

Pollen Recognition for Allergy and Asthma Management Using GIST Features

Natalia Khanzhina^(✉) and Evgeny Putin

Computer Technologies Lab, ITMO University,
49 Kronverksky Pr, 197101 St. Petersburg, Russia
nehanzhina@gmail.com, putin.evgeny@gmail.com

Abstract. In this paper we propose a way of managing allergy and asthma based on pollen recognition using images from an optical microscope. GIST descriptors are extracted as features. Our research can help to automate a time-consuming process of pollen grains classification, which is usually performed by highly qualified palynologists, and to create a real-time system of immediate notification about high atmospheric allergenic pollen concentration. Standard machine learning methods are applied and results are compared on different pollen datasets. The best model is support vector machine with 95.2% of accuracy on 9 pollen species and 98.3% on 5 pollen species.

Keywords: Allergy management · Asthma management · Image recognition · GIST · Machine learning · Dimension reduction · Pollen grains · Image preprocessing

1 Introduction

Today almost 30% of people have allergies, 8% have asthma. The most frequent origin of allergies and one of the causes of asthma is pollen. The number of people suffering of pollinosis varies between 10–15% among different countries, this number increased by 34% over last ten years because of urbanization, environmental effects of human, and also because pollen can cover long distances by air [24].

In order to manage allergies and asthma symptoms it is necessary to determine the start of the pollen dispersion. Accurate knowledge of prevalent aeroallergens can improve the diagnosis and treatment of patients. Pollen information is the key as it enables a timely start of the preventive and symptomatic treatment of seasonal allergy problems. Thus, a great need exists to catch airborne pollen and to determine immediately whether it is an allergy-causing plant species pollen or not. For these goals there exist more than 600 pollen counting stations all over Europe and only about 20 stations in Russia, where palynologists and volunteers spend much time for manual pollen operation using microscopes [24]. However, manual operation cannot provide information relevant enough for patients. For instance, 24% of adults and 40% of children in Europe cannot

travel freely due to the lack of information on atmospheric pollen concentrations in different regions in Europe [11, 19].

Thus, a near real-time system, which can automate the recognition of pollen species, is required. Development of such a system can be achieved on the basis of the usage of digital images from a microscope. Recently machine learning and, particularly, deep learning have proven their effectiveness in a variety of applications such as image classification [21, 32], natural language processing [7, 33], speech recognition [10, 16].

The need to automate pollen recognition was mentioned by Flenley for the first time in 1968 [12]. Since that time many attempts of such system development have been made, however, the problem is not completely solved yet. Proper classification of pollen grains allows to draw the appropriate conclusions and to solve problems faced by experts in other areas, not only aeropalynology [6, 29, 31].

Image recognition-based solution for this task consists of the following steps: pollen extraction, counting, and classification. Initially the image can include from 1 to about 50 pollen grains depending on their size and shape. Pollen extraction is the search of areas on image containing only one pollen grain per area and following pollen grain contouring. It can be obtained after preprocessing steps, described in Sect. 3.2. Counting is the quantitation of such extracted pollen grains. And classification is the determination of each pollen grain species. The final result can be presented as the percent composition of pollen species.

All researchers in this area extracted specific pollen features such as shape, brightness, texture features, and aperture [3–5, 27]. Some used a scanning electron microscope (their results vary between 77% and 97% of accuracy) [1, 3, 31], other used stacks of images of one pollen, a kind of three dimensional representation (resulting accuracy is between 93.8% and 97.5%) [3, 30, 31]. Most researchers used standard machine learning methods: support vector machine, linear discriminant analysis, random forest, artificial neural networks, k-nearest neighbors and others. Many authors are members of currently existing or past global research projects, aimed to develop an automated pollen recognition tool. For instance, the European project ASTHMA specifically dealt with allergic pollen [28].

Review of pollen recognition techniques [17] revealed, that some simple and local issues within pollen recognition might be carried out, but there were still many tasks related to deformed, clumped pollen, which were not resolved. The interest to the problem is still high. Recently published papers declared results obtained with an optical microscope to be between 87% and 99% of accuracy [6, 9, 23, 27, 29, 30]. However, only few works considered the steps of extraction and pollen counting, although they are very important parts of the problem, because manual image cropping could be tedious and automatic counting is the main goal of recognition in some cases. Our research bypasses these disadvantages. Also we use images from an optical microscope, which is much cheaper than scanning electron microscope and is widely used.

