Stacking of SVMs for Classifying Intangible Cultural Heritage Images

Thanh-Nghi Do, The-Phi Pham, Nguyen-Khang Pham, Huu-Hoa Nguyen, Karim Tabia, Salem Benferhat

Stacking of SVMs for classifying intangible cultural heritage images (Preprint version)

Thanh-Nghi Do[1,2], The-Phi Pham[1], Nguyen-Khang Pham[1], Huu-Hoa Nguyen[1]
Karim Tabia[3], Salem Benferhat[3]

[1] College of Information Technology
Can Tho University, 92000-Cantho, Vietnam
[2] UMI UMMISCO 209 (IRD/UPMC)
UPMC, Sorbonne University, Pierre and Marie Curie University - Paris 6, France
[3] CRIL UMR 8188
CRIL CNRS and Artois University, France
dtnghi@cit.ctu.edu.vn

Abstract. Our investigation aims at classifying images of the intangible cultural heritage (ICH) in the Mekong Delta, Vietnam. We collect an images dataset of 17 ICH categories and manually annotate them. The comparative study of the ICH image classification is done by the support vector machines (SVM) and many popular vision approaches including the handcrafted features such as the scale-invariant feature transform (SIFT) and the bag-of-words (BoW) model, the histogram of oriented gradients (HOG), the GIST and the automated deep learning of invariant features like VGG19, ResNet50, Inception v3, Xception. The numerical test results on 17 ICH dataset show that SVM models learned from Inception v3 and Xception features give good accuracy of 61.54% and 62.89% respectively. We propose to stack SVM models using different visual features to improve the classification result performed by any single one. Triplets (SVM-Xception, SVM-Inception-v3, SVM-VGG19), (SVM-Xception, SVM-Inception-v3, SVM-SIFT-BoW) achieve 65.32% of the classification correctness.

Keywords: Images of the intangible cultural heritage in the Mekong Delta · Image classification · Visual features · Support vector machines · Stacking.

1 Introduction

The Aniage project[1] focuses on high dimensional heterogeneous data based animation techniques for Southeast Asian Intangible Cultural Heritage (ICH) digital content. It aims to develop novel techniques and tools to reduce the production costs and improve the level of automation without sacrificing the control

[1] https://www.euh2020aniage.org

from the artists, in order to preserve the performing art related ICHs of Southeast Asia. The classification of ICH images[2] is the work package in the AniAge project.

The main aim is to automatically classify the image into one of predefined ICH categories. It requires to collect high quality ICH images organized by their categories (classes/labels) and study vision approaches to classify ICH images. To pursue this goal, we build an images dataset of 17 ICH categories by querying a text-based web search engine of Google, followed which we manually annotate them. And then, we explore popular vision approaches to deal with the classification task of ICH images. The extraction of visual features are performed by three popular handcrafted features such as the scale-invariant feature transform (SIFT [15,16]) and the bag-of-words model (BoW [22,13,2]), the histogram of oriented gradients (HOG [7]), the GIST [18]. Recent pre-trained deep learning networks, including VGG19 [21], ResNet50 [10], Inception v3 [23], Xception [5] are used to extract invariant features from ICH images. And then, Support vector machines (SVM [24]) models are learned from visual features to classify ICH images. The numerical test results on 17 ICH dataset show that SVM models learned from Inception-v3 and Xception features give good accuracy of 61.54% and 62.89% respectively. We propose to stack SVM classifiers using different visual features to improve classification results given by any single one. Triplets (SVM-Xception, SVM-Inception-v3, SVM-VGG19), (SVM-Xception, SVM-Inception-v3, SVM-SIFT-BoW) achieve 65.32% of the classification correctness.

The paper is organized as follows. Section 2 describes how to collect a dataset of ICH images and how to build classification models from vision approaches. Section 3 shows the experimental results before conclusions and future works presented in section 4.

2 Classification of intangible cultural heritage images

The classification system of ICH images in Fig. 1 follows the usual framework for the classification of images. It involves three steps: 1) building the high quality dataset of images, 2) extracting visual features from images and representing them, and 3) training SVM classifiers.

