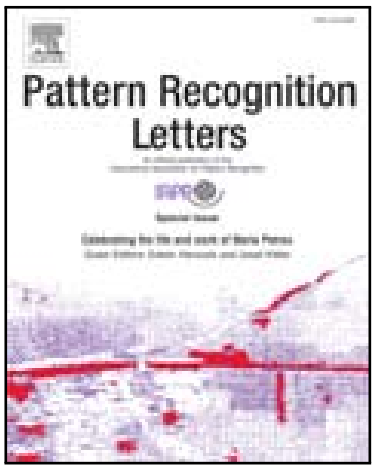
Categorizing plant images at the variety level: Did you say fine-grained?

Julien Champ, Titouan Lorieul, Pierre Bonnet, Najate Maghnaoui, Christophe Sereno, Thierry Dessup, Jean-Michel Boursiquot, Laurent Audeguin, Thierry Lacombe, Alexis Joly



Categorizing plant images at the variety level: did you say fine-grained ?

Julien Champ, Titouan Lorieul, Pierre Bonnet, Alexis Joly

PII: S0167-8655(16)30106-4
DOI: 10.1016/j.patrec.2016.05.022
Reference: PATREC 6546

To appear in: *Pattern Recognition Letters*

Received date: 24 March 2015
Accepted date: 23 May 2016

Please cite this article as: Julien Champ, Titouan Lorieul, Pierre Bonnet, Alexis Joly, Categorizing plant images at the variety level: did you say fine-grained ?, *Pattern Recognition Letters* (2016), doi: 10.1016/j.patrec.2016.05.022

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- two new datasets were created for the evaluation of plant varieties recognition
- an experimental study was conducted with today's best performing techniques
- recognizing rice seeds variety appears to be feasible in controlled environment
- recognizing grape varieties from their leaves is still an open problem
- results show that convolutional neural networks perform the best on such problems

Pattern Recognition Letters
journal homepage: www.elsevier.com

Categorizing plant images at the variety level: did you say fine-grained ?

Julien Champ[a], Titouan Lorieul[b], Pierre Bonnet[c], Alexis Joly[b,**]

[a]*Inra, LIRMM, Montpellier, France*
[b]*Inria, LIRMM, Montpellier, France*
[c]*Cirad, AMAP, Montpellier, France*

ABSTRACT

This paper addresses the problem of categorizing plant images at the variety level, i.e. at a finer taxonomic grain than state-of-the-art studies usually working at the species level. It therefore introduces two new evaluation datasets of agro-biodiversity interest, each being related to concrete scenarios on large-scale plant resources. They have been chosen so as to involve very different acquisition protocols and visual patterns in order to evaluate if state-of-the-art image classification techniques can generalize to such specific contexts and avoid the cost of building specific ad-hoc solutions. The first one is a collection of 2 071 pictures of loose rice seeds built from 95 accessions kept in a bank of seeds. The second one is a collection of 2 037 pictures of grape leaves taken in the fields and belonging to 34 varieties among the most commonly ones used in viticulture. Both datasets exhibit a very low inter-class variability resulting in two challenging fine-grained classification tasks, even for expert human operators. A baseline experimental study was conducted on the two datasets using the two most effective families of classification techniques in the state-of-the-art, i.e. convolutional neural networks on one side and fisher vectors-based discriminant models on the other side. It shows that the achieved classification performance is very different between the two problems. It is actually pretty bad for the grape leaves collection but much better in the case of the rice seeds collection for which the acquisition protocol was much more constrained and the morphological variability more visible. The conclusion is that automatically identifying plant varieties might already be feasible for some specific scenarios and in controlled environments but that it is still an open problem in the general case.

1. Introduction

Sustainable development of agriculture as well as biodiversity conservation are strongly related to our knowledge of the identity, geographic distribution and uses of plants. Unfortunately, such basic information is often only partially available for professional stakeholders, teachers, scientists and citizens, and often incomplete for ecosystems that possess the highest plant diversity (such as Mediterranean and tropical regions). One of the big challenge, expressed as the taxonomic gap, is that identifying plants is usually impossible for the general public, and also often a difficult task for professionals such as farmers or foresters and even for the botanists themselves. Using image-based identification and collaborative data management tools is considered as one of the most promising solution to help bridging the taxonomic gap (Joly et al. (2014b); Caputo et al. (2013); Cai et al. (2007a); Joly et al. (2014a); Spampinato et al. (2010); Kumar et al. (2012)). For centuries, plants have actually been classified according to their morphology, and then in many cases on their visual appearance. With the recent advances in digital devices/equipment, automatizing such classification process thanks to computer vision techniques therefore appears as the most straightforward solution.

The study of the visual appearance of plants is however much less advanced when speaking about agricultural varieties and genetic resources. Indeed, plants breeding has been practiced since thousands of years by farmers, nurserymen and others plant producers. Theses centuries of selection has allowed to develop specific plant characteristics (from wilds plants populations) for specific needs, such as better plant

**Corresponding author: Tel.: +0033-467-149-772
e-mail: alexis.joly@inria.fr (Alexis Joly)

production, pest resistance, or water use efficiency. Due to this work, humanity is now able to store, access and exchange thousands of crops varieties or horticultural plants. Describing and analysing the morphological, physiological or ecological variations of the closely related varieties belonging to a single original species is however much more difficult than studying classical inter-species variations. This problem is increasingly studied with the recent advances in plant genomic and the availability of related data but it is usually not studied from the perspective of image data and automatic visual analysis tools. The emergence of automatized varieties identification tools based on visual contents might be useful in many different use cases such as (i) living or dry plant collections management (for field conservatory, nursery plant production), (ii) control of plant material transfer, (iii) field prospecting (with the aim to well characterize the terroir of a plant production), etc.