Extracted features are described in Sect. 2.1. Applied dimension reduction techniques are described in Sect. 2.2. To achieve the goals of extraction and

counting we use a preprocessing algorithm, which is described in Sect. 3.2. Applied classifiers are described in Sect. 3.3. The experiments are described in Sect. 3.4. Results are discussed in Sect. 4.

2 Proposed Approach

2.1 GIST Features

We choose GIST descriptors [8, 26] as image features, which allows to avoid specific-purpose feature extraction. GIST is a low-dimensional scene representation. In other words, it is a kind of edges distribution histogram. An image is divided into equal parts using a grid (4×4 in our case). Edge distributions are computed on 3 scales of the image separately for every part. Edge distribution corresponds to the response of the part to every edge orientation (which has 8 or 4 values). We use color images, so this is applied to every color channel. As a result of GIST extraction, 960 descriptors were obtained. In general, the number of GIST features can be arbitrary.

2.2 Dimension Reduction

Due to the high number of GIST descriptors, dimension reduction (DR) is required. The following methods were used.

ReliefF. ReliefF is a member of the Relief algorithm family, which is a filtering feature selection technique, extended on M-classes classification. Relief is based on near-hit and near-miss measures, values of which form the weight for each feature. If the value of the weight is smaller than some threshold, this feature is rejected [34]. Weights vector is computed according the following formula:

$$w_i = \sum_{k=1}^p \left(\delta \left(x_k^i, near_miss(x_k)^i \right)^2 - \delta \left(x_k^i, near_hit(x_k)^i \right)^2 \right) \quad (1)$$

where $i = 1 \dots n$; n is the number of features; p is the number of objects; and $\delta(a, b)$ is the Kronecker delta.

The number of features selected by applying ReliefF is 300.

Mutual Information. Mutual information (MI) implies feature relative importance. It relies on entropy of a feature and its conditional entropy related to every class of objects [20]:

$$I(x, y) = H(x) - H(x|y) \quad (2)$$

where I is the relative importance; $H(x)$ is the entropy of a feature; $H(x|y)$ is the conditional entropy.

The number of features selected by applying MI is 300.

Principal Component Analysis. Principal Component Analysis (PCA) is a feature extraction method. It finds a projection to a linear manifold minimizing distance of the points to the manifold [22]. 95% of origin variance of the data were used.

3 Experiments

3.1 Materials

Current research is carried out not only on allergenic plant, but also on honey plant pollen. The approach can be easily generalized to be applied to any plants dataset. The dataset includes 9 species, almost 1800 images in total. The dataset is original, never used before, made using optical microscope Olympus BX51 with Olympus DP71 image viewing system. All the pollen types were collected mostly from Russia, Perm Krai. In the Perm region, the aeropalynological profile is typical for central Russia. On average, the concentration of allergenic pollen grains in the air of Perm is lower than in other European geographical regions. Since 2010, the aeropalynological data of the Perm region have been included in the Russian pollen monitoring program. Pollen traps are located in the city center [24].

An example of an image from the dataset is presented in Fig. 1. The example shows that an image can contain stains, or debris, which are cause of wrong segmentation.

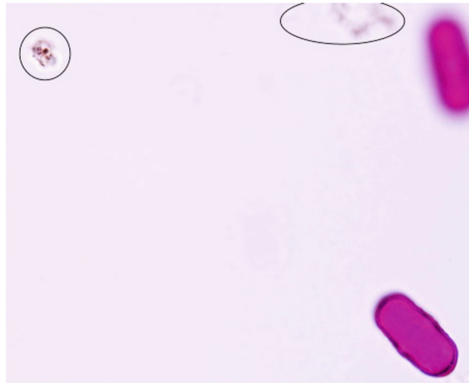











Fig. 1. Input image example

Some examples of each pollen species are presented in Table 1.

