That will be equivalent to having the better definition, showing a greater amount of characteristic details of each particular pollen grain.

It can be seen in Table 1 how there are different classes that seem very similar. This could be because, in those cases, the pollen grains came from the same fam-

Table 1. Examples of pollen of each class and of unwanted particles (trash).

class	samples	class	samples
Avena sativa		Avena sterilis	
Calocedrus decurrens		Olea europaea	
Cypress		Phalaris minor	
Dactylis glomerata		Lolium rigidum	
Quercus rotundifolia		trash	

ily: *Lolium rigidum*, *Dactylis glomerata*, *Phalarais minor* belong to Gramineae family. However, this is not true for *Avena sativa* and *Avena sterilis* which also belong to the Gramineae family. Besides, *Calocedrus decurrens* and *Cypress* belong to Cupressaceae family and in the images appear quite different. For this reason, it is very difficult to discriminate among pollen of the same family, even for skilled persons. Therefore, in some experiments the pollen belonging to the same family is treated as the same class.

3 Methods

Two processes have to be differentiated: first the process followed to segment the pollen grain images in the video sequences, and second the process followed to classify these images.

3.1 Segmentation

Videos from standard glass microscope slides containing the grains of pollen were recorded. These videos were recorded by moving the microscope (that is, changing the focal plane) in the direction perpendicular to the slide (Axis Z) for about 100 μm, depending on the pollen sizes and the slides conditions. Furthermore, these conditions determined the speed at which the microscope was moved, because, at this stage of the project development, the movement was carried out by hand. In this way, each video has about 120 to 180 frames. Afterwards, these videos were analyzed to segment the pollen grains contained in them, with the ultimate goal of finding the best image for each pollen grain that appears in each video. Due to the change in the focal plane, it is possible that, when a video has more than one pollen grain, the best frame for segmenting each grain does not match, as can be seen in Fig. 1.

The process is shown in Fig. 2. First, the algorithm calculates the co-occurrence matrix for each frame in each video. The contrast of each frame can be measured from this co-occurrence matrix and, thus, it is possible to select the frame with higher contrast. This frame is used to segment the possible grains of pollen contained in the image. And after that, once the position of each grain of pollen is known, the algorithm analyzes again the video, but now looking each time in the position where the grains of pollen are.

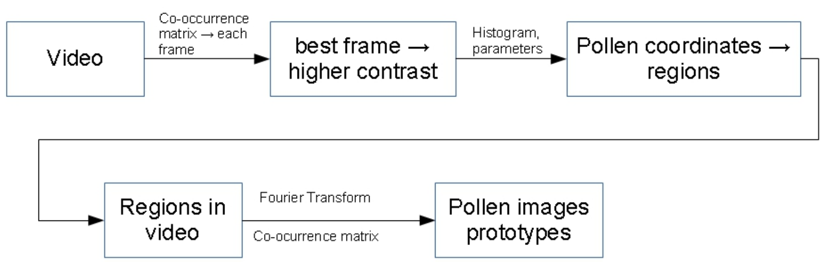


Fig. 2. Outline of segmentation algorithm of pollen grains.

To start the process of segmentation of pollen grains, the frame (color image) with higher contrast is converted to grayscale. And, the pollen grains are segmented by using the histogram and taking into account a series of parameters such as shape factor (pollen grains have a more or less rounded shape), size for each pollen class and others based on the gray value features in each region as deviation and entropy (pollen grains have a semi-translucent inner structure (cytoplasm) and it is not homogeneous in its gray levels).

To find the best frame for each pollen grain the algorithm uses two different processes. One of them consists in using the co-occurrence matrix, equal than to look for the frame with higher contrast, but now, focusing at the regions where the pollen grains are. And the other one consists in using the Fourier transform to find the frame with higher contrast focusing also at the same regions that the previous one. In that way, the algorithm obtains two different sets of images for grains pollen prototypes. Obviously, artifacts (unwanted particles as dust and detritus) are taken as well in this process. These artifacts will be used as image prototypes for background and unwanted particles.

3.2 Classification Algorithms

For training our system, a classical scheme to train classifiers was followed, which consists in to use three independent sets of patterns: one for training, one for avoid the overtraining and another for testing [9]. For generating each of these sets was applied ICA to the pollen images prototypes. We have used Neural Networks and Support Vector Machine (SVM) [10] classifiers.