2.1 The dataset of intangible cultural heritage images

Firstly, we need to build the dataset of ICH images in the Mekong Delta, Vietnam. Fig. 2 shows an images sample of 17 ICH categories. Our proposal is to collect ICH images from Google due to the availability of this biggest public repository. It just does image search by textual query being key words related to 17 ICH categories and retrieve them. However, there are still noisy and irrelevant images. And then we do the manual post-processing stage and tagging images to obtain the high quality images organized by their ICH categories. Table 1 presents the dataset description with a total of 7409 images.

[2] http://aniage.ctu.edu.vn/myproj

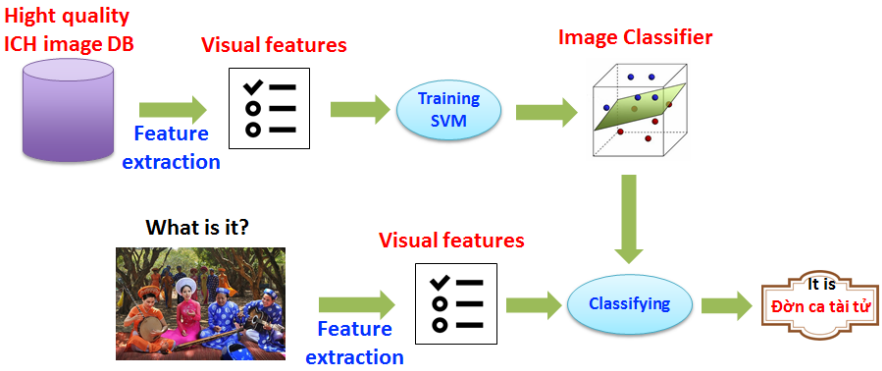

Fig. 1. Framework for classifying ICH images

Fig. 2. Images of 17 ICH categories

Table 1. Dataset description of 17 ICH categories

No	Category	#Images
1	Đờn ca tài tử Nam Bộ	513
2	Nghệ thuật Chăm riêng chà pây Khmer	185
3	Nghề Dệt chiếu	642
4	Lễ hội cúng biển Mỹ Long	398
5	Nghệ thuật sân khấu Dù kê Khmer	404
6	Lễ hội Ok om bok Khmer	465
7	Lễ hội vía Bà Chúa Xứ Núi Sam	405
8	Đại lễ Kỳ yên Đình Tân Phước Tây	223
9	Lễ hội vía Bà Ngũ Hành	569
10	Lễ hội Làm chay	365
11	Nghề đóng xuồng ghe Long định	281
12	Nghề Đan tre	641
13	Lễ cúng Việc lề	447
14	Lễ hội Đua bò Bảy Núi	449
15	Lễ hội Nghinh Ông	523
16	Lễ hội anh hùng Trương Định	361
17	Văn hóa chợ nổi Cái Răng	538
	Total	7409

2.2 Visual approaches for classifying intangible cultural heritage images

Visual approaches perform the classification task of ICH images via two key steps. The first one is to extract visual features from images and represent them. Followed which, the second one is to train SVM models to classify images.

Three popular methods for handcrafted features include the scale-invariant feature transform (SIFT [15,16]) and the bag-of-words model (BoW [22,13,2]), the histogram of oriented gradients (HOG [7]), the GIST [18].

Scale-invariant feature transform: The SIFT descriptors [15,16]) and the bag-of-words model (BoW) are the most commonly image represenation for tasks of images classification [22,13,2]. The SIFT method detects the appearance of the object at particular interest points, invariant to image scale, rotation, and also robust to changes in illumination, noise, and occlusion.

Histogram of oriented gradients: The HOG descriptors are used for human detection [7]. The HOG method computes the distribution of local intensity gradients or edge directions to describe local object appearance and shape within an image. The combined distributions form the image representation. The HOG descriptor is invariant to geometric and photometric transformations, except for object orientation.