We used two versions of the dataset: full, which contains similar shape species, and partial, which contains mostly different shape species (top 5 rows of the table).

Table 1. Preprocessed images examples

Species	Images	Total
Trifolium hybridum		200
Archangelica officinalis		200
Dianthus deltoides		199
Fagopyrum esculentum		200
Chamerion angustifolium		198
Dianthus deltoides		110
Bunias orientalis		198
Salix alba		199
Tilia cordata		60

All images were normalized by RGB-values, according to the following formula:

$$I_N = (I - Min) \frac{newMax - newMin}{Max - Min} + newMin \tag{3}$$

where I stands for old pixel color value and I_N is a new value.

Cross-validation was used to evaluate the results. Its idea is to divide the dataset into disjoint training and validation subsets K different ways, the accuracy is evaluated as the mean accuracy.

We used 10-fold cross-validation and the experiments were conducted on a computer with an Intel Core i7-3770 CPU with 16 GB of RAM.

3.2 Preprocessing

We performed three preprocessing steps:

1. The first step of preprocessing is noise reduction, including Gaussian blur, dilation and erosion functions.
2. The next step is image double- and low-thresholding applied to hue and saturation channels. Such combination shows high result on images with color gradient or hotspots.
3. The last step is the segmentation and localization provided by Canny edge detector and Hu-moments [18].

The resulting sequence of preprocessing steps is presented in Fig. 2.

The extraction (segmentation) accuracy is 73%. The result is not great, the main cause of that is clumped pollen grains (Fig. 3). This is a separate complicated issue and an object of further research.

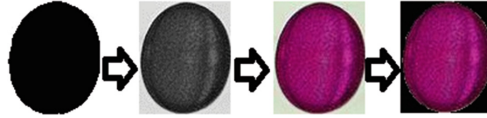


Fig. 2. Image modifications during preprocessing

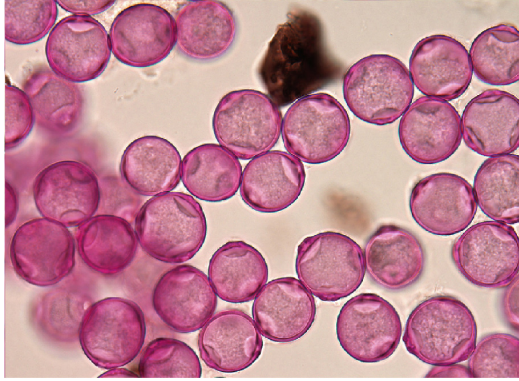


Fig. 3. Clumped pollen example

From here we will call the dataset which passed the preprocessing steps as the preprocessed dataset.

3.3 Models

The following 6 machine learning techniques were used in the research for classification [2, 13–15, 25].

1. Logistic regression (LR). A simple machine learning technique of linear classification.
2. K-nearest neighbors (kNN). This is a metric classification technique, which defines object class by its k nearest neighbors.
3. Support vector machine (SVM). It solves the problem of nonlinearly separable input vectors by projection of the low-dimensional training data into a higher dimensional feature space where they can be easily separated. The projection is achieved using kernel functions.
4. Decision trees (DT). The main idea is to recursively set up a tree over the feature space. The feature space is split with a feature value and then both subsets are split the same way recursively until the tree leaf has the minimum number of class targets for making a decision.
5. Random forest (RF). A classifier ensemble method based on bagging. Several independent models make decisions, then the common decision is determined by voting in case of classification problem and by averaging in case of regression problem.

Table 2. The results on partial dataset

Model	Origin features	PCA	Relieff	MI
LR	75.2 ± 4.7%	52.6 ± 4.8%	63.2 ± 2.9%	69.5 ± 4.6%
kNN	82.6 ± 3.4%	80.6 ± 3.3%	81.5 ± 3.6%	82.1 ± 2.5%
SVM	73.3 ± 4.7%	78.1 ± 3.8%	69.4 ± 3.5%	73.1 ± 4.9%
DT	79.5 ± 3.0%	73.7 ± 3.0%	79.1 ± 3.0%	78.4 ± 2.2%
RF	84.7 ± 3.8%	77.9 ± 3.1%	85.6 ± 3.5%	83.9 ± 3.0%
GB	83.1 ± 3.2%	76.2 ± 2.9%	84.3 ± 3.7%	82.4 ± 3.1%

- 6. Gradient boosting (GB). This is a modern machine learning technique of classifiers ensemble. It minimizes the training error of classifiers linear composition by gradient descent.