However, categorizing plant images at the variety level is a problem that is much harder than the more classically studied problem of identifying plants at the species level. Working at such a finer taxonomic grain is actually much more difficult because of the very low inter-class variability and the fact that the visual patterns to be used for discriminating the varieties can be very specific and very different from a species to another one. The real grand challenge is thus not to design ad-hoc computer vision techniques for each use case but to evaluate in what way generic image classification techniques can deal with very different contexts and acquisition protocols set up by the biologists themselves. In this paper, we therefore introduce two image collections of agro-biodiversity interest with the aim to evaluate the feasability of automated plant varieties recognition from their visual appearance. They have been chosen so as to involve very different acquisition protocols and visual patterns in order to evaluate if state-of-the-art image classification techniques can generalize to such specific contexts and avoid the cost of building specific ad-hoc solutions. The acquisition protocol of each collection was carefully designed jointly with biologists and stakeholders so as to target innovative usage scenarios within existing workflows. The produced image datasets exhibit a very low inter-class variability resulting in two challenging fine-grained classification tasks, even for expert human operators (see section 3 and 4). As a second contribution, we then present the results of an experimental study that was conducted on the two datasets using the two effective families of classification techniques for solving fine-grained image classification problems, i.e. convolutional neural networks on one side and fisher vectors-based discriminant models on the other side (see section 6).

More generally, the scientific contribution of this paper lies in two main points. First of all, the production of specific visual training data and knowledge is now becoming one the most central problem in computer vision. Image classification has actually now been proved to be solved on large generalist corpora, but there is still a gap before being able to recognize the spectrum of millions or even billions of entities lying in the long tail of data occurrences (i.e classes with very few or none training samples available). Experimental studies such as the one conducted in this paper provides some essential findings on the genericity and transfer learning abilities of state-of-the-art computer vision techniques in this regard. Secondly, the paper proves for the first time that automatically identifying plant varieties from their visual appearance is feasible for some groups and usage scenarios and this opens the door to further investigations in the domain.

2. Related works

Content-based image retrieval and computer vision approaches are considered as one of the most promising solutions to help bridging the taxonomic gap, as discussed in Gaston and O'Neill (2004); Cai et al. (2007b); Trifa et al. (2008); Spampinato et al. (2012); Joly et al. (2014a). We therefore see an increasing interest in this trans-disciplinary challenge in the multimedia community (e.g. in Nilsback and Zisserman (2008); Goëau et al. (2011b); Cerutti et al. (2011); Mouine et al. (2012); Kebapci et al. (2011); Hsu et al. (2011). Some recent studies have been done on closely related species (based on leaves analysis such as bean species Larese et al. (2014), or seed analysis such as in the studies of Acacia species and sub-species Sivakumar et al. (2013), Diplotaxis species Grillo et al. (2012), or pea varieties Smykalova et al. (2011)). Orru et al. (2012) have meanwhile invested on the efficiency evaluation of image analysis methods (for morpho-colorimetric feature extraction) compare to molecular analysis on grape variety seeds in the aim to discriminate taxonomical groups. Nevertheless, most of the analysis actually realised are not able to deal with an important number of varieties, such as in our case. Beyond the raw identification performances achievable by state-of-the-art computer vision algorithms, recent visual search paradigms actually offer much more efficient and interactive ways of browsing large flora than standard field guides or online web catalogs (Ellison et al. (2013)). Smartphone applications relying on such image-based identification services are particularly promising for setting-up massive ecological monitoring systems, involving thousands of contributors at a very low cost.

A first step in this way has been achieved by the US consortium behind LeafSnap[1], an i-phone application allowing the identification of 184 common american plant species based on pictures of cut leaves on an uniform background (see Kumar et al. (2012) for more details). Then, the French consortium supporting Pl@ntNet (Joly et al. (2014a)) went one step beyond by building an interactive image-based plant identification application that is continuously enriched by the members of a social network specialized in botany. Inspired by the principles of citizen sciences and participatory sensing, this project quickly met a large public with more than 300K downloads of the mobile applications (Goëau et al. (2013a, 2014a)). A related initiative is the plant identification evaluation task organized since 2011 in the context of the international evaluation forum CLEF[2] and that is based on the data collected within

[1] http://leafsnap.com/

[2] http://www.clef-initiative.eu/

Pl@ntNet. In 2011, 2012 and 2013 respectively 8, 11 and 12 international research groups participated to this large collaborative evaluation by benchmarking their images-based plant identification systems (see Goëau et al. (2011a, 2012, 2013b) for more details). The data used during these 3 first years can be accessed online[3]. Contrary to previous evaluations reported in the literature, the key objective was to build a realistic task very close to real-world conditions (different users, areas, periods of the year, important species number, etc.). The 2014th and 2015th edition of the task were organized within a newly created lab of CLEF called LifeCLEF[4] and dedicated to the identification of living organisms in general (with an audio-based bird identification task and a video-based fish identification task in addition to the image-base identification task). Details of the participants and the methods used in the runs are synthesised in the overview working notes of the task (Goëau et al. (2014b, 2015)).

From a computer vision and technological perspective, our work is more generally related to *image classification*. Most popular methods for this problem are typically based on the pooling of local visual features into global image representations and the use of powerful classifiers in the resulting high-dimensional embedded space such as linear support vector machines (Lazebnik et al. (2006); Perronnin et al. (2010)). The Bag-of-word representation (BoW) notably remains a key concept although the raw initial scheme of (Sivic and Zisserman (2003)) is now outperformed by several alternative new schemes (Lazebnik et al. (2006); Jiang et al. (2007); Perronnin and Dance (2007); van Gemert et al. (2010); Jégou et al. (2012)). Its principle is to first train a so called visual vocabulary thanks to an unsupervised clustering algorithm computed on a given training set of local features. The produced partition is then used to quantize the visual features of a given new image into *visual words* that are aggregated within a single high-dimensional histogram. Partial geometry can be embedded in the image representation by using the Spatial Pyramid Matching scheme of (Lazebnik et al. (2006)). As it relies on vector quantization, the BoW representation is however affected by quantization errors. Very similar visual features might be split across distinct clusters whereas more dissimilar ones might be affected to the same visual word. This results in both mismatches and potentially irrelevant matches. To alleviate this problem, several improvements have been proposed in the literature. The first one consists in expanding the assignment of a given local feature to its nearest visual words (Jiang et al. (2007); Philbin et al. (2008); van Gemert et al. (2010); Jégou et al. (2012)). This allows reducing the number of mismatches without degrading much the encoding time. Other researchers have investigated alternative ways to avoid the vector quantization step, using sparse coding (Yang et al. (2009)) or locality-constrained linear coding (Wang et al. (2010)). Such methods optimize the affectation of a given local feature to a few number of visual words thanks to sparsity or locality constraints on the global representation. Another alternative is to use aggregation-based models such as the improved Fisher Vector of Perronnin and Dance (2007) or the VLAD encoding scheme (Jégou et al. (2012)). Such methods do not only encode the number of occurrences of each visual word but also encode additional information about the distribution of the descriptors by aggregating the component-wise differences. When used with discriminative linear classifiers, such high-dimensional representations benefit of both generative and discrimination approaches leading to state-of-the-art classification performances on fine-grained classification benchmarks (Gosselin et al. (2014)).

