International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 14939-14951

http://www.ripublication.com

Multi View K-SVD Dictionary Learning for Plant Identification

Soodabeh Safa and Fatimah Khalid

Lili Nurliana Abdullah and Azreen Azman
Department of Multimedia,
Faculty of Computer Science and Information Technology,
University Putra Malaysia, Serdang, 43300, Malaysia
soodabeh.safa@gmail.com, Patiaziz@gmail.com, liyana@upm.edu.my,
azreen.azman@gmail.com

Abstract

These days, various forms of data can be seen in real word. Images as a popular type of data are widely used in many resources such as web pages, video, etc. Meanwhile, each images as a source of data is comprised of many features. Different methods of feature extraction present occasionally different descriptions of the data, each can be considered as one view of data. In this study, K-Singular Value Decomposition (K-SVD) method which is a popular method for sparse dictionary learning is developed for multi view data. Sparse dictionary learning has been applied in multi view data identification problem to find out which classes belongs to basis vector. Different feature extraction methods such as Scale-Invariant Feature Transform (SIFT) and GIST of a scene are used for plant identification to describe images of leaves in many studies, but in most of them, these different views of data are merged as high dimensional features. In this study K-SVD, as a dimension reduction algorithm is developed for sparse dictionary learning for multi view data create low dimensional visualization of images in plant identification scope. The experimental results show improvement in accuracy compared to regular K-SVD and early fusion of features.

keywords: Multi view data, K-SVD algorithm, Plant identification, Sparse dictionary learning, Regularizer, Basis vector (atom)

1. Introduction

Multi-view data is referred to a data with different feature generation methods [1] and [2]. For example, consider an image consisting of several views: Background information and objects in the image that everyone is considered as one view. Also

various feature extraction methods with extracting features in a different way, can considered as different views. Several aspects of the multi-view data have been used in [3] and [4]. These features are extracted by different methods such as SIFT, color histogram, the distribution of edges, wavelet texture [3] and [5]. There are different approaches to deal with such multi-view data. Dictionary Learning for Sparse representation to transfer the feature vector to a good space is recently used by many researchers.

2. Related Work

In this study $\mathcal{X} = \{x^1, \dots, x^N\}$ is assumed as a training data set. Number of training data set is shown by S. Every data has P view. Every view is one feature vector. In the other word we have

$$x^i = \left(x_1^i, \dots, x_p^i\right), i \in \{1, \dots, N\}. \quad (1)$$

Each of these views are described by a feature extraction method. In this study we used SIFT and GIST features for describing images.

Different methods for multisource data fusion can be divided into three categories, based on the time of fusion:

- Early Fusion: The approach of using concatenated high-dimensional features
- Middle Fusion: Model is trained simultaneously considering separation of scenes
- Late Fusion: The approach of replacing the high-dimensional learning task by multiple low-dimensional learning tasks and then fusing the results.

The early and late fusions have their own advantages and disadvantages. The curse of dimensionality in early fusion and the difficulty in determining proper weights for late fusion are examples of each method's drawbacks.

Different methods have been proposed for multi-view data fusion in related works that is mentioned as follows:

1) Bagging and Boosting methods
2) Multiple Kernel Combination methods: Such models train a weight for each type of features. When multiple types of features are combined together, all features from the same type are weighted equally. In [6] and [7] researchers trained weights in a supervised way based on soft margin support vector machine kernel. In [9] researchers reviewed all methods based on combination of kernels.
3) Structured Sparsely Regularization: In the approaches based on structured sparsely regularization, ℓ_1 (Norm 1) and ℓ_2 (Norm 2) are used for structured dictionary learning [8]. In [3] weights are trained for labels prediction in K tasks simultaneously using weighting methods. so weights vector is as :

$$W = \left[w_1^1, \dots, w_1^K; \dots; w_p^1, \dots, w_p^K\right], (2)$$

$$\min_{W} \sum_{k=1}^{K} f_k(w^k, b_k) + 2\gamma_1 \sum_{k=1}^{K} ||w_k||_G + \cdots . (3)$$

The group lasso regularizer is added in (3) to impose interrelationship of views and features as :

$$||w_i||_G = \sum_{j=1}^{p} ||w_j^p||_2$$

The first part of equation (3) is related to error in to prediction tasks. With ℓ_1 and ℓ_2, in the model, features belong to each weight in a similar way. Researchers use these mentioned norms in a way that for each task, just particular view that is related to that task is considered. Researchers applied this idea in the [10] in multitask clustering problems. Applying ℓ_2 norm on the groups of weights (on the atoms) causes group action in weighting process.