3.4 Results for Different Feature Sets and Different Machine Learning Models

Each table shows combinations of dimension reduction and classification methods. Each cell in the resulting tables contains the mean accuracy of 10-fold cross-validation and its standard deviation, which follows after the plus/minus sign. The each DR method best accuracy is highlighted in bold.

Table 2 shows results comparison on the partial dataset. The best accuracy is provided by the RF model with Relieff DR method, it is 85.6 ± 3.5%.

Table 3 shows results comparison on the partial preprocessed dataset. The best accuracy is provided by the SVM model with MI DR method, the accuracy is 98.3 ± 2.1%.

Table 4 shows results comparison on the full dataset. The best result is provided by the RF model with no DR, the accuracy is 78.5 ± 3.8%.

Table 5 shows results comparison on the full preprocessed dataset. The best accuracy is provided by the SVM model with PCA DR method, the accuracy is 95.2 ± 1.7%.

Table 3. The results on partial preprocessed dataset

Model	Origin features	PCA	Relieff	MI
LR	94.8 ± 2.2%	91.7 ± 2.0%	93.0 ± 3.2%	93.8 ± 2.9%
kNN	92.8 ± 2.2%	93.2 ± 2.7%	94.5 ± 2.9%	95.1 ± 3.1%
SVM	95.3 ± 1.9%	97.0 ± 1.2%	97.7 ± 2.0%	98.3 ± 2.1%
DT	79.4 ± 3.5%	81.9 ± 3.6%	84.2 ± 4.3%	84.9 ± 3.8%
RF	91.6 ± 3.2%	93.4 ± 3.0%	95.7 ± 3.4%	96.2 ± 3.6%
GB	92.7 ± 4.0%	94.6 ± 3.9%	97.1 ± 4.8%	97.9 ± 4.2%

Table 4. The results on full dataset

Model	Origin features	PCA	ReliefF	MI
LR	67.1 ± 3.2%	44.9 ± 3.2%	60.5 ± 2.9%	61.2 ± 3.0%
kNN	73.6 ± 3.5%	69.1 ± 3.0%	74.6 ± 3.3%	74.3 ± 2.8%
SVM	69.9 ± 3.4%	68.8 ± 3.0%	61.8 ± 2.5%	64.4 ± 2.4%
DT	67.7 ± 5.4%	64.8 ± 2.7%	67.6 ± 3.3%	67.3 ± 3.0%
RF	78.5 ± 3.8%	72.4 ± 2.6%	76.7 ± 3.5%	76.6 ± 2.7%
GB	78.0 ± 2.1%	71.8 ± 1.9%	76.6 ± 2.8%	77.1 ± 3.6%

Table 5. The results on full preprocessed dataset

Model	Origin features	PCA	ReliefF	MI
LR	93.4 ± 2.1%	89.6 ± 2.2%	89.8 ± 2.0%	91.5 ± 1.5%
kNN	92.6 ± 2.0%	91.8 ± 1.5%	92.8 ± 1.8%	88.2 ± 2.5%
SVM	93.9 ± 2.6%	95.2 ± 1.7%	91.2 ± 1.4%	91.7 ± 2.4%
DT	71.9 ± 2.7%	77.5 ± 3.1%	72.6 ± 4.1%	64.8 ± 4.1%
RF	91.9 ± 1.8%	87.9 ± 2.6%	91.5 ± 2.0%	86.2 ± 2.7%
GB	93.3 ± 2.2%	90.2 ± 2.3%	92.9 ± 1.8%	89.7 ± 2.4%

One can see from the tables that models trained on the partial 5-classes dataset achieve much better accuracies than on the full dataset. Models trained on preprocessed datasets are significantly better than models trained on non-preprocessed datasets in terms of accuracy. Thus, preprocessing is one of the most important steps of the approach.