A radically different approach to image classification is the use of *deep convolutional neural networks*. Rather than extracting the features according to hand-tuned or psycho-vision oriented filters, such methods directly work on the image signal. The weights learned by the first convolutional layers allows to automatically build relevant image filters whereas the intermediate layers are in charge of pooling these raw responses into high-level visual patterns. The last fully connected layers work more traditionally as any discriminative classifier on the image representation resulting from the previous layers. Deep convolutional neural networks have been recently proved to achieve better results on large-scale image classification datasets such as ImageNet (Krizhevsky et al. (2012)) and do attract more and more interest in the computer and multimedia vision communities. A known drawback of Deep Convolutional Neural Networks is however that they require a lot of training data mainly because of the huge number of parameters to be learned. Their performances on fine-grained classification are consequently more controversial and they are still often outperformed by local features based approaches, as shown in our experiments. Besides, it is important to notice that they inspire the investigation of new deep learning models making use of more traditional visual features embedding methods (e.g. Simonyan et al. (2013)).

3. Rice seeds dataset

3.1. Context

The French research unit on Genetic Improvement and Amelioration of Mediterranean and Tropical Plants (Agap) manages some major collections of plant genetic resources (lucerne, grapevine, rice, sorghum, cotton, groundnut, etc.), amounting to date to almost 60,000 accessions. These collections are made up of different categories of genetic resources: French national collections and heritage resources, original materials, species related to the cultivated or wild species worked on, resources of a scientific nature. In particular, these collections contain many materials intended for genetics and genomics studies, notably for the model species *Medicago truncatula* Gaertn., along with collections of rice insertion lines. The facilities managing these collections are labelled as biological resource centres through the ISO 9001 certification initiatives and the extension of standard Afnor NF 96900.

[3] http://publish.plantnet-project.org/project/plantclef
[4] www.lifeclef.org

The seed and biological resources laboratory (LSRG), in Montpellier, more specifically manages a large collection of 1717 varieties of *Oryza sativa* L. coming from more than 50 countries world wide. This large collection host nevertheless a small fraction of the 40 000 rices varieties used across the world. Each variety is preserved through one or few accessions, each accession being formed by a set of rice seeds kept in their hulls (from few tens of seeds to several thousands depending on the variety). The seeds are regularly renewed through germination allowing to keep a high average germination rate of about 90%. Before the image acquisition campaign initiated by the work presented in this paper, only few illustrations of the seeds in collection did exist.

3.2. Usage scenario

Automatically identifying the rice seeds varieties thanks to their visual appearance might be useful for several scenarios. First of all, a visual control is already performed by the operators in charge of managing seeds collections. This allows detecting mistakes or inconsistencies in the labelling of the new received accessions (i.e. new in-the-field samples). This process is however very imprecise as it relies on rough visual attributes of the seeds such as *color:brown*, *shape:long* or *hairiness:hairy*. Secondly, the automatic identification of the variety could be useful for entity resolution purposes (taxonomic name resolution). In many field contexts, the variety of a cultivated plant is actually only known through its vernacular name, i.e the local name given by the farmers of that region of the world. Mapping that names on a taxonomic referential is usually a very hard task and prevent from collecting accurate data from that cultures (which represent a potentially huge source of information).

3.3. Image acquisition

The image acquisition set up (illustrated in Figure 1) was a standard reflex camera fixed on a tripod at a fixed distance (DIST=10.4cm) from the petri box containing the rice seeds. The camera was equipped with a Compact Macro Lens (EF 50mm 1:2.5, Sigma), an MR-14 EX flash and a remote trigger. The images used for the training set (see next section for more details) were acquired by a different person, a different day, and with a different camera than the ones used as test images. This allows making the benchmark more realistic and closer to a real-world scenario where a person would have to reproduce the image acquisition protocol. For the selection of the varieties to be included in the dataset, we focused only on the ones having their rough visual attributes informed in the database (i.e. 550 varieties). We then selected 95 varieties through a randomized greedy algorithm guarantying that all visual attributes are represented in the dataset. The table 1 gives a synthesis of the visual and morphological diversity of this dataset, as for example the rounded shape of the variety 16, the brown color of the variety 23, or the presence of long hairs on variety 25.

3.4. Evaluation dataset

The final evaluation dataset to be shared with the scientific community, referred as **CiradRiceSeed** is composed of two parts, a training set and a testing set. The train dataset is compound of 1 506 images produced by the same operator, while the 565 images of the test dataset (based on the same varieties) were produced by an other operator. This was done in order to reproduce a real world scenario, in which an external actor realise new data in the aim to get possible varieties names. The varieties list is provide in table 2.