For the first time this trait is used for structured sparse representation and so far used widely for multitask application in [3], [8] and [18]. In [12] researchers applied related method to structured sparse representation. Figure 1 illustrates the frame work of proposed method in this study.

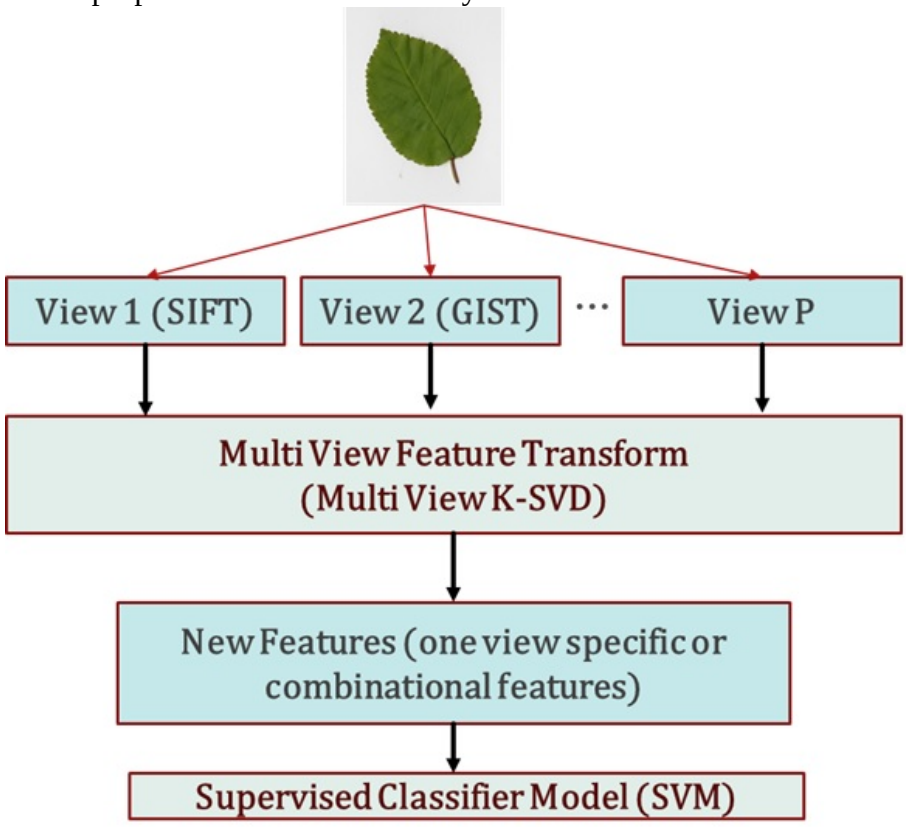

Figure 1. The frame work of proposed method

3. Feature Extraction

A. SIFT Descriptor

As SIFT descriptors are widely used in pattern recognition, signal processing and computer vision and suitable for classification, in this study we extracted SIFT descriptors.

For SIFT descriptors, instead of orderless bag of features, spatial pyramid [11] is followed. This technique works by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-region. The result of spatial pyramid is simple and computationally efficient compared to regular bag of features representation. There are different ways for SIFT feature detection and representation: regular grid and interest point detector. In this study we used the first one. Figure 2 shows an example of dense interest points on a regular grid in a leaf image. SIFT descriptors of 16*16 pixel patches computed over a grid with spacing of 8 pixels. Pyramid SIFT features, have 3 levels with 50 k-means clustering center that lead to 1050 vector size as follows:

$$1*50 + 4*50 + 16*50 = 1050$$

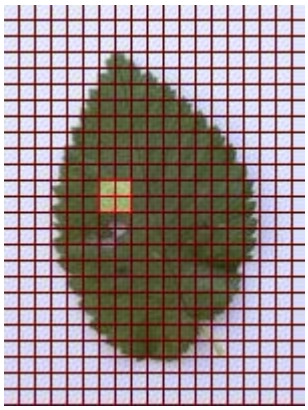

Figure 2. Example of dense interest points on a regular grid

B. GIST of a scene

We also used Torralba's GIST descriptor [12]. The GIST descriptor has recently shown good results for image classification and image search. The vector size for GIST features are 512. The image is divided into a 4-by-4 grid for which orientation histograms are extracted. GIST descriptor is similar in spirit to the local SIFT descriptor.

