Robust validation steps for clip classification

Mariano Rodríguez, Liam Schoneveld, Vincent Garcia

ROBUST VALIDATION STEPS FOR CLIP CLASSIFICATION

Mariano Rodríguez, Liam Schoneveld, and Vincent Garcia

Unique Entertainment Experience SAS, France

ABSTRACT

The classic approach to clip classification consist of computing scores per class and identifying the most frequent top-score classes. Usually, predictions are always reported; and consequently, we are either right or wrong. However, in most applications, erroneous predictions will carry substantial negative effects. Ideally, we should communicate only correct predictions and abstain from reporting erroneous ones. Two novel methods are shown to come near this goal. Their main advantages are: on the one hand, most false positives are successfully identified; on the other hand, the best levels of accuracy are maintained, if not improved.

***Index Terms*—** image classification, classifiers, neural networks, validation step, a contrario, NFA.

1. INTRODUCTION

Image classification is a fundamental task in computer vision that categorizes images by executing predefined operations on their pixels. Video classification can be sometimes very similar to image classification. Indeed, a digital video can be viewed as a discrete sequence of frames (i.e. images) sampled at a certain frequency. Typically, the classification process relies on different components. The main component consist of cutting edge classifiers which are, for the most part, derived from deep learning techniques. The direct output of these classifiers are usually very good nowadays. Still, it is a common practice to impose some additional conditions in order to validate predictions from the classifier: a validation step. This step is usually very simple, and it comes in to improve performance and to evaluate the confidence of the classifier.

Several datasets have been made available in the literature to keep track of best performing models for various classes: MNIST [1], ImageNet [2], Oxford IIIT Pets [3], CIFAR-100 [4], Caltech-UCSD Birds [5], among others. Many successful deep learning architectures have been proposed and tested on these datasets for image classification: LeNet-5 [6], AlexNet [7], VGGNet [8], GoogLeNet (Inception-V1) [9], ResNet-50 [10], NASNet [11], ViT [12], among many others. Figure 1 shows the VGG16 [8] architecture setup for classification over the ImageNet [2] dataset.

In comparison to classifiers, the validation step has received less attention by the scientific community. A possible explanation is that in order to increase performance it usually works better to improve the classifier itself. Nevertheless, in some cases improving the classifier is not an option (e.g. using a third-party classifier, new training data is not available yet, the cost of the training is expensive, etc); and even when enjoying performance from a state-of-the-art classifier, the validation step should end up improving the robustness of the method. When classifiers fail, they often tend to mis-classify each class into several different others, seemingly randomly. This noise can be used to identify lack of confidence from the classifier. A suitable framework to validate events in the presence of noise is

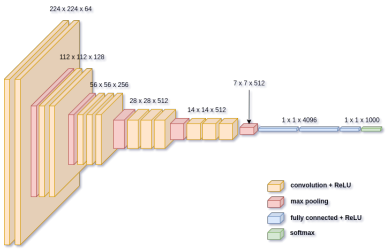


Fig. 1: The VGG16 [8] architecture for classification over the ImageNet [2] Large Scale Visual Recognition Challenge dataset.

the *a-contrario* theory introduced in [13]. The *a-contrario* methodology has already been successfully used in diverse computer vision applications [14–22]. They all have in common the proposition of a metric followed by an evaluation of agreement.

In this paper, we propose two robust validation steps for video clip classification that attempt to identify false positives by assessing confidence. If there is not enough agreement among predicted classes per frame, any resulting clip classification is invalidated and re-assigned to a virtual *unknown* class. The *unknown* class signals for no reporting i.e. no classification is to be communicated to the outside world. The proposed methods analyze top scores per frame up to predefined levels, thus enabling for multiple rank detections. Coherently, classes detected at top ranks (e.g. $\arg\max$) are considered more significant than poorly ranked classes.

This paper is organized as follows. The terminology as well as two main scenarios are described in Section 2. The two novel classifiers with *a-contrario* validation are proposed in Section 3. Experimental results are presented in Section 4. Section 5 concludes the paper.

