Evaluation of Deep Image Descriptors for Texture Retrieval

Bojana Gajic[1], Eduard Vazquez[2] and Ramon Baldrich[1]
[1]*Computer Vision Center, Universitat Autonoma de Barcelona, Bellaterra, Spain*
[2]*Cortexica Vision Systems, Suite 704 Capital Tower, 91 Waterloo Road, SE1 8RT, London, U.K.*
bgajic@cvc.uab.es, eduard.vazquez@cortexica.com, ramon@cvc.uab.es

Keywords: Texture Representation, Texture Retrieval, Convolutional Neural Networks, Psychophysical Evaluation.

Abstract: The increasing complexity learnt in the layers of a Convolutional Neural Network has proven to be of great help for the task of classification. The topic has received great attention in recently published literature. Nonetheless, just a handful of works study low-level representations, commonly associated with lower layers. In this paper, we explore recent findings which conclude, counterintuitively, the last layer of the VGG convolutional network is the best to describe a low-level property such as texture. To shed some light on this issue, we are proposing a psychophysical experiment to evaluate the adequacy of different layers of the VGG network for texture retrieval. Results obtained suggest that, whereas the last convolutional layer is a good choice for a specific task of classification, it might not be the best choice as a texture descriptor, showing a very poor performance on texture retrieval. Intermediate layers show the best performance, showing a good combination of basic filters, as in the primary visual cortex, and also a degree of higher level information to describe more complex textures.

1 INTRODUCTION

Convolutional Neural Networks (ConvNets) have revolutionised the field of Computer Vision in the last few years. Becoming a breakthrough in the accuracy achieved since 2009 and most remarkably since imageNet 2012 (Russakovsky et al., 2015) with AlexNet (Krizhevsky et al., 2012), ConvNets have been constantly evolving with new architectures as VGG (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015) or Residual Networks (He et al., 2015) among others. Representations learnt by ConvNets have overtaken classic hand-crafted descriptors such as scale-invariant feature transform (SIFT) (Lowe, 2004) and all its variations or speeded up robust features (SURF) (Bay et al., 2006). The main advantage is that ConvNets learn non-linear transformations and adapt, whereas hand-crafted descriptors were designed a priori thinking on image properties that could in principle generalise to any object or class.

Whereas the topic of classification is one of the main focuses of attention in existing literature along with localisation (Ren et al., 2015), segmentation (Zheng et al., 2015) or tracking (Wang and Yeung, 2013), low-level representations as obtained by ConvNets have attracted limited attention. The question is whether these deep representations are equally valid to describe a low-level property such as texture, typically represented with hand-crafted descriptors as Local Binary Pattern methods (Ojala et al., 2002), Gabor and wavelet based approaches (Wu et al., 2000) among many others (Mikolajczyk and Schmid, 2005).

The nature of the representations learnt in ConvNets gradually varies from the lower to the last layer. It goes from simple representations to more complex ones in subsequent layers, gradually increasing its semantical meaning (Zeiler and Fergus, 2014). In other words, we see a gradual increase from low-level to high-level representations. Discriminative models aim to find the borders between classes, not to model joint probability distributions as a generative model aims to do (Jordan, 2002). This semantical meaning relates much closely to the classes to be recognised, better separating one class to the other in the feature space, being one of the factors explaining the great success of ConvNets in the aforementioned tasks.

In this context, where semantics and classes are closely related to the representations learnt we want to address one question: how good are the low-level representations learnt? Some of the weights learnt in lower layers are similar to common approaches used to describe textures such as Gabor filters and other low-level representations (Wu et al., 2000). With this knowledge, the assumption would be that the lower

Gajic B., Vazquez E. and Baldrich R.
Evaluation of Deep Image Descriptors for Texture Retrieval.
DOI: 10.5220/0006129302510257
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 251-257
ISBN: 978-989-758-226-4

layers should be better to represent textures than the latest ones. This subject has received little attention in recently published literature, although some works can be found exploring this topic (Gan et al., 2015), (Cimpoi et al., 2014), (Cimpoi et al., 2016). For the problem that this paper explores, results obtained in the second and third publications are of great interest, since the authors defined as a *gold standard texture representation* a descriptor derived from the fifth convolutional layer of a VGG network (Simonyan and Zisserman, 2014). That is, that opposed to what intuition might say, the initial layers seem to be worse for texture. The evaluation of this representations is tested in the paper by running a classification experiment. Nonetheless, textures are a perceptual property, therefore less related with semantics. Can be then concluded, using a classification experiment, that the best descriptor for textures is derived from a high-level representation with a remarkable semantical component?