GIST: The GIST descriptors proposed by [18] are used for images retrieval. The GIST method uses Gabor filters to extract the set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the spatial structure of a scene.

Recent deep learning networks such as VGG19 [21], ResNet50 [10], Inception v3 [23], Xception [5] are pre-trained on ImageNet dataset [8]. These deep learning networks are used to extract invariant features from ICH images.

VGG19: The VGG19 network architecture [21] consists of 19 weight layers for large scale image recognition. The VGG19 network uses only 3x3 convolutional layers stacked on top of each other to develop depth. The max pooling layers are used to reduce volume size. From the input layer to the last max pooling layer are used as features extraction of images.

ResNet50: The ResNet50 network architecture [10] is designed with 50 weight layers for image recognition. The ResNet50 develops extremely deep networks by proposed micro-architecture modules (called network-in-network). Furthermore, network layers try to fit a residual mapping instead of desired one. From the input layer to the last pooling layer or the last convolutional layer are used to extract image features.

Inception-v3: The "Inception" module proposed by [23] is to learn multi-level features for image classification. The main idea uses 1x1, 3x3 and 5x5 convolutions within the Inception module of the network. And then these Inception modules are stacked on top of each other. The reduction of volume size bases on 1x1 convolutions. From the input layer to the last pooling layer or the last convolutional layer are regarded as features extractor for images.

Xception: The "Xception" network proposed by [4] is an extension of the Inception architecture. The Xception replaces the standard depthwise separable convolution (the depthwise convolution followed by a pointwise convolution) by the new modified one without any intermediate activation being the pointwise convolution followed by a depthwise convolution. Features extraction for images is performed by layers from the input layer to the last pooling layer or the last convolutional layer.

Support vector machines: For a binary classification problem depicted in Fig. 3, the SVM algorithm proposed by [24] tries to find the best separating plane furthest from both class +1 and class -1. To pursue this aim, the training SVM algorithm simultaneously maximize the margin (or the distance) between the supporting planes for each class and minimize errors.

The binary SVM solver can be extended for dealing with the multi-class problems (c classes, $c \geq 3$). The main idea is to decompose multi-class into a series of binary SVMs, including One-Versus-All [24], One-Versus-One [12]. The One-Versus-All strategy (as illustrated in Fig. 4) builds c different binary SVM models where the i^{th} one separates the i^{th} class from the rest. The One-Versus-One strategy (as illustrated in Fig. 5) constructs $c(c-1)/2$ binary SVM

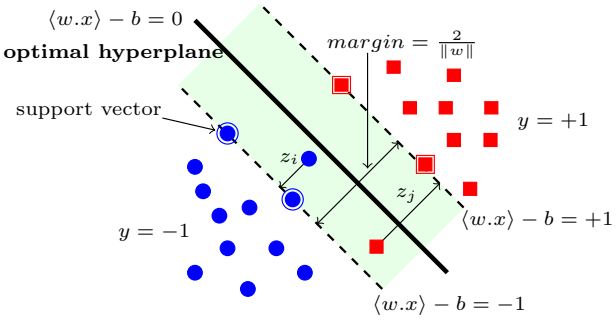

Fig. 3. Classification of the datapoints into two classes

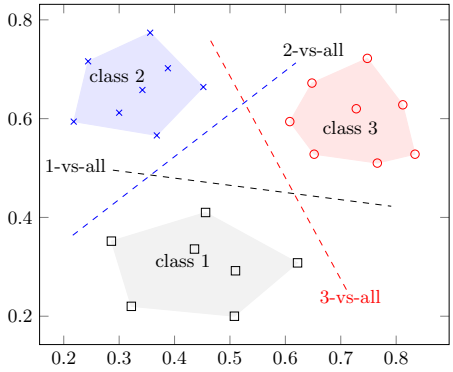

Fig. 4. Multi-class SVM (One-Versus-All)