2. THE CONTEXT

We define a video, denoted by $\mathcal{V}$, as a sequence of frames $\{\mathbf{f}_1, \cdots, \mathbf{f}_n\}$ that are equally spaced in time. A frame is formally defined as a function that usually belongs to $\mathbb{R}^2 \mapsto \mathbb{R}$ (a grayscale image) or to $\mathbb{R}^2 \mapsto \mathbb{R}^3$ (a color image). We denote the set of frames as $\mathbb{F}$. A clip $\mathcal{C}$ is defined as a small contiguous subset of the video $\mathcal{V}$,

$$\mathcal{C} \coloneqq \{\mathbf{f}_a, \cdots, \mathbf{f}_b\} \subset \mathcal{V},$$

where $1 \leq a < b \leq n$. Let also $\mathcal{P} \coloneqq \{\mathbf{p}^1, \cdots, \mathbf{p}^N\}$ denote the set of classes or properties. We assume that each frame $\mathbf{f}$ can be classified into one and only one class $\mathbf{p}$, i.e. there exists a function ψ, called frame classifier, such that, $\psi(\mathbf{f}) = \mathbf{p}$. Similarly, a clip can be classified if there exists a clear predominant class among its frames. The frame classifier ψ is often assumed to depend explicitly on a score function, $\phi : \mathbb{F} \mapsto \mathbb{R}^N$, that assigns to each frame $\mathbf{f}$ a score

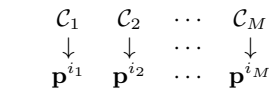


Fig. 2: Scenario A. Estimating classes per clip, i.e., given a $\mathcal{C}_l$ we are to predict the class $\mathbf{p}^{i_l}$ to which it belongs.

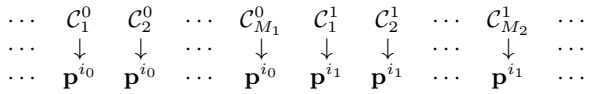


Fig. 3: Scenario B. Given a clip $\mathcal{C}_l$ we are to predict its class $\mathbf{p}^{i_l}$ knowing that this class is probably associated with several other clips. The number of groups and the number of clips in each group are unknown.

per class, $\phi(\mathbf{f}) = (s_1, \cdots, s_N)$. The score function is frequently approximated by a neural network.

Based on ϕ, let us define the greedy frame classifier as:

$$\psi_g(\mathbf{f}) \coloneqq \mathbf{p}^{\arg\max \phi(\mathbf{f})}. \tag{1}$$

For simplicity in this paper, the $\arg\max$ operator is viewed as a function i.e. $\arg\max\colon \mathbb{R}^N \mapsto \{1, \cdots, N\}$. In practice, whenever two or more classes share the maximal score, the $\arg\max$ function follows a predefined set of rules in order to deterministically return only one *argument of the maxima*. Likewise, a consensual clip classifier can be defined based on scores per frame and a simple consensual step,

$$\varphi_0(\mathcal{C}) \coloneqq \mathbf{p}^j \tag{2}$$

where

$$j = \underset{i=1,\cdots,N}{\arg\max} \sum_{\mathbf{f}\in\mathcal{C}} \mathbf{1}_{\{i=\arg\max \phi(\mathbf{f})\}}. \tag{3}$$

Finally, a common approach to assess confidence is to impose a minimum number of rank one observations from the greedy frame classifier. We call it the α-consensual clip classifier and it is formally defined as:

$$\varphi_\alpha(\mathcal{C}) = \begin{cases} \mathbf{p}^j, & \alpha < \sum_{\mathbf{f}\in\mathcal{C}} \mathbf{1}_{\{j=\arg\max \phi(\mathbf{f})\}} \\ \mathbf{p}^u, & \text{otherwise} \end{cases} \tag{4}$$

where j is defined as in Equation 3 and $\mathbf{p}^u$ denotes an added extra class signaling unconfidence.

Two main scenarios are addressed in this paper when classifying clips. Let us describe them properly.