4. The K-SVD Algorithm

Feature extraction by learning a transform from the previous space is possible. In some transforms, such as the Fourier transform, the atoms are fixed and independent of the context. On the other hand, in other transforms, like Principal Component

Analysis **(**PCA) and sparse representation, the atoms are context dependent. In other words, in these types of transforms, the atoms are extracted from the domain samples. As an example, in the discrete Fourier transform, the atoms are pre-fixed and it makes no difference for each data set with the same dimensions of the feature space. However, in PCA transform, the data covariance matrix is calculated first. Then making use of eigenvectors of this matrix, the transform for the desired domain is achieved. Extracting the atoms in the sparse representation is called "Dictionary Learning" or "Code Word Learning".

In the dictionary learning, the dictionary should be trained in the way, the data view is sparse data into a new space. With applying sparseness condition on the new space, the high level atoms is extracted, on the other hand it can reproduce all the signals with little error, the patterns will be extracted feature vectors. The basic formula of dictionary learning optimization problem for X training data set is as follows.

– (4)

The dictionary learning optimization problem is a convex problem in the case of D or α is a, is fixed, but it is not convex if they are both unknown. Figure 3 shows the K-SVD model.

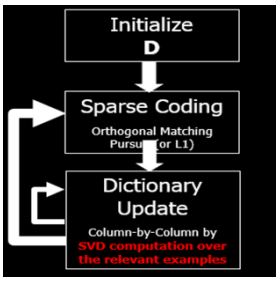

Figure 3. The K-SVD model

The K-SVD algorithm is generalizing the K-means clustering process. K-SVD is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the atoms in the dictionary to better fit the data. The update of the dictionary columns is combined with an update of the sparse representations.

The advantage of this method of learning multi view dictionary compared to using the group-lasso restriction is that by the direct application of distinct views of feature space, the group lasso restriction is omitted from the optimization problem.

The discussed algorithms for learning dictionary, for their simplicity, speed as well as efficiency seem attractive to most researchers. The task of compound atoms is the analysis of the correlation between various views. The common patterns among the data in the view are extracted by single atoms. By the elimination of redundancies and the analysis of high frequency patterns between feature vectors, the dictionary

learning methods for sparse representation prevent over-fitting model classifier. Figure 4 and 5 show the K-SVD algorithm.

The K-SVD algorithm

Find the best dictionary to represent the data samples $\{x^i\}_{i=1}^N$ as sparse compositions, by solving

$\min_{D,\alpha}\{||X - D\alpha||_F^2\}$ subject to $\forall_i, ||\alpha^i||_0 \leq T_0, where\ D = [d^1, \dots, d^m]$

$\in R^{n\times m}, \alpha^i = [\alpha^i(1), \dots, \alpha^i(m)]^T \in R^m.$

Initialization: Set the dictionary matrix $D^0 \in R^{n\times m}$ with ℓ^2normalized columns. Set $J = 1$.

Repeat until convergence (stopping rule):

- *Sparse Coding Stage*: Use any pursuit algorithm to compute the representation vectors α_i, for each example x_i, by approximating the solution of

  $i = 1,2, \dots, N$ $\min_{\alpha^i}\left\{\left||x^i - D\alpha^i|\right|_F^2\right\}$ subject to $\left||\alpha^i|\right|_0 \leq T_0.$

- Codebook Update Stage :For each column $k = 1,2, \dots, m\ in\ D^{(J-1)}$, Update it by

-Define the group of examples that use this atom,

$G^k = \{i|1 \leq i \leq N, \alpha^i(k) \neq 0\}.$

-Compute the overall representation error matrix, E^k, by

$E^k = Y - \sum_{j\neq k} d^j\alpha(j).$

-Restrict E^k by choosing only the columns corresponding to G^k, and obtain E_R^k.