Independent Component Analysis, (ICA). ICA is a statistical generative model whose objective is to explain the original data ($\mathbf{X}$) using statistically independent random vectors ($\mathbf{S}$). $\mathbf{X}$ can be modeled as $\mathbf{X} \approx \mathbf{AS}$, where $\mathbf{S}$ is the matrix of latent independent components and $\mathbf{A}$ is the mixing matrix.
This technique can be used for feature extraction since the components of $\mathbf{X}$ can be regarded as characteristics representing the objects (patterns) [6]. FastICA algorithm [11] was used to build ICA base functions of the pollen image space. Once a suitable ICA model has been created, it can be expressed, as also with other transformations (wavelets or Gabor filters [12]), each sample of an image (I) located in (x, y), that is, the pixel gray-scale values (point luminances) in an image, as a linear superposition of some base functions $A_j(x, y)$ (rows of the mixing matrix $\mathbf{A}$) called features,

$$I(x, y) = \sum_{j=1}^{q} A_j(x, y) s_j \tag{1}$$

where the s_j are image-dependent coefficients. While the features $\mathbf{A}_j$ are the same for all patches, all prototypes used to generate the ICA base functions. That is, by estimating an image basis using an ICA algorithm, it can be obtained a basis adapted to the data that can model the image space.

The optimal ICA model order is estimated based on the performance of a classifier. Where, the input vectors to this classifier are generated in the following way: first, two different input window sizes, 32×32 and 64×64, are considered; hence each sample used to build the ICA model was resized accord to the size of the model. The resizing was done using the bilinear interpolation algorithm provided by the OpenCV library [13]. For each input size were built 11 ICA models providing a given number of features in the range of 10 to 60.

Neural Networks, (NN). The neural classifier was built with a feed-forward multilayer perceptron which has a single hidden layer. The neural network weights were adjusted with a variant of the classical Back-Propagation (BP) algorithm named Resilient Back-Propagation (Rprop) [14]. Rprop is a local adaptive learning scheme performing supervised batch-learning in a multilayer perceptron with faster convergence than the standard BP algorithm.

In this work, the Stuttgart Neural Network Simulator environment SDK [15] was used to generate and train the neural network classifiers. To avoid local minimum during the training process, each setting was repeated four times, changing the initial weights in the net at random. Furthermore, the number of neurons in the hidden layer was allowed to vary between 50 and 650 in steps of 50 selecting the network that provides the highest success rate over the test subset.

Support Vector Machines, (SVM). LibSVM [16] library was used with the radial basis function kernel showed in eq. 2. To find the optimal configuration of the classifier, the best values to use for parameters γ and C were studied, varying them between 0.125 and 4 in a nested loop which doubles the value of the corresponding parameter in each round.