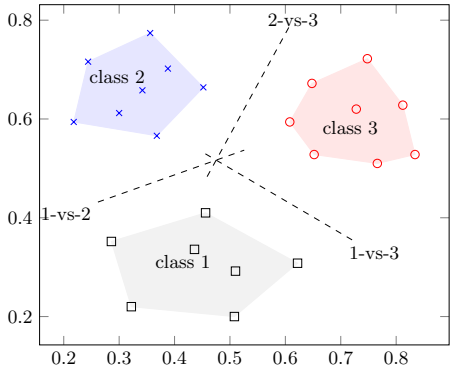

Fig. 5. Multi-class SVM (One-Versus-One)

models for all the binary pairwise combinations of the c classes. The class is then predicted with the largest distance vote. In practice, the One-Versus-All strategy is implemented in LIBLINEAR [9] and the One-Versus-One technique is also used in LibSVM [3].

SVM algorithms use different kernel functions [6] for dealing with non-linear classification tasks. The commonly non-linear kernel functions include a polynomial function of degree d, a radial basis function (RBF).

3 Experimental results

In this section, we present experimental results of different visual approaches for classifying ICH images. We implement them in Python using library Keras [4] with backend Tensorflow [1], library Scikit-learn [19] and library OpenCV [11]. All experiments are conducted on a machine Linux Fedora 23, Intel(R) Core i7-4790 CPU, 3.6 GHz, 4 cores and 32 GB main memory and the Nvidia GeForce GTX 960M 2GB GPU.

The image dataset of 17 ICH categories in the Mekong Delta, Vietnam is randomly split into the trainset (6001 images), the validation set (667 images) and testset (749 images). We use the trainset to build visual classification models. Then, results are reported on the testset using the resulting visual classification models.

3.1 Tuning parameters

We use the validation set to tune parameters for building visual classification models on the trainset. With methods for feature extractor and image representation, only handcrafted features SIFT and BoW model needs tuning the number of clusters (visual words) well-known as the parameter of $k$means algorithm [17]. We try to vary the number of visual words from 1000 to 5000 for finding the best experimental results. And then, the results are unchanged while increasing the number of visual words over 2000. Therefore, we use 2000 visual words for the BoW model.

With SVM models, we propose to use RBF kernel functions because it is general and efficient [14]. There is need to tune the hyper-parameter γ of RBF function $[K\langle x_i, x_j\rangle = exp(-\gamma||x_i - x_j||^2)]$ and the cost C (a trade-off between the margin size and the errors) to obtain the best correctness. Finally, we find out best parameters' SVM in Table 2 for visual classification models.

3.2 Classification results for 17 ICH categories

We obtain classification results of visual approaches in Table 3 and Fig. 6. The highest accuracy is bold-faced and the second one is in italic. In the comparison among visual classification approaches, we can see that methods for handcrafted features extraction such as SIFT-BoW, HOG, GIST are not suited for classifying ICH images. Recent deep networks (excepting ResNet50) for extracting invariant

Table 2. Hyper-parameters for training SVM models

No	Feature extraction method	γ	C
1	SIFT and BoW	0.0001	1000000
2	HOG	0.1	100000
3	GIST	5	1000000
4	VGG19	0.005	100000
5	ResNet50	0.005	100000
6	Inception-v3	0.005	100000
7	Xception	0.005	100000

features from ICH images are most accurate results. Typically, Xception and Inception v3 achieve 62.89% and 61.54% in terms of overall classification accuracy, respectively. We also try to tune these pre-trained deep networks by re-training about their 10% of layers from our image trainset but obtained results can not be improved even degraded.