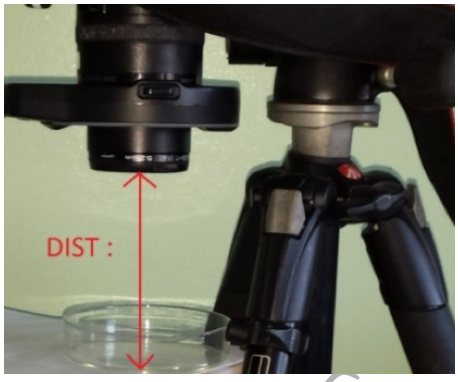


Fig. 1. Image acquisition setup for the CiradRiceSeed dataset

4. Grape Leaves dataset

4.1. Context

The French Vine and Wine Institute (IFV), through the intermediary of the National Plant Material Pole located in Southern France, is a vine selection establishment that coordinates clonal selection and conservation actions carried out throughout the country. It ensures the conservation of a quite unique collection in the world of hundreds of vine varieties at the "*Domaine de l'Espiguette*", and about 4000 clones. A large part of this material is inscribed in the national official catalogue[5]. The main objectives of the IFV is to improve the potential of the wine plant material while maintaining all of the characteristics and identity of each variety. The IFV contributes to the conservation of all vine varieties grown in France by both the sanitary quality of plants, and the consideration of the diversity of varieties and types of wines. This activity allows a strong link between research and production. The IFV selection center is located in the "*Domaine de l'Espiguette*", occupying 80 hectares in the town of Grau du Roi, Languedoc-Roussillon. This site was chosen for the national conservatory clones because its soils consist only of pure Mediterranean sand, containing neither phylloxera or nematode vectors of fanleaf virus. More than 38 hectares are planted with clones collections. Based on DNA results, it is estimated that around 5000 varieties exist worldwide, and many of them are closely related (This et al. (2006)). This important number, reflect the long history of domestication that probably started 8000 BP (before present). Morphological identification of this varieties is mainly done with adult leaves analysis, with some specific normalized leaf descriptors, as described in OIV (2009).

[5] http://plantgrape.plantnet-project.org/

Table 1. Image samples of the CiradRiceSeed dataset for 5 random varieties

	Train			Test
Variety 4				
Variety 16				
Variety 23				
Variety 25				
Variety 84				

4.2. *Usage scenario*

Ampelography is a field of botany dedicated to the identification and classification of grapevines. Traditionally this has been done by comparing the shape, colour and texture of the vine leaves and grape berries. Recently, DNA fingerprinting has played a major role in the development of this discipline, nevertheless it has not permitted to widely extend grapevines identification to non-expert peoples. Automatically identifying the grape varieties thanks to the visual appearance of their leaves might be for this reason useful for several scenarios. First of all, it is important to remember that grape identification may have a very important impact, notably on financial aspect, as most of the wines in the world are based on the cultivars names. Vine variety identification is then practice, (i) by state agents in charge of the reglementation in vigor, (ii) nurserymans that want to confirm names of the commercialized material, (iii) winegrowers who want to identify old plants conserved in unused plots, etc... All theses contexts could get a significant benefit of an automated multimedia identification system.

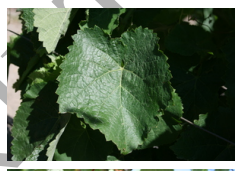
4.3. *Image acquisition*

In the aim to evaluate such solution, we organised in June 2012 the collect of visual data on a selection of grape varieties growing at the Espiguette estate. 11 different people, each of them with different camera, were mobilized for this experimentation. They were organised by binomials, in order to collect by binomial about 60 different images of laves for each variety. Variety were selected according to their economic importance in French viticulture. Leaves were selected at the adult stage, in healthy condition, and the provided instructions were to (i) try avoiding strong lightness contrasts (with a part of the leaf in shadow and and other one in full sun condition) (ii) try to avoid the visual overlap between several leaves. Photographers were free to choose their camera parameters, but to pay attention to put one leaf at the image center, to photograph it in landscape format, to maximize it surface on the picture (in order to ovoid to take picture from a too long distance). Most of the pictures were taken directly on the living plants, but a part of them were took on a cut leaf put on the ground (with pure sand in this case).

4.4. *Evaluation dataset*

The final evaluation dataset to be shared with the scientific community, referred as **IFVGrapeLeaves** is composed of two parts, a training set and a testing set. The table 1 give a synthesis of the visual and morphological diversity of this dataset. It is then possible to see variation (i) between leaves in terms of leaf shape, lobes number, or tooth size on the leaf margin, (ii) but also in terms of background color, light conditions, or image quality. The full dataset is compound of 2037 images of 34 grape varieties. In order to divide it in train an test sets, we kept all the pictures of one photograph for train (generally the photograph who took the most important

Table 4. Image samples of the IFVGrapeLeaves dataset for 5 random varieties

	Train			Test
Variety 4 (Chenin)				
Variety 9 (Gamay)				
Variety 12 (Gros Manseng)				
Variety 19 (Meunier)				
Variety 25 (Roussanne)				

number of images for or specific variety), and we putted all the others pictures of the binomial in test dataset. This avoids for the same variety, to have picture of the same author in both datasets. The varieties list is provide in table 3.

5. Baseline fine-grained image classification systems

We did experiment two families of image classification techniques that are known to provide state-of-the-art classification performances in fine-grained recognition challenges (Gosselin et al. (2014); Joly et al. (2014b)) and that perform the best within the two last editions of the PlantCLEF international evaluation campaign Goëau et al. (2014b, 2015) (whatever the type of view or acquisition protocol e.g. leaf scans, in-the-field leaf pictures, flower pictures, bark pictures, etc.).

5.1. Fisher vectors & Support Vectors Machine

Fisher vectors (FV) were first introduced in image classification by Perronnin and Dance (2007) and proved to be very efficient in the fine-grained classification task later on (Gosselin et al. (2014)). According to recent surveys such as Huang et al. (2014), they are the best performing pooling strategy currently available. We will only recall here the main steps used to extract Fisher vectors, for detailed explanations of the theoretical derivation and for performance analysis we redirect the readers to Sánchez et al. (2013). The pipeline for computing the Fisher vector describing an image consists in:

- Dense extraction of local features: descriptors, usually SIFT descriptors, are extracted on densely sampled overlapping patches on several scales.
- PCA transformation: the descriptors are then decorrelated and compressed using a Principal Component Analysis reducing the dimensionality (usually to 80 for SIFT features).
- Feature space density estimation: the distribution of features is modeled as a Gaussian Mixture Model (GMM) that is learned using the popular Expectation Maximisation algorithm. We thus obtain a probability distribution of the form of $u(x) = \sum_{k=1}^{K} w_k u_k(x)$ where u_k follows a Gaussian distribution, $u_k \sim \mathcal{N}(\mu_k, \Sigma_k)$ with μ_k the mean and Σ_k the covariance matrix which is diagonal because the features are decorrelated, and w_k is the weight of the k-th Gaussian, they satisfy $\sum_k w_k = 1$.
- Encoding: the features are encoded and pooled using

$$\mathcal{G}_{\mu_k} = \frac{1}{\sqrt{w_k}} \sum_{i=1}^{N} \gamma_k(x_i) \frac{x_i - \mu_k}{\sigma_k}$$

$$\mathcal{G}_{\sigma_k} = \frac{1}{\sqrt{w_k}} \sum_{i=1}^{N} \frac{\gamma_k(x_i)}{\sqrt{2}} \left(\left(\frac{x_i - \mu_k}{\sigma_k} \right)^2 - 1 \right)$$

where all the dicitions and squaring are element-wise operation and where $\gamma_k(x) = \frac{w_k u_k(x)}{\sum_{k'=1}^{K} w_{k'} u_{k'}(x)}$. Theses $2K$ vectors

Table 2. Variety list of the CiradRiceSeed dataset

ID	Variety name	ID	Variety name
0	KANIRANGA	48	LAC 3 (1)
1	DA 9	49	P 335
2	DA 5	50	9 AB
3	T 1	51	T 1 (IRR 74)
4	CO 18	52	NHTA 13
5	SURJAMKUHI	53	T 43
6	DHOLA AMAN	54	DJUBUH
7	JC 73-4	55	SAN THOU GEE
8	RTS 5	56	KHAO TONG
9	KUN MIN TSIEH HUNAN	57	MACK FAY DENG
10	CO 25	58	KH CHETANG
11	JHONA 349	59	KHAO XENG
12	BAGUAMON 14	60	LA KHONE DENG
13	MALAGKIT PIRURUTONG	61	9 D
14	RTS 16	62	13 A
15	PADI RAOEKANG	63	257
16	HU LO TAO	64	483
17	JC 91	65	KU 48
18	LAMBAYQUE 1	66	IRAT 8
19	ARC 10497	67	RAM TULASI SEL.
20	KAKANI 2	68	HERIKA (LONG FORM)
21	ACHILLE	69	CALORO
22	63-105	70	SEMO
23	ZAKPALE	71	ZAKPALE 4
24	268 B / PR 22-3-2	72	KESSIRIME B
25	377	73	KEDIALA OUADEO B
26	TAITUNG 328	74	MAROVEL BEIDARI B
27	KOIRAO BALEO	75	PAGAIYAHAN
28	COLUMBIA 2	76	EH IA CHIU
29	SHINHAKABURI (1)	77	DA 23
30	ZAKPALE 3	78	DA 16
31	HAO MET NHAY	79	RATHUWEE
32	MACK HING HOM	80	JC 148
33	CHIANAN 8 MA-3	81	JC 178
34	SOMCAU 70 A	82	NAM SA GUI 19
35	9 A	83	MOROBEREKAN
36	P 269	84	ARC 10317
37	8 A	85	DA 13
38	3 E	86	BASMATI 370
39	14 SEN 8 A	87	PRATAO
40	AGNAKE G	88	TA MAO TAO
41	DV 29	89	JC 1
42	HAGINOMAE MOCHI	90	JC 92
43	MALADY (MDG 439)	91	TSIPALA 421
44	GOUE L	92	DOM ZARD
45	GOUE CC	93	MEHR
46	KOREMOUTOU L	94	ARELATE
47	TELE 1		

are concatenated to produce the final representation of dimension $2dK$.

- Post-processing: the vectors are L2-normalized and then they are power-normalized using $sign(x).|x|^{\gamma}$ where the power parameter γ is found by cross-validation.

Usually, the supervised classification of Fisher Vectors is performed by using a linear classifier as it has been shown that using kernelized techniques on such high-dimensional features does not improve the performances. In our experiments, we used the Support Vector Machine (SVM) algorithm implemented within the LibLinear library (Fan et al. (2008)).

5.2. Convolutional neural networks

Convolutional Neural Networks (CNN) have been mainly used since the 90's for their performances in digit classification. But since a few years, they appear to have now surpassed all state of the art methods for large-scale image classification (Krizhevsky et al. (2012)).

In these experimentations we have used Caffe (Jia et al. (2014)), a Deep Learning Framework, allowing us to use state of the art CNN architectures and models. For the two datasets presented in this work, we have chosen in the Caffe model Zoo the "GoogLeNet GPU implementation" model, based on Google winning architecture in the ImageNet 2014 Challenge Szegedy et al. (2014), and we finetuned this model on our datasets.

The GoogLeNet architecture consists of a 22 layers deep network with a softmax loss as the classifier on top, and is composed of three "inception modules" stacked on top of each other. Each intermediate inception module is also connected to an auxiliary classifier during training, so as to encourage discrimination in the lower stages in the classifier, increase the gradient signal that gets propagated back, and provide additional regularization. These auxiliary classifiers are only used during the training part, and then discarded.

Table 3. Variety list of the IFVGrapeLeaves dataset

ID	Variety name		
1	Arrufiac.B	18	Mauzac.B
2	Bourboulenc.B	19	Merlot.N
3	Cabernet.franc.N	20	Meunier.N
4	Chardonnay.B	21	Mourvedre.N
5	Chenin.B	22	Négrette.N
6	Clairette.B	23	Petit.Courbu.B
7	Colombard.B	24	Petit.Manseng.B
8	Duras.N	25	Piquepoul.blanc.B
9	Fer.N	26	Roussanne.B
10	Gamay.N	27	Sauvignon.B
11	Grenache.gris.G	28	Semillon.B
12	Grenache.N	29	Syrah.N
13	Gros.Manseng.B	30	Tannat.N
14	Lauzet.B	31	Tempranillo.N
15	Len.de.l'El.B	32	Ugni.Blanc.B
16	Macabeu.B	33	Vermentino.B
17	Marsanne.B	34	Viognier.B

Table 5. Parameters of the train stage for the CNN

	base_lr	max iterations	Step size
CiradRiceSeed	0.00001	10000	1000
IFVGrapeLeaves	0.0001	2000	200

6. Experiments

6.1. Setup

For a fair evaluation of both visual classification methods, we first cropped and resized all images to a resolution of 256x256 pixels, that is the size of the images given as input to the first layer of the convolutional neural network as described in section 5.2.