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2, \gamma > 0) \qquad (2)$$

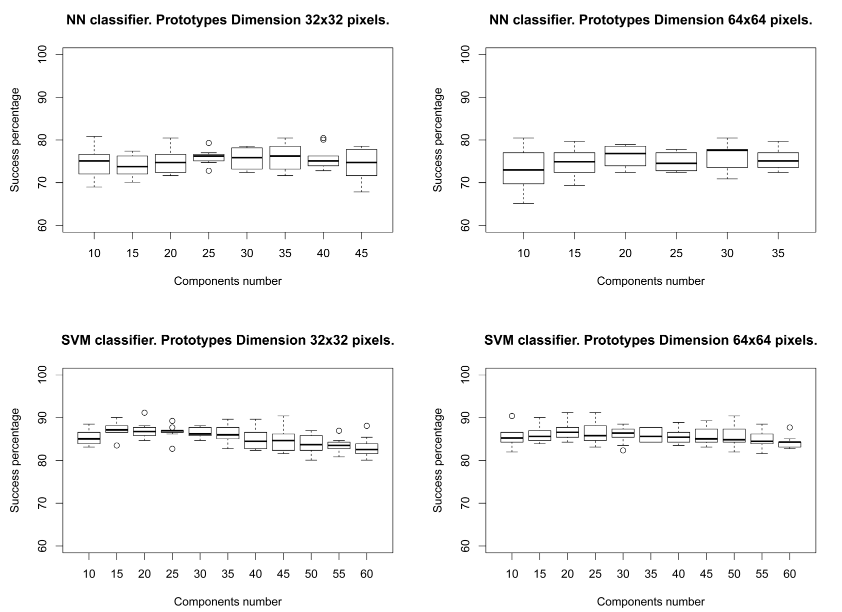
4 Results

Due to the nature of the dealt problem was decided to do two experiments. In the first (Experiment 1) all classes shown in Table 1 are considered independently each from others. That is, in Experiment 1, ten classes are considered. While in the second (Experiment 2), *Dactylis glomerata*, *Phalaris minor* and *Lolium rigidum* are grouped in a same class named *Gramineae*; *Avena sativa* and *Avena sterelis* are also grouped in a same class named *Avena*, although, they also belong to the *Gramineae* family. And finally, *Calocedrus decurrens* and *Cypress* are grouped in a same class named *cypress*, although images of those classes appear somewhat different. In that way, only six classes are considered in the Experiment 2.

A double training process was done, on the one hand NN classifiers were trained and, on the other hand, SVM classifiers. After training process were obtained the results shown in Tables 2 and 3, where can be seen the results obtained over the test subsets. To perform the test, the same sets prototypes in a *10-fold cross validation* configuration for the two types of classifiers were used. Therefore, the results are shown using box plots. The total number of prototypes was 1660 with a average distribution in the 10-fold as follow: 167.9 for *trash*, 134.4 for *Cypress*, 174.7 for *Calocedrus decurrens*, 175.1 for *Lolium rigidum*, 76.5 for *Dactylis glomerata*, 141.2 for *Phalaris minor*, 121 for *Avena sativa*, 105.6 for *Avena sterelis*, 298.1 for *Quercus rotundifolia* and 264.9 for *Olea europeae*.

First of all, it has to be said that the training times for the neural nets classifiers were longer than for the SVM classifiers (in average, about 25 minutes for SVM against more than 48 hours for NN). For this reason, after to do a significant

Table 2. Classification results depending on components number obtained for the **Experiment 1**.



number of trainings with NN and to test that the results were worse than with SVM, it was decided not to do more trainings.
In tables 2 and 3 one can see the performance obtained by the neural and SVM classifiers depending on the number of features in the inputs vectors. In that way, the results obtained, choosing the median as a standard to select the best value, are shown in Table 4.
As can be seen in these tables the performance of the SVM classifiers is better than that of the neural classifiers, about 10 percentage points in all cases. This along with the training time do that, for this problem, the election of the SVM classifiers is much more convenient than NN classifiers. Another points to consider are, on the one hand, that the success results do not seem to depend on the prototypes dimension. At least for the two sizes that were considered. Which indicates that there is information enough in the prototypes of less size. And, in the other hand, the grouping of the classes in Experiment 2 does not seem to improve very much the average success results.

5 Conclusions

The aim of this study was to test if by mean of an algorithm we were able to count and classify the allergenic airborne pollen in a semi-automatic way. Since,

Table 3. Classification results depending on components number obtained for the **Experiment 2**.

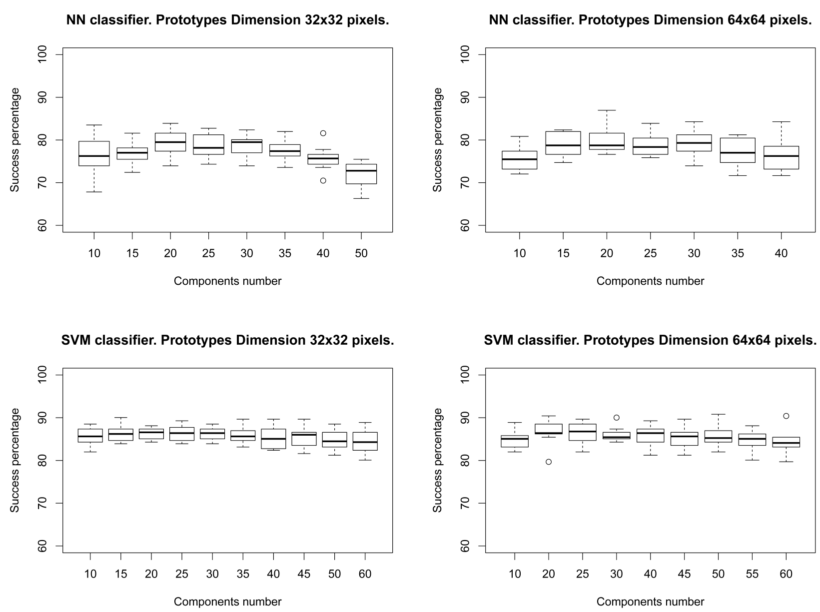


Table 4. Summary of Tables 2 and 3. Maximum value of the median in each case.