4 Discussion and Conclusion

In this paper we made an attempt to use machine learning to solve the problem of automated pollen grains images recognition. This is a very important problem due to the allergy and asthma management, the key cause of these diseases is pollen. To prevent allergy and asthma symptoms it is necessary to know the concentration of allergenic plants pollen in the air in real time. Existing pollen counting stations cannot provide rapid enough information because of manual processing. To automatize the recognition of pollen species we processed its images from optical microscope. We used GIST descriptors as the feature vector and applied several dimension reduction methods (PCA, MI, ReliefF). This approach gave 98.3% of maximum accuracy on the partial preprocessed dataset, which contains only 5 pollen species. The best classification model is SVM with a polynomial kernel.

That is a new approach relating to this problem, because other authors mostly used specific-purpose features focused on pollen grains nature. Usage

of GIST allows to generalize our solution minimizing the accuracy loss. GIST descriptors are a kind of universal features.

We studied four versions of the dataset to see if pollen grains shape strictly assigns GIST values and to compare preprocessed and initial dataset GIST results.

We found out that the GIST-based approach works much better with the preprocessed dataset, which contains only one pollen grain per image.

We used three dimension reduction techniques and compared their results pairwise with machine learning models.

In future research we will make an attempt to use a convolutional neural network, which is a very promising technique [21], never used by other researchers within this problem. Also we plan to improve pollen the extraction stage, especially in order to resolve the issue of clumped pollen.

The final goal of this research is to develop a program for pollen recognition and bring it to the state of a real-time system, which will cut the cost on pollen operations in half.

Acknowledgments. Authors thank Andrey Filchenkov and Daniil Chivilikhin for suggestions and useful comments. This work was financially supported by the Government of Russian Federation, Grant 074-U01.

References

1. Allen, G.: An automated pollen recognition system. Masters thesis, Institute of information Sciences and Technology, Massey University (2006)
2. Bishop, C.M.: Pattern Recognition and Machine Learning, 1st edn. Springer, New York (2006)
3. Boucher, A., Hidalgo, P.J., Thonnat, M., Belmonte, J., Galan, C., Bonton, P., Tomczak, R.: Development of a semi-automatic system for pollen recognition. *Aerobiologia* **18**(3–4), 195–201 (2002). <http://dx.doi.org/10.1023/A:1021322813565>
4. Chen, C., Hendriks, E.A., Duin, R.P., Reiber, J.H.C., Hiemstra, P.S., Deweger, L.A., Stoel, B.C.: Feasibility study on automated recognition of allergenic pollen: grass, birch and mugwort. *Aerobiologia* **22**(4), 275–284 (2006). <http://dx.doi.org/10.1007/s10453-006-9040-0>
5. Chica, M.: Authentication of bee pollen grains in bright-field microscopy by combining one-class classification techniques and image processing. *Microsc. Res. Tech.* **75**, 1475–1485 (2012). <http://dx.doi.org/10.1016/j.jfoodeng.2012.03.028>
6. Chudyk, C., Castaneda, H., Leger, R., Yahiaoui, I., Boochs, F.: Development of an automatic pollen classification system using shape, texture and aperture features. In: Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB, pp. 65–74 (2015)
7. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011)
8. Computer graphics laboratory courses. https://courses.graphics.cs.msu.ru/pluginfile.php/81/mod_resource/content/1/cv2013.09_cbir.pdf