2.1. Scenario A. Classes per clip

Given a collection of clips, we are asked to predict the class associated with each clip, see Figure 2. In this scenario, we deal with clips in an independent manner, i.e., we assume no connection between clips. Examples of this scenario are numerous and they naturally kick in whenever there is need of tags per clip (e.g. main object being focused by the camera, city or celebrity identification, type of background, etc).

2.2. Scenario B. Classes per group of clips

As for the previous scenario, a collection of clips is available and we are to predict a class per clip. In addition, clips are supposed to come in groups, and frames within each group should share a common and unique class. The number of groups as well as the number of clips in each group is unknown beforehand. Figure 3 shows a representation of this scenario.

This scenario is typical when dealing with clips coming from a long video in which changes of the targeted events are expected to occur from time to time. Examples of this situation are: game identification (when people record themselves playing video games they often play one game for several minutes and possibly change to another after a while); spoken language identification (when talking people are expected to speak the same language for a while, and possibly switch to another later on); etc.

3. A CONTRARIO VALIDATION

The proposed validation procedure is based on the *a contrario* theory [13], which relies on the non-accidentalness principle [23, 24]. Informally, this principle states that there should be no detection in noise. In our context, we assess the existence of a causal relation between several outputs from the score function ϕ.

3.1. Detecting meaningful classes

In a uniformly distributed random world, you would not expect to observe a class more frequently than any other. Thus, we propose to identify classifier confidence with specific anomalies in this random world. To evaluate confidence, we measure the probability of having equal or more frequent appearances of a class than what was observed. This means that the real assumption in this subsection is a certain similarity between the right tails from the resulting distributions: one derived from our forthcoming assumption and the other coming from the true distribution over ϕ-ranked vectors. Indeed, if we observe a measure to the left hand side of the right tail, a precise probability value is not needed as it will not get tagged as meaningful anyway.

Given a sequence of frames $\mathcal{F} \coloneqq \{\mathbf{f}_1, \cdots, \mathbf{f}_M\}$ and the score function ϕ, we define a causality function as follows. Our stochastic model $\mathcal{H}_0$ used to evaluate accidentalness assumes a uniform distribution over ϕ-ranked vectors. Let $R_k^{\phi(\mathbf{f})}$ be a random variable (r.v.) representing all classes associated to the k highest ϕ-scores from frame $\mathbf{f}$ i.e. $\text{rank}_k \phi(\mathbf{f})$. For example, an observed value of $R_1^{\phi(\mathbf{f})}$ corresponds to ψ_g, the greedy frame classifier from Equation 1. A simple calculation states that

$$\mathbb{P}_{\mathcal{H}_0}\left(\mathbf{p} \in R_k^{\phi(\mathbf{f})}\right) = \frac{k V_{N-1}^{k-1}}{V_N^k}, \tag{5}$$

where N is the total number of classes and V_N^k is the variations of k elements among N. In order to assess the accidentalness of ϕ over $\mathcal{F}$, we propose to base the validation on the following r.v.:

$$C_{\mathcal{F},\phi}^k(\mathbf{p}) = \sum_{\mathbf{f}\in\mathcal{F}} \mathbf{1}_{\left\{\mathbf{p}\in R_k^{\phi(\mathbf{f})}\right\}}, \tag{6}$$

where $\left\{R_k^{\phi(\mathbf{f})}\right\}_{\mathbf{f}\in\mathcal{F}}$ are assumed independent and identically distributed (i.i.d.). Thus, $C_{\mathcal{F},\phi}^k(\mathbf{p})$ is a binomial r.v. whose number of trials equals M and the probability of success for each trial appears in Equation 5. Finally, the causality function is defined as a realization of $C_{\mathcal{F},\phi}^k(\mathbf{p})$, i.e. $c_{\mathcal{F},\phi}^k(\mathbf{p}) = \sum_{\mathbf{f}\in\mathcal{F}} \mathbf{1}_{\left\{\mathbf{p}\in r_k^{\phi(\mathbf{f})}\right\}}$.