In this paper, we aim to shed some light on this problem by proposing an experiment more closely related with perception as it is image retrieval. The difference with other works focused on retrieval (Babenko et al., 2014) and the present work, is that here we perform a psychophysical experiment to evaluate texture retrieval in which we asses the capability to describe textures by different layers.

This paper is organised as follow. In section 2 we explain the motivation behind running a psychophysical experiment based on image retrieval as opposed to a classification experiment. Subsequently, in section 3 the psychophysical experiment is defined and results obtained are presented in section 4. Finally, conclusions are drawn in section 5.

2 CLASSIFICATION VERSUS RETRIEVAL FOR LOW-LEVEL DESCRIPTORS

The fact that the best descriptor for texture is derived from a layer with strong semantical meaning, therefore, goes against the initial intuition after observing the weights learnt in ConvNets as clearly shown in (Zeiler and Fergus, 2014). Texture is a low-level perceptual property. We know that we perceive orientations and scales in a particular way in V1 (Rust et al., 2005) and that it explains the way we process images at a low level. In this early stage, semantics do not play an important role. Actually, texture perception is not as much related to semantic meaning, but a continuous space where just some areas can be named. Can then a strongly semantical representation be the best one to describe textures?

There is no doubt that the representation found in (Cimpoi et al., 2016) works well to classify a specific set of textures. However, it is not clear that such representation is the best way to describe a perceptual, low-level world. Image retrieval does not include classes, and therefore is a different task than that of classifying semantic sets, being much more suitable to evaluate a perceptual descriptor.

3 EXPERIMENTAL SETUP

The experiment presented in this article aims to verify the assessment that the representations derived from the latest deep convolutional layers of a ConvNet are in fact the most suitable for textures. The authors in (Cimpoi et al., 2016) affirm that most of the performance gain is realised in the very last few layers. To avoid the introduction of other variables and different settings, we are limiting our psychophysical experiment following the procedures and using the same data as used in the referred work. The difference, as explained in section 1, is that we evaluate the quality of the descriptors in the context of image retrieval.

3.1 Texture Descriptors

In order to get the best representations of texture on images, we follow (Cimpoi et al., 2016) and use deep convolutional features as local image descriptors. These features are extracted from each of the outputs of convolutional layers of ConvNets pretrained on ImageNet ILSVRC data (Deng et al., 2009). We considered both VGG-M and VGG-VD network architectures, as proposed in the referred work. In this experiment, we want to evaluate the suitability of different representations for texture retrieval. The number of layers in VGG-M suffices to this aim. Conclusions extracted with current experiment can be extended to deeper architectures, such as VGG-VD, since in terms of representations for each layer, it follows exactly the same principles.

As detailed in (Cimpoi et al., 2016), representations extracted from each convolutional unit are pooled into a Fisher vector (Perronnin and Dance, 2007) representation with 64 Gaussian components. The dimensionality of the representation space range from 12k to 65k, depending on the layer. As the number of dimensions of these descriptors is too high, they are supposed to be very redundant. Consequently, the descriptors are compressed by principal

component analysis (PCA) to size of 4096 dimensions, as proposed in (Cimpoi et al., 2016). After dimensionality reduction, the descriptors are L2 normalized.

To summarise, in the experiment we compare five different descriptors, one per convolutional layer. Our aim is to verify if the descriptors that obtain better results on classification provide better retrieval results.

3.2 Dataset

In the process of finding the best dataset for texture retrieval task, we considered currently available texture datasets. Many of them were concentrated on material (FMD (Sharan et al., 2013), KTH-TIPS2 (Caputo et al., 2005), CUReT (Dana et al., 1999)). The only dataset that contains texture images is Describable Textures Dataset (DTD) (Cimpoi et al., 2014). This dataset is created in the wild and it contains 5640 texture images that are annotated with adjectives selected in a vocabulary of 47 English words. These words are chosen from a larger set of 98 words that people commonly use to describe textures, proposed in (Bhushan et al., 1997). This work is mainly focused on the cognitive aspects of texture perception, including perceptual similarity and the identification of directions of perceptual texture variability. The words that describe surface shape, or do not give information about visual aspects are removed from the set (messy, corrugated etc.) and some words with similar meanings are merged into one (coiled, spiralled and corkscrewed). Examples of images from different classes of DTD are shown in Fig 1.

