# Feature Selection for Big Visual Data: Overview and Challenges

Verónica Bolón-Canedo[1], Beatriz Remeseiro[2], and Brais Cancela[1]

[1] Department of Computer Science, Universidade da Coruña (Spain)
vbolon@udc.es, brais.cancela@udc.es
[2] Department of Computer Science, Universidad de Oviedo (Spain)
bremeseiro@uniovi.es

**Abstract.** The unprecedented amount of visual data that is available nowadays has created new research opportunities and challenges in the areas of computer vision and machine learning. When dealing with large scale datasets, with a huge number of samples and features, the use of feature selection plays an important role for dimensionality reduction whilst allowing model interpretation, data understanding and knowledge extraction. This manuscript is focused on feature selection as applied to big visual data, including both traditional and deep approaches, and tries to give an overview of the cutting-edge techniques to deal with large-scale vision problems and identify technical challenges in the field.

**Keywords:** feature selection, visual information, big data, deep learning

## 1 Introduction

The amount of visual data is exponentially growing day by day, mainly because the increasing availability of cameras of all types [10]. On the one hand, people daily acquire, transmit and share both images and videos through the Internet and different social networks. On the other hand, huge amounts of visual information is being stored and processed by different agents in the industrial world with different purposes, such as automatic inspection in manufacturing applications [25] or detecting events in visual surveillance systems [14].

This huge amount of visual data has led to new problems that pose a challenge for machine learning and computer vision researchers. Image datasets have grown not only in the number of examples, but also in the number of features that describe them. In this situation, it might be reasonable to think that having more features would give us more information and thus better results. However, this is not happening because of the *curse of dimensionality*, a colorful term coined by Richard Bellman back in 1957 to describe the difficulty of optimization by exhaustive enumeration on product spaces [1]. It refers to the different phenomena that appear when analyzing high-dimensional datasets (with hundreds or thousands of features) that do not occur in low-dimensional settings.

A possible solution to this situation is the use of dimensionality reduction techniques, aiming at reducing the number of input features for a given problem.

In particular, this manuscript is focused on the use of feature selection methods in the area of big visual data. Note that traditional approaches of computer vision can be seen as two-step methods in which a set of image properties is first computed to obtain a descriptor, i.e. a feature vector that then feed machine learning algorithms that may include feature selection methods and/or learning models. But when dealing with big visual data, it is quite common to use deep learning algorithms due to their computational power and generalization ability [19]. As opposed to traditional methods, deep learning techniques can be seen as models that perform feature computation and inductive learning as part of the same process. This paper addresses both approaches, traditional and deep, and presents an overview of different techniques as well as the main challenges that need to be faced in the near future.

The rest of this manuscript is structured as follows: Section 2 includes an overview of feature selection applied to large-scale computer vision problems, Section 3 deals with deep learning approaches for feature selection, Section 4 includes some technical challenges, and Section 5 closes with the conclusions.

## 2  Feature Selection

As stated in the introduction, dimensionality reduction techniques are used to reduce the number of input features. These dimensionality reduction techniques usually come in two flavors: *feature selection* and *feature extraction*. On the one hand, feature extraction techniques achieve dimensionality reduction by *combining* the original features (i.e. the features extracted from the images that are fed as input to the FS algorithms). In this manner, they are able to generate a set of *new* features, which is usually more compact and of stronger discriminating power. It is the typical choice in applications such as image analysis, signal processing, or information retrieval. On the other hand, feature selection achieves dimensionality reduction by *removing* the irrelevant and redundant features. Due to the fact that feature selection maintains the *original* features, it is especially useful for applications where the original features are important for model interpreting and knowledge extraction.

As mentioned above, feature extraction is the preferred approach when dealing with image analysis, and therefore there are plenty of works analyzing the application of these methods to this field [33, 24, 11, 16]. However, although not so common, feature selection has also been applied to image analysis, and it is expected to be more important in the future, now that the interpretability of the results is gaining importance. Therefore, the remaining of this paper will be focused on feature selection (FS).

FS methods are typically divided into three major approaches according to the relationship between a FS algorithm and the inductive learning method used to infer a model [12]: filters, which are independent of the induction model; and wrappers and embedded methods, which involve a learning algorithm to determine the useful features. When dealing with Big Data, the preferred approach

is to use filters, since they are advantageous in terms of computational cost. For more details, refer to the specialized books [12, 3].