To assess the accidentalness of frequent appearences of a class $\mathbf{p}$, we need to evaluate the survival function of our binomial r.v. in Equation 6 at time $c_{\mathcal{F},\phi}^k(\mathbf{p})$: $\mathbb{P}_{\mathcal{H}_0}\left(C_{\mathcal{F},\phi}^k(\mathbf{p}) \geq c_{\mathcal{F},\phi}^k(\mathbf{p})\right)$. When this probability is small enough, there exists evidence to reject the

null hypothesis and declare the class $\mathbf{p}$ meaningful. However, one needs to consider that usually multiple classes are tested. If 100 tests are performed, for example, it would not be surprising to observe an event that appears with probability 0.01 under random conditions. Thus, the number of tests N_T needs to be included as a correction term, as it is done in the statistical multiple hypothesis testing framework [25],

$$N_T = |\bigcup_{\mathbf{f}\in\mathcal{F}} r_k^{\phi(\mathbf{f})}|.$$

Following the *a contrario* methodology [13], we define the *Number of False Alarms* (NFA) of a class $\mathbf{p}$ as:

$$\mathrm{NFA}_{k,\mathcal{F},\phi}(\mathbf{p}) = N_T \cdot \mathbb{P}_{\mathcal{H}_0}\left(C^k_{\mathcal{F},\phi}(\mathbf{p}) \geq c^k_{\mathcal{F},\phi}(\mathbf{p})\right). \quad (7)$$

Classes with NFA $\leq \varepsilon$, for a predefined ε value, are accepted as valid. One can show [13] that under $\mathcal{H}_0$, the expected number of classes with NFA $\leq \varepsilon$ is bounded by ε. As a result, ε corresponds to the mean number of false detections under $\mathcal{H}_0$. We set the value $\varepsilon = 10^{-3}$, meaning that, under $\mathcal{H}_0$, only 1 out of 1000 draws ends up with a false detection.

Algorithm 1 introduces our proposal to detect classes from a sequence of frames $\mathcal{F}$ and a score function ϕ. Let us now address scenarios A and B from Subsections 2.1-2.2.

Algorithm 1 DETECTCLASSES

input:
$\mathcal{F}$ - Sequence of frames.
ϕ - Score function.
$k_{\max}$ - Maximal accepted rank.
ε - NFA threshold.
start:
list_of_detections = $\emptyset$
foreach $k \in \{1, \cdots, k_{\max}\}$ **do**
  foreach $\mathbf{p} \in \bigcup_{\mathbf{f}\in\mathcal{F}} r_k^{\phi(\mathbf{f})}$ **do**
    if $\mathrm{NFA}_{k,\mathcal{F},\phi}(\mathbf{p}) < \varepsilon$ **then**
      ▷ $\mathbf{p}$ is detected as meaningful
      append $(k, \mathrm{NFA}_{k,\mathcal{F},\phi}(\mathbf{p}), \mathbf{p})$ to list_of_detections
return list_of_detections

3.2. Detections in Scenario A

Under conditions similar to those from scenario A, Algorithm 2 is able to classify a collection of clips either into the set of available classes or into the additional *unknown* class. The added *unknown* class, denoted as $\mathbf{p}^u$, is used whenever the score function ϕ seems unconfident.

3.3. Detections in Scenario B

Under the hypothesis of scenario B, Algorithm 3 makes global and per clip detections in order to assess causality. This is extremely useful when some clips do not have enough frames to detect confidence from the score function ϕ. The idea behind it is simple. First, a global strong detection is more reliable than weak per-clip detections. Indeed, it is easier to assess confidence when more frames are involved. Finally, as global detections lose the per clip information, we use their strongly detected classes to validate weak detections per clip.

Algorithm 2 can be used in this scenario as well, but it will not exploit the fact that clips come in groups. Algorithm 3 is indeed more adapted to this situation. The reader will notice that if a class $\mathbf{p}$ is not strongly (i.e. globally) detected as meaningful by Algorithm 3, then no clip association is possible for this class, even if there exists a clip $\mathcal{C}$ such that $\forall \mathbf{f} \in \mathcal{C},\ \mathbf{p} = \mathbf{p}_{\arg\max \phi(\mathbf{f})}$. Conversely, the most meaningful class in a clip (i.e. weakly meaningful) might not necessarily be the one assigned to the clip by Algorithm 3. Indeed, a less meaningful class in that clip might be more meaningful globally, and therefore have a better chance to emerge as the class assigned to the clip. Additionally, and depending upon the application, it could make sense to attach the *unknown* class $\mathbf{p}^u$ to any clip for which two or more global detections are weakly detected.