3.3 Layout of the Experiment

To perform the psychophysical experiment, one image from each of the 47 texture categories in DTD dataset was randomly selected. We retrieved the 5 most similar images for each descriptor. Top 5 results might contain images that are not from the same class as the query. The test should not evaluate classes of the retrieved images.

Images from the dataset are ranked using the Euclidean distance between the query image and retrieved image representations. Only the first 5 images are included in the test. In order to compare texture descriptors based on feature maps of different convolutional layers, for each query image, we extracted five retrieval responses independently.

Since there is no ground-truth that would enable measuring how good retrieval result are, we created a survey to get subjective responses using an online platform. This way of testing is flexible because it allows people from different places and countries to do the survey. Although the test is performed in uncalibrated conditions, we fixed the necessary conditions to assure validity of the results. The main conditions are colour and size, as these are properties that can change texture perception. Colour change has been tested in different monitors without relevant changes in the answers. Regarding size, which can change texture perception, all subjects are asked to perform the experiment on desktop PC using at least 19” screen.

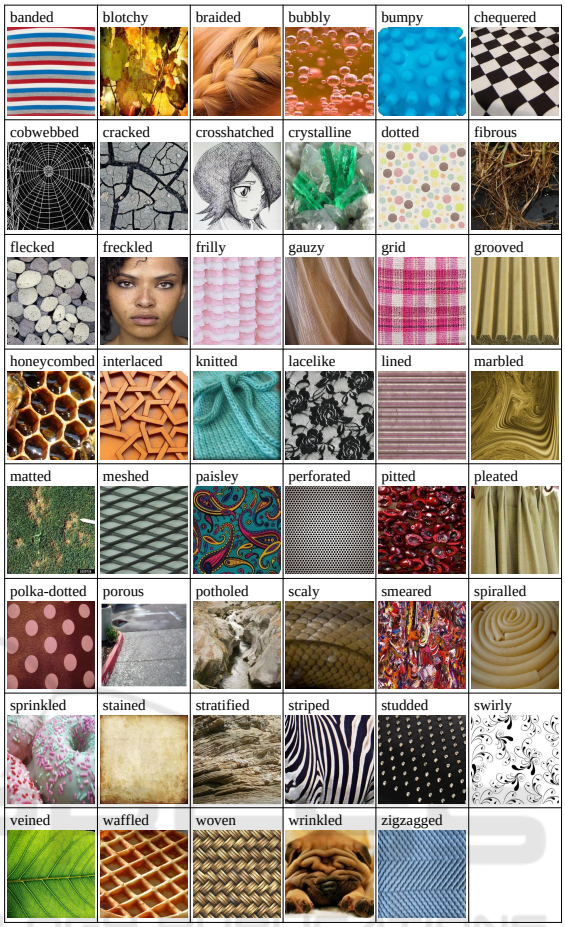

Figure 1: Examples from Describable Texture Dataset.

For each of 47 query images, we created a question that offers five different retrieval responses (one based on each convolutional layer) where every response contains top 5 images. We asked participants to compare these responses and to choose at least one good and one bad result. For easy comparison, results of descriptors based on feature maps from different convolutional layers are presented one below each other. In order to prevent the subjects from finding patterns in results in different questions, rows with responses were randomly shuffled for each query image.

3.4 Specifications of the Survey

The initial instructions to the subjects prior to the test were:

”Imagine that you are searching for an image on internet, where the texture is what you are interested in. In this survey we present some possible solutions, while your task is to rate if our proposed solutions are good or bad. There is not a specific task, just to have some results that reminds the query as much as possible.

The query image is present in the beginning of each row and it is bordered by a red rectangle. Each row presents one possible solution. Each solution contains five ranked images, extracted by a particular ”system”. The closer an image is to the query image (left one), the more similar the system considers it.

You can consider a good answer those rows that you would be comfortable with when looking for the query image. It’s up to you to decide which criteria is more relevant in every single case.

Please, rate at least one retrieval result as good and at least one as bad per each question. If you think that a row is neither good or bad leave the corresponding answer empty.

Thank you!”

This test was verified by four people with agreement.