	SVM		**NN**	
Prototypes dim.	Experiment 1	Experiment 2	Experiment 1	Experiment 2
32	87.16→15 comp	86.59→20 comp	76.24→25 comp	79.50→20 comp
64	86.59→20 comp	86.78→20 comp	77.58→30 comp	79.31→30 comp

we need the samples already prepared. And from the obtained results, we could say that we have developed an algorithm able to segment well the pollen in the image samples. And the classifier algorithm using SVM carries out about a 90% success. Which, taken into account nature of the problem, is a very promising result. As future work, we have to improve the algorithm to segment the pollen images doing it more adaptable to the samples conditions. And, with the station built, we have also to capture each time more image prototypes of all classes to do more robust the classifier.

Acknowledgments

Presented work in this paper has been funded by *"Research program at the Uni. of Extremadura, 2013. Action VII: P. Int. to Research and Technological Development"* and "Junta de Extremadura" through GR10018, partially by ERDF.

References

1. D'Amato, G., Cecchi, L., Bonini, S., Nunes, C., Annesi-Maesano, I., Behrendt, H., Liccardi, G., Popov, T., Cauwenberge, P.V.: Allergenic pollen and pollen allergy in Europe. Allergy **62** (2007) 976–990
2. S-otero, M.P.D., Gonzlez, A., Rodrguez-Damin, M., Cernadas, E.: Computer-aided identification of allergenic species of Urticaceae pollen. Grana **43**(4) (2004) 224–230
3. Gottardini, E., Rossi, S., Cristofolini, F., Benedetti, L.: Use of Fourier transform infrared (FT-IR) spectroscopy as a tool for pollen identification. Aerobiologia **23**(3) (sep 2007) 211–219
4. Holt, K., Allen, G., Hodgson, R., Marsland, S., Flenley, J.: Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. Review of Palaeobotany and Palynology **167**(34) (2011) 175–183
5. Rodríguez-Damin, M., Cernadas, E., Formella, A., Fernndez-Delgado, M., S-Otero, P.D.: Automatic detection and classification of grains of pollen based on shape and texture. IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews **36**(4) (July 2006) 531–542
6. Hyvärinen, A., Hurri, J., Hoyer, P.O.: Natural Image Statistics. A Probablisctic Approach to Early Computational Vision. Springer-Verlag London Limited (2009)
7. Anjali, P., Ajay, S.: A review on natural image denoising using independent component analysis (ICA) technique. Advances in Computational Research **2**(1) (2010) 06–14
8. Fyfe, R.M., de Beaulieu, J.L., Binney, H., Bradshaw, R.H.W., Brewer, S., Flao, A.L., Finsinger, W., Gaillard, M.J., Giesecke, T., Gil-Romera, G., Grimm, E.C., Huntley, B., Kunes, P., Kuhl, N., Leydet, M., Lotter, A.F., Tarasov, P.E., Tonkov, S.: The European Pollen Database: past efforts and current activities. Vegetation History and Archaeobotany **18**(5) (September 2009) 417–424
9. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
10. Vapnik, V.N.: The Nature of Statistical Learning Theory. Second edn. Statistics for Engineering and Information Science. Springer (2000)
11. Ripley, B.: FastICA Algorithms to perform ICA and Projection Pursuit. http://cran.r-project.org/web/packages/fastICA/fastICA.pdf (February 2009)
12. Gonzalez, R.C., Woods, R.E.: Digital image processing. Third edn. Prentice-Hall, Upper Saddle River, NJ (2008)
13. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. first edn. O'Reilly Media (2008)
14. Riedmiller, H., Braun, H.: A direct adaptive method for faster backpropagation learning. The RPROP algorithm. IEEE International Conference on Neural Networks **1** (1993) 586–591
15. Zell, A., Mache, N., Huebner, R., Mamier, G., Vogt, M., Schmalzl, M., Herrmann, K.: SNNS (Stuttgart Neural Network Simulator). Neural Network Simulation Environments **254** (1994) 165–186
16. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3) (2011) 27:1–27:27