Table 3. Overall classification accuracy for 17 ICH categories

No	Visual approach	Accuracy (%)
1	SVM-SIFT-BoW	33.87
2	SVM-HOG	32.93
3	SVM-GIST	37.25
4	SVM-VGG19	50.47
5	SVM-ResNet50	34.14
6	SVM-Inception-v3	*61.54*
7	SVM-Xception	**62.89**

3.3 Stacking of SVM classifiers for classifying 17 ICH categories

We propose to use voting scheme [25] among visual models to improve classification correctness for ICH images. The main idea is to combine multiple visual classifiers learned for the classification task by weighted voting between the prediction of each visual classifier VC_i as illustrated in equation (1).

$$Majority-vote\{w_1 * pred(x, VC_1) + w_2 * pred(x, VC_2) + \ldots + w_k * pred(x, VC_k)\} \quad (1)$$

Voting schemes always use the visual classifier SVM-Xception because this model gives the best result. Followed which, other visual models are included in voting schemes with the hope that the models can complement one another in the classification. Table 4 and Fig. 7 show results obtained by weighted voting schemes.

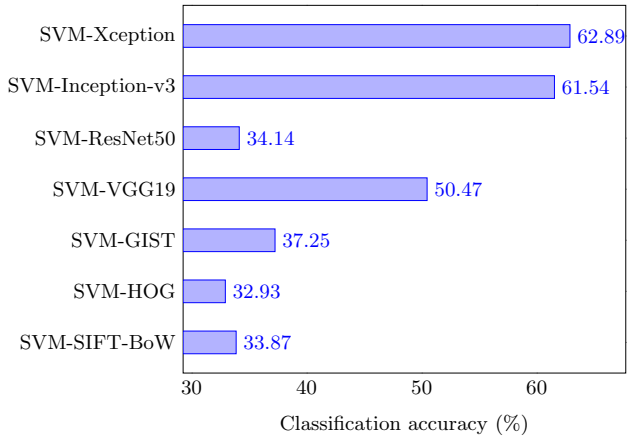

Fig. 6. Overall classification accuracy for 17 ICH categories

The couple of SVM-Xception and SVM-Inception-v3 improves 1.62% and 2.97% of classification correctness against SVM-Xception and SVM-Inception-v3, respectively.

The improvements of the triplet SVM-Xception, SVM-Inception-v3 and SVM-VGG19 over each single visual classifier are 2.43%, 3.78% and 14.85%, respectively.

The triplet SVM-Xception, SVM-Inception-v3 and SVM-SIFT-BoW achieves the best accuracy as the triplet SVM-Xception, SVM-Inception-v3 and SVM-VGG19. It also improves 31.45% of the accuracy compared to SVM-SIFT-BoW.

Table 4. Overall classification accuracy of voting schemes for 17 ICH categories

No	Voting scheme	Accuracy (%)
8	0.85*SVM-Xception + 0.15*SVM-SIFT-BoW	63.29
9	0.8*SVM-Xception + 0.2*SVM-HOG	63.02
10	0.8*SVM-Xception + 0.2*SVM-GIST	63.02
11	0.75*SVM-Xception + 0.25*SVM-VGG19	63.56
12	0.75*SVM-Xception + 0.25*SVM-ResNet50	63.02
13	0.65*SVM-Xception + 0.35*SVM-Inception-v3	64.51
14	0.55*SVM-Xception + 0.225*SVM-Inception-v3 + 0.225*SVM-VGG19	**65.32**
15	0.65*SVM-Xception + 0.22*SVM-Inception-v3 + 0.13*SVM-SIFT-BoW	**65.32**

4 Conclusion and future works

We have presented visual approaches for classifying images of the intangible cultural heritage (ICH) in the Mekong Delta, Vietnam. We collect an images