-Apply SVD decomposition $E_R^k = U\Delta V$. Choose the updated dictionary column d^k to be the first column of U . Update the coefficient vector $\alpha^{(R)}(k)$ to be the first column of Vmultiplied by $\Delta(1,1)$.

- Set *J=J+1*

Figure 4. The K-SVD algorithm, The problem definition

The K-SVD algorithm

Find the best dictionary to represent the data samples $\{x^i\}_{i=1}^N$ as sparse compositions, by solving

$\min_{D,\alpha}\{||X - D\alpha||_F^2\}$ subject to $\forall_i, ||\alpha^i||_0 \leq T_0, where\ D = [d^1, \dots, d^m] \in R^{n\times m}, \alpha^i = [\alpha^i(1), \dots, \alpha^i(m)]^T \in R^m.$

Initialization: Set the dictionary matrix $D^0 \in R^{n\times m}$ with ℓ^2normalized columns. Set $J = 1$.

Repeat until convergence (stopping rule):

- *Sparse Coding Stage*: Use any pursuit algorithm to compute the representation vectors α_i, for each example x_i, by approximating the solution of

  $i = 1,2, \dots, N$ $\min_{\alpha^i}\left\{\left||x^i - D\alpha^i|\right|_F^2\right\}$ subject to $\left||\alpha^i|\right|_0 \leq T_0$.

- Codebook Update Stage :For each column $k = 1,2, \dots, m\ in\ D^{(J-1)}$, Update it by

-Define the group of examples that use this atom,

$G^k = \{i|1 \leq i \leq N, \alpha^i(k) \neq 0\}$.

-Compute the overall representation error matrix, E^k, by

$E^k = Y - \sum_{j\neq k} d^j\alpha(j)$.

-Restrict E^k by choosing only the columns corresponding to G^k, and obtain E_R^k.

-Apply SVD decomposition $E_R^k = U\Delta V$. Choose the updated dictionary column d^k to be the first column of U . Update the coefficient vector $\alpha^{(R)}(k)$ to be the first column of Vmultiplied by $\Delta(1,1)$.

- Set *J=J+1*

Figure 5. The K-SVD algorithm

5. Proposed Multi View K-SVD Algorithm

In the proposed method, just apply on the part of feature vector that is relevant to correlated views of that atom.

The priority of this method for multi view dictionary learning in compare to group lasso is that with direct applying of distinct features in different views the group lasso is deleted from optimization problem.

In order to develop K-SVD algorithm for considering distinct views of feature space (P view), the basis vectors or "atoms" of the dictionary is divided into (P+1)

groups. In the first P dictionary, the atoms each has value in just one view (figure 6). The dictionary includes a mixture of atoms of all views :

(5)

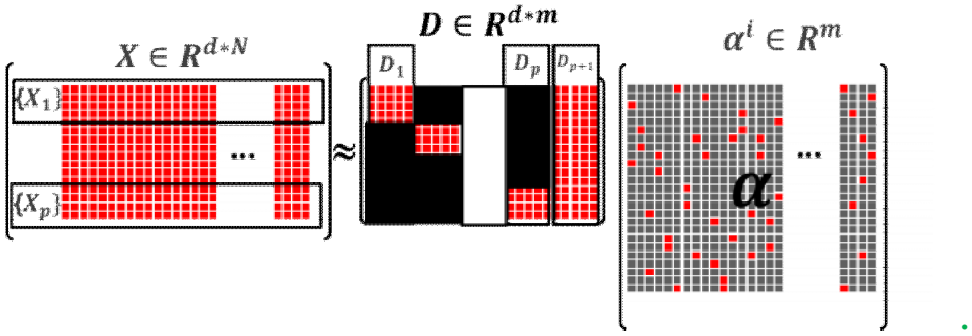

Figure 6. The dictionary presentation

The atoms of Dj dictionary for j<=p can have value just in the locations corresponding to the same view. The sparse representation of i data is:

(6)

For multi view learning we change the K-SVD algorithm as figure 7 and 8.