Algorithm 2 ACONTRARIOCLIPCLASSIFIER

input:
$\{\mathcal{C}_1, \cdots, \mathcal{C}_N\}$ - Clips to classify.
ϕ - Score function.
$k_{\max}$ - Maximal accepted rank per clip.
ε - NFA threshold per clip.
start:
foreach $\mathcal{C} \in \{\mathcal{C}_1, \cdots, \mathcal{C}_N\}$ **do**
  $\mathcal{D}(\mathcal{C}) =$ first class in lexicographical order w.r.t. (rank,nfa) from
    DETECTCLASSES$(\mathcal{C}, \phi, k_{\max}, \varepsilon)$.
    ▷ if DETECTCLASSES returns $\emptyset$, we then set $\mathcal{D}(\mathcal{C}) = \mathbf{p}^u$
return $\mathcal{D}$

Algorithm 3 ACONTRARIOGROUPCLASSIFIER

input:
$\{\mathcal{C}_1, \cdots, \mathcal{C}_N\}$ - Clips to classify.
ϕ - Score function.
$k^g_{\max}$ - Global maximal accepted rank.
$k_{\max}$ - Maximal accepted rank per clip.
ε^g - Global NFA threshold.
ε - NFA threshold per clip.
start:
$\mathcal{D}(\mathcal{C}_i) = \mathbf{p}^u \quad \forall i$ ▷ initialize $\mathcal{D}$
global_detections = DETECTCLASSES$(\cup_{i=1}^N \mathcal{C}_i, \phi, k^g_{\max}, \varepsilon^g)$
SORT(global_detections) ▷ in-place lexicographical order w.r.t. (rank,nfa)
foreach (k, nfa, $\mathbf{p}$) in global_detections **do**
  foreach $\mathcal{C} \in \{\mathcal{C}_1, \cdots, \mathcal{C}_N\}$ **do**
    if $\mathcal{D}(\mathcal{C}) = \mathbf{p}^u$ **and**
    $\mathbf{p}$ in DETECTCLASSES$(\mathcal{C}, \phi, k_{\max}, \varepsilon)$ **then**
      $\mathcal{D}(\mathcal{C}) = \mathbf{p}$
return $\mathcal{D}$

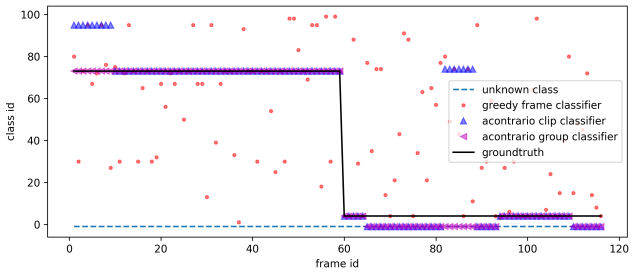
4. EXPERIMENTAL RESULTS

We have used these validation steps within our company (Unique Entertainment Experience) to identify false positives while improving performance. Our approach is simple: we prefer not to communicate a prediction if we know there is a high chance of it being a false positive. Our models not being public yet, we choose to present results based on simulated scenarios A and B from Subsections 2.1-2.2. We simulate these scenarios by ensuring all frames in a clip share a unique class; all frames belonging to a test set from CIFAR-100 [4].