For the Fisher Vector's method, we used a sampling step of 3 pixels and 5 different scales to extract the patches on which the SIFT descriptors were computed. These descriptors were L2-normalized and square-rooted before being reduced to 80 dimensions using PCA similarly to Gosselin et al. (2014). The

Table 6. Classification rates of the evaluated runs on the CiradRiceSeed dataset

System used	Classification rate		
	Top 1	Top 3	Top 5
CNN 1 (= without ImageNet)	8.84	14.86	19.82
CNN 2 (= with ImageNet)	52.38	75.39	86.37
CNN 3 (= with ImageNet + colorimetric data augmentation)	84.95	98.40	99.29
CNN 4 (= with ImageNet + Histogram equalization)	**88.67**	**98.93**	**99.82**
Fisher Vectors	72.57	88.14	94.51
Fisher Vectors + Histogram equalization	76.81	93.98	97.17

Table 7. Classification rates of the evaluated runs on the IFVGrapeLeaves dataset

System used	Classification rate		
	Top 1	Top 3	Top 5
CNN 1 (= without ImageNet)	3.75	5.81	12.59
CNN 2 (= with ImageNet)	9.32	23.12	35.83
CNN 3 (= with ImageNet + colorimetric data augmentation)	**11.50**	**27.60**	**39.70**
CNN 4 (= with ImageNet + Histogram equalization)	11.01	22.76	34.86
Fisher Vectors	9.20	22.28	30.99
Fisher Vectors + Histogram equalization	10.05	21.55	32.57

Gaussian mixture was configured to learn $K = 1024$ visual words which is a good compromise between memory & CPU usage vs. classification performances (as discussed in Sánchez et al. (2013) or Gosselin et al. (2014)). Using more words (e.g. 2048 or 4096) might provide slightly better results but requires switching to more powerful hardware architectures and heavy computation times. As suggested in Gosselin et al. (2014), we applied a power-law normalization of the fisher vectors and did learn the value of the parameter α (i.e. the power value) for each dataset by cross-validation. The regularization parameter C of the SVM (which controls the trade-off between achieving a low error on the training data and minimising the norm of the weights) was also learned by cross-validation for each dataset. The cross-validation was performed using a stratified shuffling procedure with 5 iterations: during each iteration, 10% of the data was randomly sampled to constitute the validation set and the rest was kept for training while preserving the proportion of the different classes in each set.

For the Convolutional Neural Network, we used several strategies that appeared to improve the results within our experiments. The first one (CNN 1) consisted in training the CNN described in section 5.2 from scratch, exclusively on the training data of each dataset and without the use of any external data. As for all other configurations, it makes use of the data augmentation technique implemented within Caffe library and consisting in cropping randomly a 224x224 pixels image, and eventually mirroring it horizontally. The second strategy (CNN 2) rather consisted in fine-tuning a previously trained CNN. We therefore started with the CNN described in 5.2, trained on the popular generalist ImageNet dataset. We then removed its top layers (the fully connected ones) and train this new model using the desired dataset.

The third strategy (CNN 3) is similar to the second one but integrates an additional data augmentation step aimed at improving the robustness of the classifier to the heterogeneous acquisition conditions and camera settings (such as the white balance, the use of the flash, etc.). For each training image, 8 new images were generated by applying a set of colorimetric transformations with randomized parameters, i.e. brightness & saturation modulation in the HSL color space (multiplier factor randomized between 0.8 and 1.2), and contrast modulation (multiplier factor randomized between 0.7 and 1.3). The fourth strategy (CNN 4) targets the same objective of increasing the robustness to colorimetric variations but through a different approach consisting in pre-processing all images (train and test images are both modified) with an histogram equalization (on each RGB channel) instead of using data augmentation. As this strategy provided the best results in the experiments, we also evaluated it for the Fisher Vector method so as to allow a fair comparison.

For all the training strategies of the CNN, the batch size was fixed to 32, and the gamma value to 0.9. Other specific parameters can be found in table 5. During the training, at each iteration, the learning rate is updated according to :

$$base\_lr * gamma^{\lfloor \#iterations/step \rfloor}$$

Note that for the CNN1 configuration, for which the training phase is started from scratch, the base learning rate parameter (base_lr) was multiplied by 100.

6.2. Results

Tables 6 and 7 present the synthesized results of the experiments for respectively the **CiradRiceSeed** dataset and the **IFVGrapeLeaves** dataset. For each run, we provide the average success rate of the classifier on all images of the test set, considering either the best prediction for each test image (Top 1), or the 3 best predictions (Top 3), or the 5 best predictions (Top 5). A first global conclusion that we can derive from the comparison of the two tables is that the performances achieved on the **CiradRiceSeed** dataset are much better than the ones achieved on the **IFVGrapeLeaves** dataset. Whatever

the used classification approach and training strategy, the classification rate on the **IFVGrapeLeaves** dataset actually remains very low, with a maximum of 11.50% good classification rate. On the other side, the best performances on the **CiradRiceSeed** dataset are very good reaching up to 88.67% good classification rate, whereas the number of classes is much higher (95 vs. 34). This shows that the scenario addressed through the **IFVGrapeLeaves** dataset, i.e. the leaf-based identification of the varieties in the field, is a much more challenging problem than the scenario addressed through the **CiradRiceSeed** dataset (in-lab identification). Note that this does not mean that the grape's varieties could never be automatically discriminated based on the visual appearance of their leaf. We actually know that it is to some extent possible thanks to existing morphological identification keys OIV (2009). But it shows that the acquisition protocol should be more constrained (typically by scanning the leaves instead of targeting in the field photographs) or enriched by the use of other traits (e.g. grapes images). We thus plan to enrich the **IFVGrapeLeaves** dataset with such new contents for further evaluations.