9. del Pozo-Baños, M., Ticay-Rivas, J.R., Alonso, J.B., Travieso, C.M.: Features extraction techniques for pollen grain classification. *Neurocomputing* **150**, 377–391 (2015). <http://dx.doi.org/10.1016/j.neucom.2014.05.085>
10. Deng, L., Li, X.: Machine learning paradigms for speech recognition: an overview. *IEEE Trans. Audio Speech Lang. Process.* **21**(5), 1060–1089 (2013). <http://dx.doi.org/10.1109/TASL.2013.2244083>
11. European federation of asthma report. <http://www.efanet.org/air-quality/pollen>
12. Flenley, J.R.: The problem of pollen recognition, problems of picture Interpretation. In: CSIRO Workshop, pp. 141–145 (1968)
13. Friedman, H.J. Greedy Function Approximation: A Gradient Boosting Machine. IMS Reitz Lecture (1999)
14. Guggenberger, A.: Another Introduction to Support Vector Machines (2008). <https://scribd.com/document/153294663/Another-Introduction-Svm>
15. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*, 2nd edn. Springer, New York (2009). 533 pages
16. Hinton, G., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Sig. Process. Mag.* **29**(6), 82–97 (2012). <http://dx.doi.org/10.1109/MSP.2012.2205597>
17. Holt, K.A., Bennett, K.D.: Principles and methods for automated palynology. *New Phytol.* **203**(3), 735–742 (2014). <http://dx.doi.org/10.1111/nph.12848>
18. Hu, M.K.: Visual pattern recognition by moment invariants. *IRE Trans. Inf. Theor.* **8**(2), 179–187 (1962). <http://dx.doi.org/10.1109/TIT.1962.1057692>
19. International ragweed day press release. <http://www.pollens.fr/docs/CP-IRD-2015.pdf>
20. Kira, K., Rendell, L.: A practical approach to feature selection. In: *Proceedings of the 9th International Conference on Machine Learning*, pp. 249–256 (1992)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105 (2012)
22. Manning, C., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)
23. Marcos, J.V., Nava, R., Cristobal, G., Redondo, R., Escalante-Ramrez, B., Bueno, G., Dèniz, O., Gonzalez-Porto, A., Pardo, C., Chung, F., Rodríguez, T.: Automated pollen identification using microscopic imaging and texture analysis. *Micron* **68**, 36–46 (2015). <http://dx.doi.org/10.1016/j.micron.2014.09.002>
24. Minayeva, N.V., Novoselova, L.V.: Pollen monitoring in Perm Krai (Russia) experience of 6 years. *Acta Agrobotanica* **68**(4), 343–348 (2015). <http://dx.doi.org/10.5586/aa.2015.042>
25. Mitchell, T.M.: *Machine Learning*. McGraw-Hill Science/Engineering/Math, Boston (1997)
26. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **42**(3), 145–175 (2001). <http://dx.doi.org/10.1023/A:1011139631724>
27. Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M.: Automatic and online pollen monitoring. *Int. Arch. Allergy Immunol.* **167**, 158–166 (2015). <http://dx.doi.org/10.1159/000436968>
28. Projects home page. <http://www-sop.inria.fr/orion/ASTHMA/asthma/asthma.html>

29. Redondo, R., Bueno, G., Chung, F., Nava, R., Marcos, J.V., Cristóbal, G., Rodríguez, T., Gonzalez-Porto, A., Pardo, C., Déniz, O., Escalante-Ramírez, B.: Pollen segmentation and feature evaluation for automatic classification in bright-field microscopy. *Comput. Electron. Agric.* **110**, 56–69 (2015). <http://dx.doi.org/10.1016/j.compag.2014.09.020>
30. Riley, K.C., Woodarda, J.P., Hwanga, G.M., Punyasenac, S.W.: Progress towards establishing collection standards for semi-automated pollen classification in forensic geohistorical location applications. *Rev. Palaeobot. Palynol.* **221**, 117–127 (2015). <http://dx.doi.org/10.1016/j.revpalbo.2015.06.005>
31. Ronneberger, O., Burkhardt, H., Schultz, E.: General-purpose object recognition in 3D volume data sets using gray-scale invariants - classification of airborne pollen-grains recorded with a confocal laser scanning microscope. In: Proceedings of the International Conference on Pattern Recognition, vol. 2, pp. 290–295 (2002). <http://dx.doi.org/10.1109/ICPR.2002.1048297>
32. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 430–443. Springer, Heidelberg (2006). doi:[10.1007/11744023_34](https://doi.org/10.1007/11744023_34)
33. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic n-grams as machine learning features for natural language processing. *Expert Syst. Appl.* **41**(3), 853–860 (2014). <http://dx.doi.org/10.1016/j.eswa.2013.08.015>
34. Yang, Y., Pedersen, J.O.: A comparative study on dimension reduction in text categorization. In: Proceedings of the 14th International Conference on Machine Learning, pp. 412–420 (1997)