The survey was fulfilled by 10 people. Half of them were naive subjects, never involved in an image comparison task before. The other half were people linked to the computer vision field, many of them in research. None of the subjects was related to this specific work. 90% of people were between 24 and 35 years old, while the 10% were between 18 and 24. 50% of participants were female. They were people of 7 different nationalities.

3.5 Examples from the Survey

In this section, two examples of questions from the survey and the collected results are presented. As explained in section 3.3, each row of figures 2 and 3 represents a retrieval result based on features from a different convolutional layer. In this case, rows are not shuffled, so the first row is based on the first convolutional layer, second on the second and so on.

For every image, the answers from all subjects are summarised in the number of times a given layer representation was chosen as bad and how many times as good. The results for Figure 2 and 3 are shown in Table 1.

Figure 2 shows an example of texture retrieval where the first convolutional layer was rated as bad by almost all the observers. On the other hand, the retrieval based on the second and third layers were rated as good by all the observers. In this case, we can conclude that the first layer is the worst, the second and third are the best. Depending on the threshold between the number of good and bad votes, representations from the layers four and five could be considered differently. As long as all the subjects who rated the fourth layer gave it the good mark it is probable that this representation is acceptable, while the fifth layer provides descriptor that achieved more bad than good votes, so it is classified as a bad descriptor.

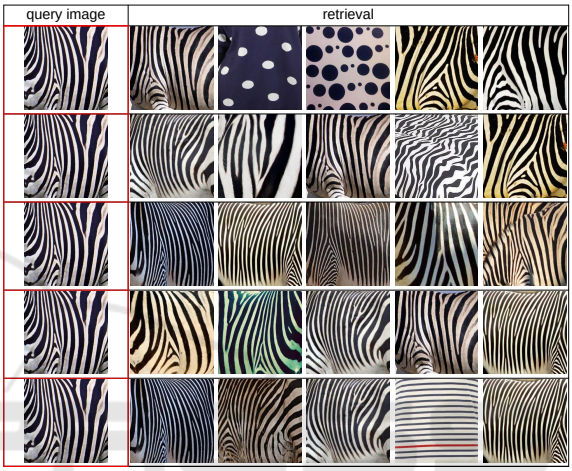

Figure 2: Example of a question from the survey with the best 2^{nd} and 3^{rd} and the worst 1^{st} layer representations.

Figure 3 shows a different behaviour of the texture descriptors. In this case, almost all the subjects agreed that the result of the fourth and fifth convolutional layers are bad. On the other hand, they agreed that the third layer gives good result. Therefore, the third layer can be regarded as the best.

Table 1: Results obtained on the two representative examples.

	Fig 2		Fig 3	
conv. layer	good	bad	good	bad
1	1	9	7	2
2	10	0	6	3
3	10	0	8	1
4	8	0	0	9
5	3	4	1	8

The main difference between the results of these two examples is the result obtained by the descriptors based on the first convolutional layer. The reason for this could be the fact that the first convolutional layer contains low-level features which are good enough in case of simple texture that does not have many details.

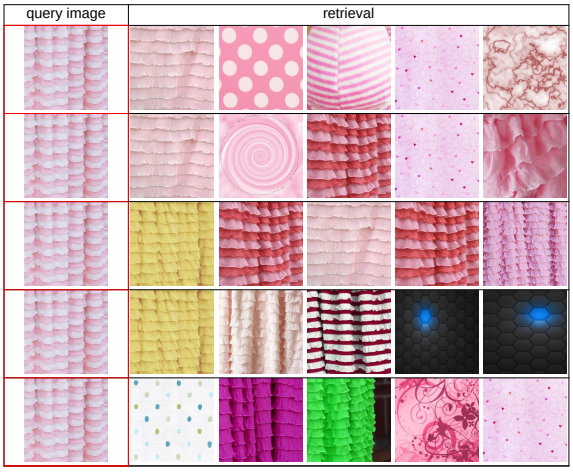

Figure 3: Example of a question from the survey with the best 3^{rd} and bad 4^{th} and 5^{th} layer representations.