On the other side the performances on the **CiradRiceSeed** dataset are clearly better to what we could expect. The visual variability across the 95 species is actually globally low and there are many varieties that are almost impossible to distinguish to the naked eye. Back to the scenarios described within section 3.2 (i.e. labelling control and entity resolution issues), we can even argue that the achieved performances might be sufficient for a practical usage. Rising an alert when the label of a given accession does not belong to the 5 best automatic predictions might for instance be a very effective control tool (the Top 5 classification rate of the best experimented classifier is actually equal to 99.29%). Beyond these scenarios, the achieved performances show that the visual features learned by the classifiers are informative enough to characterize the morphological variability of the different varieties. This open the door to new phenotyping scenarios in the future, for instance if we consider combining such analysis of the visual appearance with the analysis of genomic data or the analysis of other traits.

Let us now look more in details to the scores obtained by the different classification methods on the **CiradRiceSeed** dataset, starting with the different CNN strategies. The weak performances of CNN 1 first show, as one could expect, that the convolutional neural network is not able to learn effective visual features when it is trained on the targeted data only. It is actually well known that this technology requires much more larger training materials. The scores obtained by CNN 2 show that using a network trained beforehand (on a generalist dataset such as ImageNet) provides better performances, even if our fine-grained classification problem is much more specific. The achieved performances are however still lower than the ones achieved by the Fisher Vectors. The most likely reason, confirmed by the better performances of CNN 3 and CNN 4, is that the visual features trained by the CNN on ImageNet are still less robust than the hand-crafted SIFT features to the colorimetric variations occurring in our data (resulting from the heterogeneous acquisition conditions and camera settings of the different observers). The score of CNN 3 shows that using data augmentation as a way to increase the robustness to such transformations does improve the results. A simple histogram equalization (on each RGB channel) applied to all pictures as performed in CNN4 provides an even better improvement and allows the CNN to reach a very high classification score equal to 88.67%. This shows that the visual features trained by the CNN on ImageNet are generic enough to well characterize the pre-processed pictures even if their colorimetric distribution is far from the natural images of ImageNet.

7. Conclusion and perspectives

This paper addressed the problem of categorizing plant images at the variety level, i.e. at a finer taxonomic grain than state-of-the-art studies usually working at the species level. It therefore introduced two new evaluation datasets of agro-biodiversity interest, each being related to concrete scenarios on large-scale resources. A baseline experimental study was conducted on the two datasets to assess whether state-of-the-art classification techniques such as convolutional neural networks and fisher vectors are performant enough to answer the targeted use cases. It did show that the achieved classification performance is very different between the two problems. It is actually pretty bad for the grape leaves collection but much better in the case of the rice seeds collection for which the acquisition protocol was much more constrained and the morpholigical variability between varieties more visible. The conclusion we can draw from these raw results is that automatically identifying plant varieties might already be feasible for some specific scenarios and in controlled environments but that it is still an open problem in the general. In further works, we attempt to explore more in depth the confusion matrix of variety-level visual classifiers in order to correlate the visual similarity of the varieties with genomic data or with other phenotypical properties of the plants.

Acknowledgments

This work has been done throughout the ARCAD-FEDER project, supported by Agropolis Fondation, the French foundation for Agricultural Sciences and Sustainable Development, the European Union through the ERDF, and the Région Languedoc-Roussillon.

References

Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J., 2007a. Sensor network for the monitoring of ecosystem: Bird species recognition, in: Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on, IEEE. pp. 293–298.

Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J., 2007b. Sensor network for the monitoring of ecosystem: Bird species recognition, in: Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on, pp. 293–298. doi:10.1109/ISSNIP.2007.4496859.

Caputo, B., Muller, H., Thomee, B., Villegas, M., Paredes, R., Zellhofer, D., Goeau, H., Joly, A., Bonnet, P., Gomez, J.M., et al., 2013. Imageclef 2013: the vision, the data and the open challenges, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer, pp. 250–268.

Cerutti, G., Tougne, L., Vacavant, A., Coquin, D., 2011. A Parametric Active Polygon for Leaf Segmentation and Shape Estimation, in: 7th International Symposium on Visual Computing, Las Vegas, United States. p. 1. URL: https://hal.archives-ouvertes.fr/hal-00622269.

Ellison, A.M., Farnsworth, E.J., Chu, M., Kress, W.J., Neill, A.K., Best, J.H., Pickering, J., Stevenson, R.D., Courtney, G.W., VanDyk, J.K., 2013. Next-generation field guides .

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874.

Gaston, K.J., O'Neill, M.A., 2004. Automated species identification: why not? Philosophical Transactions of the Royal Society of London B: Biological Sciences 359, 655–667.

van Gemert, J.C., Veenman, C.J., Smeulders, A.W., Geusebroek, J.M., 2010. Visual word ambiguity. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, 1271–1283.

Goëau, H., Bonnet, P., Joly, A., 2015. Lifeclef plant identification task 2015, in: CLEF2015 Working Notes. Working Notes for CLEF 2015 Conference, Toulouse, France, September 8 - 11, 2015, CEUR-WS.

Goëau, H., Bonnet, P., Joly, A., Affouard, A., Bakic, V., Barbe, J., Dufour, S., Selmi, S., Yahiaoui, I., Vignau, C., et al., 2014a. Pl@ ntnet mobile 2014: Android port and new features, in: Proceedings of International Conference on Multimedia Retrieval, ACM. p. 527.

Goëau, H., Bonnet, P., Joly, A., Bakić, V., Barbe, J., Yahiaoui, I., Selmi, S., Carré, J., Barthélémy, D., Boujemaa, N., et al., 2013a. Pl@ ntnet mobile app, in: Proceedings of the 21st ACM international conference on Multimedia, ACM. pp. 423–424.

Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthélémy, D., Molino, J.F., Birnbaum, P., Mouysset, E., Picard, M., 2011a. The imageclef 2011 plant images classi cation task, in: ImageCLEF 2011, pp. 0–0.

Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthélémy, D., Boujemaa, N., Molino, J.F., 2012. The ImageCLEF 2012 Plant Identification Task, in: CLEF'2012: Conference and Labs of the Evaluation Forum, Rome, Italy. URL: https://hal.inria.fr/hal-00960918.