DNA microarray images are a typical application of classical feature selection [5]. This type of data consists in measuring the simultaneous expression of thousands of genes, which then can be used as inputs to large-scale data analysis. Typically, these datasets have thousands of features and very small samples (often less than 100), so feature selection plays an extremely important role and there are thousands of works in the literature acknowledging its effectiveness in this domain. The interested reader may find some works that review the FS methods used most for microarray data. Saeys et al. [28] provided a basic taxonomy of classical FS techniques and discuss their use in a number of bioinformatics applications. Lazar et al. [18] presented a survey focusing on filter methods in a unified framework. During the last decade, in addition to the application of state-of-the-art methods, an important number of new filter approaches have been proposed and applied to microarray data, which are reviewed by Bolón-Canedo et al. [5].

Many other works can be found in the literature that apply feature selection to different computer vision problems with high-dimensional datasets. Jia et al. [15] analyzed the effect of receptive field designs on image classification accuracy, and proposed a learning algorithm based on incremental feature selection that outperforms the state-of-art on the CIFAR-10 dataset. Tan et al. [31] proposed a novel adaptive scaling scheme for high-dimensional FS on Big Data, and demonstrated the effectiveness of the method by performing experimentation on different synthetic and real-world large-scale datasets, including image datasets such as MNIST. Zhao et al. [34] presented a new framework to select relevant features in multi-modal, often high-dimensional datasets by applying deep neural networks and FS with LASSO, and carried out an experimental study on three image classification datasets. Cao et al. [7] proposed a measure to evaluate the relevance of feature groups in image classification by employing the well-known feature selection filter *minimum Redundancy Maximum Relevance* (mRMR), and demonstrated the importance of using well-designed features after applying this measure to different image datasets and several image descriptors.

## 3   Deep Feature Selection

Deep learning algorithms are normally used to *extract* relevant features. By removing the last layer, one can take the final layer as a feature vector (these are the so-called *deep features*). As it is mentioned, this is a feature extraction procedure which is not the focus of this manuscript. However, as previously explained, feature selection methods maintain the original features, so they can be used for model interpretability and knowledge extraction. Assuming this idea, there are some deep learning techniques that can be included as *hybrid* feature selection methods, since they take advantage of feature extraction properties to infer which image pixels are relevant to the task. Note that these methods

are not as fast as classic feature selection approaches (see Section 2), since an intermediate step is required to obtain the relevant information.

All these algorithms try to obtain the *saliency* features, and deep learning techniques were proven to be a powerful tool to detect them. By definition, *saliency* is the characteristic that stands out compared with its neighborhood. In our study field, it means the detection of relevant features within an image, discarding those that are not necessary (e.g. foreground vs. background). This approach is somehow different from classic feature selection since here the idea is to detect those features that help the model to trigger any given output, rather than just taking a fixed subset. Although no dimensionality reduction can be performed if spatial information is used (for instance, in Convolutional Neural Networks), feature selection can be used by putting irrelevant features to zero, and thus helping the model to avoid overfitting.

Training algorithms can be classified in two different approaches: weakly-supervised and fully-supervised. In weakly-supervised algorithms, the classification task indicates what kind of information appears in the image, but not where it is located. Thus, only by visualizing the neural network node activation we can infer where the object of interest is located. On the contrary, fully-supervised algorithms requires a segmentation image to compare against it.

Simonyan et al. [30] use a two-step algorithm to obtain a saliency map. First, a deep Convolutional Neural Network (CNN) for class classification is trained. Then, for every image, it recognizes the pixels that are contributing the most to select the winner class. Two different approaches were created to deal with this information, either to select a smaller ROI within the image or to perform an object segmentation, discarding the information that is considered as background. It is very interesting because there is no information provided to the CNN about where the object is placed within the scene, and yet the network highest activations are usually correctly placed [27]. However, this effect can also be used to break the system, as it was demonstrated that incoherent images can be created to completely fool a CNN [26].