The Multi View K-SVD algorithm

Find the best dictionary to represent the data samples $\{x^i = (x_1^i, \dots, x_p^i)\}_{i=1}^N$ as sparse compositions, by solving

$\min_{D,\alpha} \left\{ ||X - D\alpha| |_F^2 \right\}$ subject to $\forall_i, \left| |\alpha^i| \right|_0 \leq T_0, where$

$D = [D_1, \dots, D_p, D_{p+1}] \in R^{n \times m}, \alpha^i = [\alpha_1^i; \dots; \alpha_p^i; \alpha_{p+1}^i] \in R^m, m = \sum_{l=1}^{p+1} m_l,$

$D_l = [d_l^1, \dots, d^{m_l}] \in R^{n \times m_l}, \alpha_l^i = [\alpha_l^i(1), \dots, \alpha_l^i(m_l)]^T$
$\in R^{m_l}\ for\ every\ l\ in\ \{1, \dots, p+1\}.$

-Define $View^k$ as a column vector:

$View^k = \left[0^{\sum_{j=1}^{l-1} \dim(view_j)}; 1^{\dim(view_l)}; 0^{\sum_{j=l+1}^{p} \dim(view_j)}\right]\ for\ k$
$= \left\{ \sum_{j=1}^{l-1} m_j + 1, \dots, \sum_{j=1}^{l} m_j \right\}\ such\ that\ l \in \{1, \dots, p\},$

$View^k = [1^n]\ for\ k = \left\{ \sum_{j=1}^{p} m_j + 1, \dots, \sum_{j=1}^{p+1} m_j \right\}.$

Figure 7. The Pseudo code of multi view K-SVD algorithm, The problem definition

Initialization: Set the dictionary matrix $D^0 \in R^{n\times m}$ with ℓ^2normalized columns that every atom only can have non-zero elements in corresponding View. Set $J = 1$.

Repeat until convergence (stopping rule):

- *Sparse Coding Stage*: Use any pursuit algorithm to compute the representation vectors α_i, for each example x_i, by approximating the solution of

$$i = 1,2,\dots,N \ \min_{\alpha^i}\left\{\left|\left|x^i - D\alpha^i\right|\right|_F^2\right\} \text{ subject to } \left|\left|\alpha^i\right|\right|_0 \le T_0.$$

- Codebook Update Stage :For each column $k = 1,2,\dots,m \ in \ D^{(J-1)}$, Update it by

-Define the group of examples that use this atom,

$$G^k = \{i | 1 \le i \le N, \alpha^i(k) \ne 0\}.$$

Compute the overall representation error matrix, E^k, by

-$E^k = Y - \sum_{j\ne k} d^j \alpha(j)$.
-Restrict E^k by choosing only the columns corresponding to G^k and rows corresponding to 1's in $View^k$, and obtain E_R^k.
-Apply SVD decomposition $E_R^k = U\Delta V$.
Define d_{View}^k as rows of d^k corresponding to 1's in $View^k$.
-Choose the updated dictionary column d_{View}^k to be the first column of U. Update the coefficient vector $\alpha^{(R)}(k)$ to be the first column of V multiplied by $\Delta(1,1)$.

- Set $J=J+1$

Figure 8. The pseudo code of multi view K-SVD

The advantage of this method of learning multi view dictionary compared to using the group-lasso restriction is that by the direct application of distinct views of feature space, the group lasso restriction is omitted from the optimization problem.

The discussed algorithms for learning dictionary, for their simplicity, speed as well as efficiency seem attractive to most researchers. The task of compound atoms is the analysis of the correlation between various views. The common patterns among the data in the view are extracted by single atoms. By the elimination of redundancies and the analysis of high frequency patterns between feature vectors, the dictionary learning methods for sparse representation prevent over-fitting model classifier.

6. Dataset

The Pl@ntLeaves dataset is used within the image CLEF2012 [13] which contained 11572 images: 6630 scans, 2726 scan-like images and 2216 photographs. The three image types illustrated with the same species *Celtis australis* L in figure 9. The most important motivation for choosing this dataset of images is the wide diversity of image rotation, scale, noise and luminance. These diversities of plant leaf images provide a major challenges to the researchers to provide a suitable automatically plant species classification methods. Statistics of the composition of the training and test data is shown in table 1.