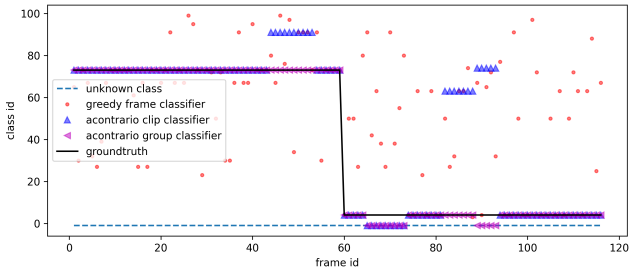
Each simulation of scenario A from Subsection 2.1 generates exactly one hundred clips. Each clip $\mathcal{C}_i$ consists of random frames

Classifiers	Scenario A						Scenario B					
	ϕ_{small}			ϕ_{vgg}			ϕ_{small}			ϕ_{vgg}		
	Acc	FP	U	Acc	FP	U	Acc	FP	U	Acc	FP	U
Greedy frame classifier (ψ_g)	30.40	69.60	-	63.48	36.52	-	30.51	69.49	-	63.17	36.83	-
Consensual clip classifier (φ_0)	60.91	39.09	-	93.07	6.93	-	60.96	39.04	-	92.87	7.13	-
α-Consensual clip classifier (φ_1)	58.99	20.60	20.41	92.84	5.58	1.58	59.02	20.98	20.00	92.70	5.84	1.46
α-Consensual clip classifier (φ_2)	41.84	4.09	54.07	87.61	2.37	10.02	42.07	4.11	53.82	87.27	2.50	10.23
α-Consensual clip classifier (φ_3)	24.91	0.73	74.36	75.55	0.72	23.73	25.15	0.81	74.04	74.54	0.77	24.69
α-Consensual clip classifier (φ_4)	13.08	0.13	86.79	58.61	0.20	41.19	12.80	0.13	87.07	57.80	0.22	41.99
ACONTRARIOCLIPCLASSIFIER$_*$	57.88	10.30	31.82	92.85	4.50	2.65	58.80	10.46	30.74	92.64	4.86	2.49
ACONTRARIOGROUPCLASSIFIER$_{**}$	-	-	-	-	-	-	58.56	3.65	37.79	95.09	1.49	3.43

Table 1: Classification metrics over one thousand random simulations of scenarios A and B from Subsections 2.1-2.2. Accuracy, false positive rate and unconfidence rate are reported for different classifiers based on two score functions: ϕ_{small} and ϕ_{vgg}. Legend: mean accuracy (Acc); mean rate of false positives (FP); mean rate of the virtual unknown class (U); non applicable (-). Our set of parameters in these experiments were: $k_{\max} = 5, \varepsilon = 10^{-3}$ (*); and $k^g_{\max} = 1, k_{\max} = 5, \varepsilon^g = \varepsilon = 10^{-3}$ (**). Cells in each column Acc and FP are colored linearly between light blue (worst scores) and dark blue (best scores); the U column is not colored.



(a) All classifiers depending on the score function ϕ_{small}.



(b) All classifiers depending on the score function ϕ_{vgg}.

Fig. 4: Visualisation of classifiers on a random simulation of scenario B from Subsection 2.2.

belonging to the i-th class from CIFAR-100. The number of frames per clip ranging randomly between 5 and 10.

Similarly, for each simulation of scenario B from Subsection 2.2, we draw several clips from two random classes. The number of clips per class and the number of frames per clip are chosen as random integers between 5 and 10. Either one or two classes are present in each simulation of scenario B.

Two score functions have been trained on CIFAR-100: the first, denoted by ϕ_{small}, consists of a small network (two convolutions and three fully connected layers) trained from scratch for 100 epochs; the second, denoted by ϕ_{vgg}, is a pretrained VGG16 model[1].

[1]available at pytorch hub: repo 'chenyaofo/pytorch-cifar-models' and model 'cifar100_vgg16_bn'.

Table 1 reports three classification metrics (accuracy, false positive rate and unconfidence rate) under scenarios A and B. Eight score-based classifiers are shown: the greedy frame classifier from Equation 1, the consensual clip classifier from Equation 2, four α-consensual clip classifiers from Equation 4, the *a-contrario* clip classifier from Algorithm 2 and the *a-contrario* group classifier from Algorithm 3. In both scenarios, the *a-contrario* clip classifier ranked among the highest accuracy scores while correctly identifying false positives; almost not sacrificing any true positive in exchange for recognizing false positives. A better compromise is achieved by the *a-contrario* group classifier in scenario B, where the false positive rate has been divided by three with respect to the *a-contrario* clip classifier. Dark blue colored cells represent the best scores per column in Table 1. The proposed methods are consistently highlighted as among the best scoring methods for both accuracy (Acc) and false positive rate (FP).