Goëau, H., Joly, A., Bonnet, P., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J.F., 2013b. The imageclef plant identification task 2013, in: Proceedings of the 2nd ACM international workshop on Multimedia analysis for ecological data, ACM. pp. 23–28.

Goëau, H., Joly, A., Bonnet, P., Selmi, S., Molino, J.F., Barthélémy, D., Boujemaa, N., 2014b. Lifeclef plant identification task 2014, in: CLEF2014 Working Notes. Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014, CEUR-WS. pp. 598–615.

Goëau, H., Joly, A., Selmi, S., Bonnet, P., Mouysset, E., Joyeux, L., 2011b. Visual-based plant species identification from crowdsourced data, in: MM'11 - ACM Multimedia 2011, ACM, Scottsdale, United States. pp. 0–0. URL: https://hal.inria.fr/hal-00642236, doi:10.1145/2072298.2072472.

Gosselin, P.H., Murray, N., Jégou, H., Perronnin, F., 2014. Revisiting the fisher vector for fine-grained classification. Pattern Recognition Letters 49, 92–98.

Grillo, O., Draper, D., Venora, G., Martínez-Laborde, J.B., 2012. Seed image analysis and taxonomy of diplotaxis dc.(brassicaceae, brassiceae). Systematics and Biodiversity 10, 57–70.

Hsu, T.H., Lee, C.H., Chen, L.H., 2011. An interactive flower image recognition system. Multimedia Tools Appl. 53, 53–73. URL: http://dx.doi.org/10.1007/s11042-010-0490-6, doi:10.1007/s11042-010-0490-6.

Huang, Y., Wu, Z., Wang, L., Tan, T., 2014. Feature coding in image classification: A comprehensive study. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36, 493–506.

Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C., 2012. Aggregating local image descriptors into compact codes. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, 1704–1716.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 .

Jiang, Y.G., Ngo, C.W., Yang, J., 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval, in: Proceedings of the 6th ACM international conference on Image and video retrieval, ACM. pp. 494–501.

Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., Yahiaoui, I., Carré, J., Mouysset, E., Molino, J.F., et al., 2014a. Interactive plant identification based on social image data. Ecological Informatics 23, 22–34.

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Planque, R., Rauber, A., Fisher, R., Müller, H., 2014b. Lifeclef 2014: multimedia life species identification challenges, in: Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Springer, pp. 229–249.

Kebapci, H., Yanikoglu, B., Unal, G., 2011. Plant image retrieval using color, shape and texture features. Comput. J. 54, 1475–1490. URL: http://dx.doi.org/10.1093/comjnl/bxq037, doi:10.1093/comjnl/bxq037.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V., 2012. Leafsnap: A computer vision system for automatic plant species identification, in: Computer Vision–ECCV 2012. Springer, pp. 502–516.

Larese, M.G., Namías, R., Craviotto, R.M., Arango, M.R., Gallo, C., Granitto, P.M., 2014. Automatic classification of legumes using leaf vein image features. Pattern Recognition 47, 158–168.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, IEEE. pp. 2169–2178.

Mouine, S., Yahiaoui, I., Verroust-Blondet, A., 2012. Advanced shape context for plant species identification using leaf image retrieval, in: Ip, H.H.S., Rui, Y. (Eds.), ICMR '12 - 2nd ACM International Conference on Multimedia Retrieval, ACM, Hong Kong, China. URL: https://hal.inria.fr/hal-00726785, doi:10.1145/2324796.2324853.

Nilsback, M.E., Zisserman, A., 2008. Automated flower classification over a large number of classes, in: Computer Vision, Graphics Image Processing, 2008. ICVGIP '08. Sixth Indian Conference on, pp. 722–729. doi:10.1109/ICVGIP.2008.47.

OIV, 2009. Oiv descriptor list for grape varieties and vitis species .

Orru, M., Grillo, O., Venora, G., Bacchetta, G., 2012. Computer vision as a method complementary to molecular analysis: Grapevine cultivar seeds case study. Comptes rendus biologies 335, 602–615.

Perronnin, F., Dance, C., 2007. Fisher kernels on visual vocabularies for image categorization, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE. pp. 1–8.

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the fisher kernel for large-scale image classification, in: Computer Vision–ECCV 2010. Springer, pp. 143–156.

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2008. Lost in quantization: Improving particular object retrieval in large scale image databases, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. pp. 1–8.

Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J., 2013. Image classification with the fisher vector: Theory and practice. International journal of computer vision 105, 222–245.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep Fisher networks for large-scale image classification, in: Advances in Neural Information Processing Systems.

Sivakumar, V., Anandalakshmi, R., Warrier, R.R., Singh, B., Tigabu, M., Nagarajan, B., 2013. Discrimination of acacia seeds at species and subspecies levels using an image analyzer. Forest Science and Practice 15, 253–260.

Sivic, J., Zisserman, A., 2003. Video google: A text retrieval approach to object matching in videos, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE. pp. 1470–1477.

Smykalova, I., Grillo, O., Bjelkova, M., Hybl, M., Venora, G., 2011. Morphocolorimetric traits of pisum seeds measured by an image analysis system. Seed Science and Technology 39, 612–626.

Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, Y.H.J., Fisher, R.B., Nadarajan, G., 2010. Automatic fish classification for underwater species behavior understanding, in: Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams, ACM. pp. 45–50.

Spampinato, C., Mezaris, V., van Ossenbruggen, J., 2012. Multimedia analysis for ecological data, in: Proceedings of the 20th ACM international conference on Multimedia, ACM. pp. 1507–1508.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going deeper with convolutions. CoRR abs/1409.4842. URL: http://arxiv.org/abs/1409.4842.

This, P., Lacombe, T., Thomas, M.R., 2006. Historical origins and genetic diversity of wine grapes. TRENDS in Genetics 22, 511–519.

Trifa, V.M., Kirschel, A.N.G., Taylor, C.E., Vallejo, E.E., 2008. Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. Journal of The Acoustical Society of America 123. doi:10.1121/1.2839017.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y., 2010. Locality-constrained linear coding for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE. pp. 3360–3367.

Yang, J., Yu, K., Gong, Y., Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 1794–1801.