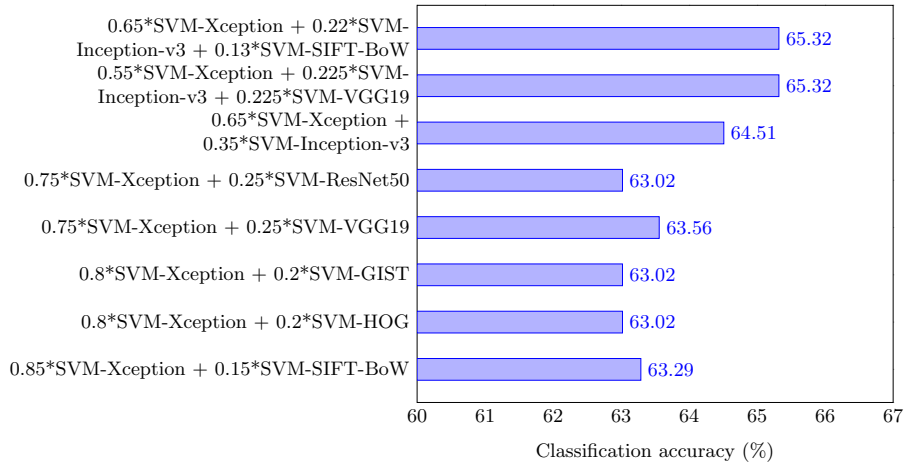

Fig. 7. Overall classification accuracy of voting schemes for 17 ICH categories

dataset of 17 ICH categories from Google and manually tagging images according to their categories. Visual approaches are used to deal with the ICH image classification. The feature extraction methods include three popular handcrafted features such as SIFT-BoW, HOG, GIST and four recent deep learning networks of invariant features like VGG19, ResNet50, Inception v3, Xception. Followed which SVM models are learned from these visual features to classify ICH images. The numerical results on 17 ICH dataset show that SVM-Xception and SVM-Inception-v3 give good accuracy of 61.54% and 62.89% respectively. We propose to use voting schemes between visual models to improve the classification result performed by any single one. Triplets (SVM-Xception, SVM-Inception-v3, SVM-VGG19), (SVM-Xception, SVM-Inception-v3, SVM-SIFT) achieve 65.32% of the classification correctness.

These visual approaches can be used to re-rank images retrieved from Google and then we select top-ranked images for automated organizing ICH images by their ICH categories. It allows us to build a large number of images for a specified ICH category. Another approach [20] for developing the images database size combines textual and visual features.

Acknowledgments This work has received support from the European Project H2020 Marie Sklodowska-Curie Actions (MSCA), Research and Innovation Staff Exchange (RISE): Aniage project (High Dimensional Heterogeneous Data based Animation Techniques for Southeast Asian ICH Digital Content), No: 691215.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irv-

ing, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Proceedings of the European Conference on Computer Vision. pp. 517–530 (2006)
3. Chang, C.C., Lin, C.J.: LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(27), 1–27 (2011)
4. Chollet, F., et al.: Keras. https://keras.io (2015)
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. CoRR **abs/1610.02357** (2016)
6. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods. Cambridge University Press, New York, NY, USA (2000)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01. pp. 886–893. IEEE Computer Society (2005)
8. Deng, J., Berg, A.C., Li, K., Li, F.: What does classifying more than 10, 000 image categories tell us? In: Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V. pp. 71–84 (2010)
9. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9**(4), 1871–1874 (2008)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015)
11. Itseez: Open source computer vision library. https://github.com/itseez/opencv (2015)
12. Kreßel, U.H.G.: Pairwise classification and support vector machines. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods, pp. 255–268. MIT Press, Cambridge, MA, USA (1999)
13. Li, F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA. pp. 524–531 (2005)
14. Lin, C.: A practical guide to support vector classification (2003)
15. Lowe, D.: Object recognition from local scale invariant features. In: Proceedings of the 7th International Conference on Computer Vision. pp. 1150–1157 (1999)
16. Lowe, D.: Distinctive image features from scale invariant keypoints. International Journal of Computer Vision pp. 91–110 (2004)
17. MacQueen, J.: Some methods for classification and analysis of multivariate observations. Berkeley Symposium on Mathematical Statistics and Probability, University of California Press (1), 281–297 (1967)
18. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision **42**, 145–175 (2001)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

20. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. IEEE Trans. Pattern Anal. Mach. Intell. **33**(4), 754–766 (Apr 2011)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
22. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: 9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France. pp. 1470–1477 (2003)
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR **abs/1512.00567** (2015)
24. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)
25. Wolpert, D.: Stacked generalization. Neural Networks **5**, 241–259 (1992)