This idea of evaluating the activation weights was also used in [20], but introducing a pool of different CNNs. Wei et al. [32] extended a similar version, introducing both deep CNN and semantic segmentation. Zhou et al. [36] introduce the concept of *class activation maps*, a weight-based heat map associated to each image. Different final layer possibilities were tested in [22], selecting Global Average Pooling as the most suitable one. Li and Yu [21] used a Conditional Random Field (CRF) as final step in order to improve the spatial coherence. Similar idea was addressed in [35], but combining a paired-CNN (two different CNNs with shared weights in some layers) to introduce both global and local context information, respectively. Pre-trained CNNs were also proven to be valid architectures to provide saliency information [13].

Despite the detection of salient regions, there is also a field of study that addresses the issue of detecting objects in the scene. It is often referred as *salient object detection*. We are not interested in the segmentation part, since it is a classification problem, far from the feature selection methods we are address-

ing in this paper. However, some remarkable salient techniques were created, and we want to mention them. A comprehensive analysis and study of existing works was reported in [6], where both a review and a benchmark study were conducted. Cheng et al. [9] used the saliency map to successfully detect the foreground objects within an image. Li et al. [23] uses a fully-supervised regularized nonlinear regression model, developing a multi-task learning scheme for inferring the correlations between saliency detection and semantic image segmentation.

## 4 Challenges

Ongoing advances in computer-based technologies have enabled researchers to collect data at an increasingly fast pace. It is more frequent than ever to have to deal with high-dimensional data, so feature selection becomes an imperative pre-processing step [4]. However, this advent of Big Data has brought an important number of challenges, both in traditional and deep learning approaches, and some of them are discussed in the following.

- **Real-time processing.** Data is being collected at an unprecedented fast pace and, correspondingly, needs to be processed rapidly. Social media networks and portable devices dominate our day-to-day, and they require sophisticated methods capable of dealing with vast amounts of data in real time, e.g. for video/image retrieval. Although online learning is a quite popular field among researchers, online feature selection has not received the same amount of attention. In this new Big Data scenario, online feature selection methods should be capable of (i) modifying the selected subset of features as new training samples arrive, and (ii) being executed in a dynamic feature space that would initially be empty but would add features as new information arrived. Online deep learning should be also a focus of attention for researchers, due to its limited progress in recent years. In this sense, the strategy of stochastic gradient descent and the update of parameters on a mini-batch basis used for training deep learning algorithms may be adapted for online learning [8].
- **Feature cost.** Typically, feature selection methods only focus on detecting the relevant features, but there are some situations in which these features have an associated cost which should not be ignored. For example, the cost for computing different image features may be not homogeneous [2], so the fact that the computational cost of extracting each feature varies implies different computational times. Time complexity is crucial with the advent of Big Data, especially in real-time processing. Although the issue of reducing costs in feature selection has received some attention in the last few years, we still definitely need new feature selection methods that can match the accuracy of state-of-the-art algorithms while reducing computational cost.
- **Interpretability.** As mentioned before, feature extraction is the preferred approach to reduce image dimensionality. However, these techniques have the limitation that the features being selected are transformations of the

original ones. Thus, when model interpretability is important, FS is necessary to discover the good features of a model. And not only to deal with images, but also on other types of big databases it is necessary to develop user-friendly visualization tools to enhance interpretability. There is an important necessity of more interactive model visualizations where (i) users can change input parameters to better interact with the model and visualize future scenarios, and (ii) users can iterate through different feature subsets rather than be tied to a specific subset chosen by an algorithm.

– **Acceptability.** In the context of interpretability previously mentioned, deep learning models are often seen as *black-boxes* difficult to interpret, which implies a problem of user acceptance in critical sectors such as health or robotics. Saliency maps (see Section 3) may be useful to understand output models by providing information about the pixels that contribute to generate this output. Other techniques have been presented to generate visual explanations from deep learning models [29], although their full deployment has not taken place yet. Thus, the benefits of using visual analytics tools need to be communicated to the potential users in order to overcome any possible barrier [17].

## 5  Conclusions

The growing size of visual datasets, composed by not only a great amount of samples (images and/or videos) but also a huge number of features, makes indispensable the use of dimensionality reduction techniques. These methods can be useful to reduce the number of input features and alleviate the computational burden with no degradation in performance. As opposed to feature extraction, feature selection is also helpful in model interpretation and knowledge extraction.