Figure 4 shows, on a random simulation from scenario B, the outputs of the greedy frame classifier and our two proposals. Notice the apparent randomness when classes are mis-classified by the greedy frame classifier, i.e. the red dots out of the groundtruth line. However, the classifier is sometimes pointing to a unique false positive; and some classes were not even observed. This means that our assumption of uniformity over ϕ-ranked vectors does not hold. Still, as commented in Subsection 3.1, this is not a dealbreaker, as our core assumption is of similar right tails of $R_k^{\phi(\mathbf{f})}$ under $\mathcal{H}_0$ and under real life conditions; which explains the success of the proposed classifiers under scenarios A and B.

5. CONCLUSIONS

In this paper we proposed two methods for clip classification. They measure confidence of the score function when classifying, allowing to accurately identify false positives. High levels of accuracy (if not the best) are kept, while detecting lack of confidence in erroneous predictions. These methods are helpful to prevent reporting when a clip is likely to be mis-classified. A small variation of this methodology could lead us to robust assignments of multiple classes per clip. This extension will be the focus of future work. All results (and more) presented in this paper are available at:
`https://rdguez-mariano.github.io/pages/valsteps`

Acknowledgments: We thank Maryan Morel, Oisín Benson, Tom Mason and Clara Gainon de Forsan de Gabriac for numerous suggestions.

6. REFERENCES

[1] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[4] A. Krizhevsky, G. Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[5] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-ucsd birds 200," 2010.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[12] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," *arXiv preprint arXiv:2106.04560*, 2021.

[13] A. Desolneux, L. Moisan, and J.-M. Morel, *From Gestalt Theory to Image Analysis*, Springer, 2008.

[14] L. Moisan and B. Stival, "A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix," *IJCV*, vol. 57, no. 3, pp. 201–218, 2004.

[15] F. Cao, J. L. Lisani, J. M. Morel, P. Musé, and F. Sur, *A Theory of Shape Identification*, Springer, 2008.

[16] J. Rabin, J. Delon, and Y. Gousseau, "A statistical approach to the matching of local features," *SIIMS*, vol. 2, no. 3, pp. 931–958, 2009.

[17] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *TPAMI*, vol. 32, pp. 722–732, 2010.

[18] R. Grompone von Gioi and V. Pătrăucean, "A contrario patch matching, with an application to keypoint matches validation," in *ICIP*, pp. 946–950. 2015.

[19] V. Pătrăucean, R. Grompone von Gioi, and M. Ovsjanikov, "Detection of mirror-symmetric image patches," in *CVPRW*, 2013, pp. 211–216.

[20] M. Rodriguez and R. G. V. Gioi, "Affine invariant image comparison under repetitive structures," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 1203–1207.

[21] M. Rodriguez, J. Delon, and J.-M. Morel, "Automatic detection of repeated objects in images," in *ICIP*, Anchorage, Alaska, United States, Sep 2021.

[22] T. Nikoukhah, J. Anger, M. Colom, J.-M. Morel, and R. Grompone von Gioi, "ZERO: a Local JPEG Grid Origin Detector Based on the Number of DCT Zeros and its Applications in Image Forensics," *Image Processing On Line*, vol. 11, pp. 396–433, 2021, `https://doi.org/10.5201/ipol.2021.390`.

[23] A. P. Witkin and J. M. Tenenbaum, "On the role of structure in vision," in *Human and Machine Vision*, J. Beck, B. Hope, and A. Rosenfeld, Eds., pp. 481–543. Academic Press, 1983.

[24] D. Lowe, *Perceptual Organization and Visual Recognition*, Kluwer Academic Publishers, 1985.

[25] A. Gordon, G. Glazko, X. Qiu, and A. Yakovlev, "Control of the mean number of false discoveries, bonferroni and stability of multiple testing," *Ann. Appl. Stat.*, vol. 1, pp. 179–190, 2007.