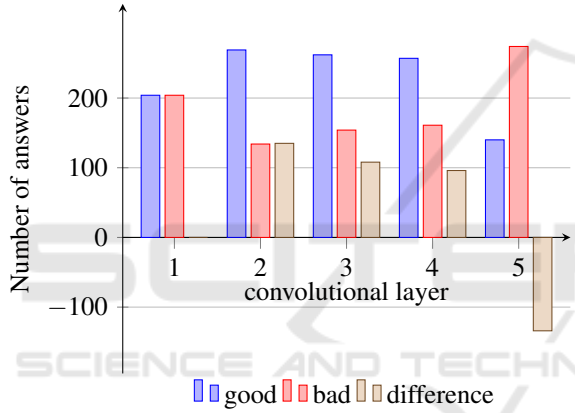

Figure 4: The results are presented for descriptors based on different convolutional layers separately. Number of "good" answers is sumed for all queries and all subjects and it is presented by the blue bars. The same procedure is done for "bad" answers and the results are presented by red bars. The differences between the number of positive and negative votes - yellow bars.

In case of images with complex texture, this representation might be too simplistic.

4 RESULTS

This section summarises and discusses the results obtained from the experiment. Results are aggregated in two different ways, as shown in Fig. 4 and Fig. 5.

Figure 4 shows how many times the representation of the image using each layer is selected as a good or bad result. The yellow bars are the difference between the number of good and bad votes. These results imply that texture is described well by the intermediate layers, while the last layer provides the worst results.

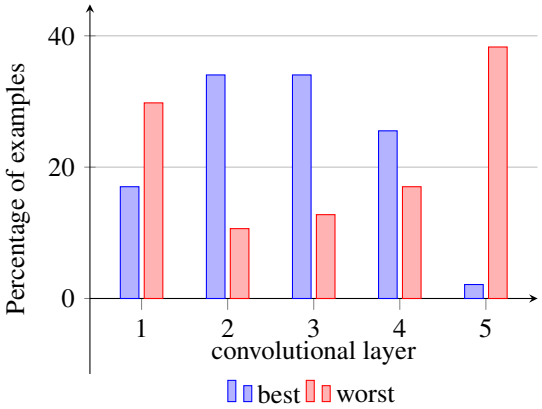

Figure 5: Chart with the results of the survey. For each question from the survey we calculate the difference between the number of "good" and "bad" votes per each descriptor. The biggest difference implies that that layer provides the best results, while the smallest difference says that the result is the worst. In case that two layers have the same difference between the positive and the negative votes, we declare both of them as bests or worsts.

The first layer introduces confusion as the number of good and bad votes is the same. The main reason for the uncertain results could confirm our assumption that texture is low-level property. Texture on some images is very simple, without many details and in these cases, representation based on the first convolutional layer is rated as good (Figure 3). In contrast, images of texture on certain objects apart from having information about texture, contain shape. As shape is known as a high-level property, this information is present in deeper layers.

Fig. 5 shows additional information useful for ranking of the descriptors. In this case, the graphic shows how many times each descriptor appeared as the best and the worst.

For each subject (i), query image (j) and the number of the layer on which the descriptor is based (k) the result of the survey can be good, bad or other. We value good answer by +1, bad by -1, while neutral answer is 0:

$$t(i,j,k) = \begin{cases} 1 & \text{if vote is good} \\ -1 & \text{if vote is bad} \\ 0 & \text{other} \end{cases}$$

For each query image from the survey and for each layer we sum the responses from all the subjects:

$$p(j,k) = \sum_{i=1}^{10} t(i,j,k)$$

The layer with the highest $p(j,k)$ score is rated as the best, while the one with the lowest score is the

worst.

$$b(j) = \arg\max_k (p(j,k))$$

$$w(j) = \arg\min_k (p(j,k))$$

For each layer we counted number of questions in which it appeared as the best and as the worst result.

$$B(k) = \sum_{j=1}^{47} (1, \text{if } k = b(j))$$

$$G(k) = \sum_{j=1}^{47} (1, \text{if } k = w(j))$$

Figure 5 shows the difference between the layers and their ranking. According to the results that are presented, it can be concluded that the descriptors based on the fifth convolutional layer do not describe texture well enough. On the other hand, second and third layers are providing the best descriptors.

The feature maps from the fifth convolutional layer contain information that is not relevant for texture description. This layer is deep and it provides high-level properties, such as shape. Our experiment confirms that shape is not a relevant parameter for texture description so the descriptors from this layer do not contain essential information.