Traditional approaches of feature selection have been successfully applied to different problems, including DNA microarray analysis and image classification. When dealing with Big Data, particularly in computer vision problems, the use of deep learning seems to be inevitable due to its computational power. However, using deep models involves sacrificing interpretation to gain abstraction and integration [29]. In this context, saliency features can be computed by deep learning techniques to detect the relevant regions within an image.

In conclusion, both traditional and deep feature selection approaches have demonstrated to be useful in different large-scale problems of computer vision. However, some technical challenges need to be addressed to increase model interpretability and user acceptability, and to allow real-time processing.

## Acknowledgments

# References

1. Bellman, R.: Dynamic Programming. Princeton University Press (1957)
2. Bolón-Canedo, V., Remeseiro, B., Sánchez-Maroño, N., Alonso-Betanzos, A.: mC-ReliefF: An Extension of ReliefF for Cost-Based Feature Selection. In: International Conference on Agents and Artificial Intelligence. vol. 1, pp. 42–51 (2014)
3. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Feature selection for high-dimensional data. Springer (2015)
4. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Systems 86, 33–45 (2015)
5. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F.: A review of microarray datasets and applied feature selection methods. Information Sciences 282, 111–135 (2014)
6. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Transactions on Image Processing 24(12), 5706–5722 (2015)
7. Cao, Z., Principe, J.C., Ouyang, B.: Group feature selection in image classification with multiple kernel learning. In: International Joint Conference on Neural Networks. pp. 1–5 (2015)
8. Chen, X.W., Lin, X.: Big Data Deep Learning: Challenges and Perspectives. IEEE Access 2, 514–525 (2014)
9. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3), 569–582 (2015)
10. Fang, Z., Hwang, J.N., Huo, X., Lee, H.J., Denzler, J.: Emergent Techniques and Applications for Big Visual Data. International Journal of Digital Multimedia Broadcasting 2017 (2017)
11. Guo, G., Fu, Y., Dyer, C.R., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Transactions on Image Processing 17(7), 1178–1188 (2008)
12. Guyon, I.: Feature extraction: foundations and applications, vol. 207. Springer (2006)
13. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. In: International Conference on Machine Learning. pp. 597–606 (2015)
14. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 34(3), 334–352 (2004)
15. Jia, Y., Huang, C., Darrell, T.: Beyond spatial pyramids: Receptive field learning for pooled image features. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3370–3377 (2012)
16. Juan, L., Gwun, O.: A comparison of SIFT, PCA-SIFT and SURF. International Journal of Image Processing 3(4), 143–152 (2009)
17. Keim, D.A., Mansmann, F., Schneidewind, J., Ziegler, H.: Challenges in visual data analysis. In: International Conference on Information Visualization. pp. 9–16 (2006)
18. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., Nowe, A.: A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(4), 1106–1119 (2012)

19. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015)
20. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5455–5463 (2015)
21. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 478–487 (2016)
22. Li, G., Yu, Y.: Visual saliency detection based on multiscale deep CNN features. IEEE Transactions on Image Processing 25(11), 5012–5024 (2016)
23. Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J.: DeepSaliency: Multi-task deep neural network model for salient object detection. IEEE Transactions on Image Processing 25(8), 3919–3930 (2016)
24. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)
25. Malamas, E.N., Petrakis, E.G., Zervakis, M., Petit, L., Legat, J.D.: A survey on industrial vision systems, applications and tools. Image and Vision Computing 21(2), 171–188 (2003)
26. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 427–436 (2015)
27. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 685–694 (2015)
28. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 618–626 (2017)
30. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
31. Tan, M., Tsang, I.W., Wang, L.: Towards Ultrahigh Dimensional Feature Selection for Big Data. Journal of Machine Learning Research 15, 1371–1429 (2014)
32. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: STC: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(11), 2314–2320 (2017)
33. Weinberger, K.Q., Saul, L.K.: Unsupervised learning of image manifolds by semidefinite programming. International Journal of Computer Vision 70(1), 77–90 (2006)
34. Zhao, L., Hu, Q., Wang, W.: Heterogeneous Feature Selection With Multi-Modal Deep Neural Networks and Sparse Group LASSO. IEEE Transactions on Multimedia 17(11), 1936–1948 (2015)
35. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1265–1274 (2015)
36. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)