(a) (b) (c)

Figure 9. Three image types illustrated with the same species *Celtis australis* L, scan (a), pseudo-scan (b) and photo categories (c)

Table 1. Statistics of the composition of the training and test data

Image Type		Pictures	Individual Plants
Scan	Train	4879	310
	Test	1760	157
Scan-like	Train	1819	118
	Test	907	85
Photograph	Train	1733	253
	Test	483	213
All	Train	8422	681
	Test	3150	455

7. Experimental Results

In this study, Pyramid SIFT features, have 3 levels with 50 and 1050 vector size. The vector size for GIST feature is 512. The feature vectors are normalized to 1. Soft Margin Support Vector Machine is used as classifier. The number of vectors in dictionary and soft margin regularization are set with 5-fold cross validation. For

multi-view dictionary the number of basis vectors or atoms in all groups are same. The results are given in Table 2.

Table 2. The accuracy results of early fusion, regular K-SVD method and proposed multi view K-SVD method with SIFT and GIST features.

Methods	Dictionary	Scan	Scan-Like	Photograph	Average
Early Fusion (SIFT +GIST)	----------	0.437	0.413	0.412	0.420
K-SVD(SIFT+GIST)	m=300 T_0=20	0.4659	0.4598	0.447	0.461
Multi View K-SVD(SIFT+GIST)	m=300 m1=m2=m3=100 T_0=20	0.5250	0.5314	0.474	0.519

8. Conclusion and Feature Work

In this study, K-Singular Value Decomposition (K-SVD) as a dimension reduction algorithm to create low dimensional visualization of images for sparse dictionary learning is developed for multi view data of plant image leaves. The experimental results of multi view K-SVD show improvement in the plant identification task compared to regular K-SVD and also early fusion of features. Scan and scan-like categories obtained significantly better results than photo images. For the feature work we intend to develop other dictionary learning algorithm for multi view data.

References

[1] Lanckriet, Gert RG, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. 2004. "Learning the kernel matrix with semidefinite programming." The Journal of Machine Learning Research 5: 27-72.

[2] Xia, Hao and Wu, Pengcheng and Hoi, Steven C. H. 2013. "Online multi-modal distance learning for scalable multimedia retrieval." Proceedings of the sixth ACM international conference on Web search and data mining 455-464.

[3] Schölkopf, Bernhard, Ralf Herbrich, and Alex J. Smola. 2001. "A generalized representer theorem." Computational learning theory. Springer Berlin Heidelberg. 416-426.

[4] Sonnenburg, Sören, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. 2006. "Large scale multiple kernel learning." The Journal of Machine Learning Research 7: 1531-1565.

[5] Lanckriet, Gert RG, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. 2004. "Learning the kernel matrix with semidefinite programming." The Journal of Machine Learning Research 5: 27-72.

[6] Hoi, Steven CH, Rong Jin, Peilin Zhao, and Tianbao Yang. "Online multiple kernel classification." Machine Learning 90, no. 2 (2013): 289-316.

[7] Lin, Yen-Yu, Tyng-Luh Liu, and Chiou-Shann Fuh. 2008. "Dimensionality reduction for data in multiple feature representations." Advances in Neural Information Processing Systems 21 961-968.

[8] Kim, Seyoung, and Eric P. Xing. 2010. "Tree-guided group lasso for multi-task regression with structured sparsity." Proceedings of the 27th International Conference on Machine Learning (ICML-10): 543-550.

[9] McFee, Brian, and Gert Lanckriet. 2011. "Learning multi-modal similarity." The Journal of Machine Learning Research 12: 491-523.

[10] Liu,Weifeng, Dacheng Tao, Jun Cheng, and Yuanyan Tang. 2014. "Multiview hessian discriminative sparse coding for image annotation." Computer Vision and Image Understanding 118 50-60.

[11] 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 (Vol. 2, pp. 2169-2178).

[12] Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision, 42(3), 145-175.

[13] http://www.imageclef.org/2012/plant.