5 CONCLUSIONS AND DISCUSSION

The increasing complexity learnt in the layers of a ConvNet clearly helps the task of classification. The ability to describe low-level features has been associated to lower layers. Texture is a perceptual property just partially related to semantics. It has been associated to the V1, the primary visual cortex, which responds to basic stimuli, in a similar way as lower layers in a ConvNet do. Therefore, recent findings suggesting that the last layer of the VGG network is the best to describe textures seem counter-intuitive. In this paper, we have presented a psychophysical experiment to assert the suitability of different layers of the VGG ConvNet for texture retrieval. We have used the same dataset presented in (Cimpoi et al., 2014), the Describable Textures Dataset, to run the retrieval experiment. The reason underlying is to see whether a good representation for classification is actually a good representation for retrieval, where categories are not present, a task more related to low-level perception.

Results obtained show a great agreement among all the subjects in that the last layer is actually the worst for the task at hand. This result agrees with the common theory regarding textures, namely, that these do not hold a strong semantical meaning. The fifth layer encodes high-level information, therefore being less adequate to describe textures. As shown in Fig.2, this representation does not properly encode basic texture properties such as scale or orientation.

Another interesting result is that the first layer, although performing remarkably better than the fifth, does not seem to be the best choice to represent textures. Interestingly, results suggest that the filters learnt in such layer might actually be too basic to perform good enough in the Describable Textures Dataset. The conclusion from this experiment is that intermediate layers, which encode more complex representations after the nonlinearities learnt from basic filters, are the best to describe textures. These results show that textures do have a degree of higher-level information. Not a surprising finding given the nature of the dataset. Actually, textures do go further than just basic stimuli. Consider for instance the texture 'floral', as can be found on textiles. This texture does contain a degree of high level information and therefore higher layers should better encode such concepts. Henceforth, a trade-off between low and high representations is theoretically needed to represent basic textures and some higher-level representations at the same time.

The adequacy of the Describable Textures Dataset as a sufficient representation to describe textures is an important question not tackled in this paper. If textures are just partially related to semantics, to what extent a dataset strictly derived from semantics can effectively represent textures? This is not a criticism of the dataset since it clearly states that it is related to 'describable' textures. The problematic relates more to the conclusion that a good texture representation is one that can perform a good classification task in such dataset, or, to this effect, to any dataset derived from semantics.

Finally, an important conclusion of this paper is that a good representation for classification does not necessarily have to be a good representation for retrieval. This is due to the nature of representations learnt by a discriminative model, where the main aim is to draw the borders between classes in the feature space. The relations between elements of the same class are not tackled directly on discriminative learning. Such task relates much closely to representation learnt by generative models, where the joint probability is modelled. Consequently, a good direction to find a better representation for texture in the context

of Deep Learning might be to exploit unsupervised approaches.

ACKNOWLEDGEMENTS

This work has been supported by Spanish MINECO project TIN2014-61068-R.

REFERENCES

Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. (2014). Neural codes for image retrieval. In *European Conference on Computer Vision*, pages 584–599. Springer.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

Bhushan, N., Rao, A. R., and Lohse, G. L. (1997). The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images. *Cognitive Science*, 21(2):219–246.

Caputo, B., Hayman, E., and Mallikarjuna, P. (2005). Class-specific material categorisation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1597–1604. IEEE.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613.

Cimpoi, M., Maji, S., Kokkinos, I., and Vedaldi, A. (2016). Deep filter banks for texture recognition, description, and segmentation. *International Journal of Computer Vision*, 118(1):65–94.

Dana, K. J., Van Ginneken, B., Nayar, S. K., and Koenderink, J. J. (1999). Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics (TOG)*, 18(1):1–34.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Gan, Y., Cai, X., Liu, J., and Wang, S. (2015). A texture retrieval scheme based on perceptual features. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 897–900. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Perronnin, F. and Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Rust, N. C., Schwartz, O., Movshon, J. A., and Simoncelli, E. P. (2005). Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–956.

Sharan, L., Liu, C., Rosenholtz, R., and Adelson, E. H. (2013). Recognizing materials using perceptually inspired features. *International journal of computer vision*, 103(3):348–371.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Wang, N. and Yeung, D.-Y. (2013). Learning a deep compact image representation for visual tracking. In *Advances in neural information processing systems*, pages 809–817.

Wu, P., Manjunath, B., Newsam, S., and Shin, H. (2000). A texture descriptor for browsing and similarity retrieval. *Signal Processing: Image Communication*, 16(1):33–43.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. (2015). Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537.