Pattern Recognition Letters
journal homepage: www.elsevier.com

Boosting image classification through semantic attention filtering strategies

Eduardo Fidalgo[a,c,**], Enrique Alegre[a,c], Victor González-Castro[a,c], Laura Fernández-Robles[b,c]

[a]*Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León, Spain*
[b]*Departamento de Ingenierías Mecánica, Informática y Aeroespacial, Universidad de León, Spain*
[c]*Researcher at INCIBE (Spanish National Cybersecurity Institute), León, Spain*

ABSTRACT

Saliency Maps, frequently used to highlight significant information, can be combined with other paradigms, such as Bag of Visual Words (BoVW), to improve image description when the saliency regions correspond closely with the objects of interest. In this paper, we present three attention filtering strategies based on their saliency map that improve image classification using the BoVW framework, Spatial Pyramid Matching (SPM) and Convolutional Neural Networks (CNN) features. Firstly, we demonstrate how the blurring factor used in the Hou's image signature algorithm determines what information remains and impacts to the obtained accuracy in image classification. Next, we propose AutoBlur, a simple but effective approach to automatically select this factor. Then, based on AutoBlur, we introduce two variants of our approach SARF (Semantic Attention Region Filtering), to semantically remove non-relevant regions through a Mean Shift segmentation. The first one is based on the intersection of the Hou's image attention areas with its Mean Shift segmentation, while the second one discards regions using a key point voting system that relies on the Euclidean distance. The experiments carried out showed that the methods of Semantic Attention Filtering that we are proposing could be successfully used with both BoVW, SPM and CNN's in most of the evaluated situations. In the five datasets assessed, all the three proposed methods outperform the baseline when using BoVWs in almost every case. For Spatial Pyramid Matching, the behaviour is similar, finding that the baseline is superior to our proposals in only one of the datasets used. In the case of CNN's, our filtering proposal outperforms the baseline in two datasets, being very similar to it in the other cases.

1. Introduction

The importance of assigning a category automatically to an image from a set of categories, i.e., image classification, has grown enormously in the last few years.

Convolutional Neural Networks (CNN) are the current state-of-the-art in image classification (song Tang et al., 2017) and are known to outperform tradicional machine learning methods fed with hand-crafted features (Krizhevsky et al., 2012). However, CNN are large and complex models that require a huge number of training data to avoid over-fitting. Even though there are techniques to carry out data augmentation (Krizhevsky et al., 2012), there are some cases where there is not enough data because it is challenging to obtain, e.g. medical imaging applications, or due to other reasons (Wesam et al., 2017).

Regardless this drawback, it has been demonstrated how the use of CNN features extracted from a pre-trained network on a big dataset, like ImageNet (Russakovsky et al., 2015) can still obtain state of the art results (Paulin et al., 2015) on even small datasets. However, even in these situations, if the CNN features do not contain relevant information for the intended image classification tasks, CNN results can be improved with a proper selection of areas of interest in the image.

In both previous scenarios, i.e. small number of training images or challenging image classification tasks, it could be necessary to extract "hand-crafted" features from the image to *describe* it and thereafter feed traditional machine learning techniques –, e.g., Support Vector Machines (SVM) or Random Forests – with these feature vectors, i.e., descriptors.

**Corresponding author
e-mail: efidf@unileon.es (Eduardo Fidalgo), ealeg@unileon.es (Enrique Alegre), victor.gonzalez@unileon.es (Victor González-Castro), l.fernandez@unileon.es (Laura Fernández-Robles)

Several works present methods for learning local descriptors as an alternative to supervised CNN, such as convolutional kernel networks (Paulin et al., 2015). Among all, the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) used within the Bag of Visual Words (BoVW) framework has been successfully applied for extracting features in image classification problems, either standalone or combined with other descriptors (Fidalgo et al., 2016, 2017).

The Bag of Visual Words (BoVW) model (Csurka et al., 2004) represents an image based on the frequency of the appearance of patches within it. Each patch is described by means of a feature vector, which is frequently obtained by means of SIFT algorithm. There is an important drawback in this framework: it discards the spatial order of local descriptors, which decreases the quality of the image description. To overcome this problem, Lazebnik et al. (Lazebnik et al., 2005) proposed an extension of BoVW, the Spatial Pyramid Matching (SPM), which divides the image into smaller subregions and computes the histograms of their local features.

BoVW or SPM descriptors computed on patches from the entire image contain information from both the object of interest, henceforth the foreground, and from the background. However, only the foreground is meaningful for a good description of the object of interest. Therefore, using features from the whole image may result in suboptimal classifications (Borji and Itti, 2013).

Visual Saliency, i.e., the subjective perceptual quality which makes some items to grab our attention immediately, is an effective way to deal with this drawback. Describing only the objects of interest within the image would boost the description methods and, therefore, improve recognition performance (Vig et al., 2012). For this reason, automatically selecting salient regions across the images can be very useful.

Modelling this visual saliency recently attracted the interest of the computer vision community (Mathe and Sminchisescu, 2015; Wang et al., 2016). Cheng et al. (2015) proposed a regional contrast-based salient object detection algorithm, consisting of assigning saliency values both to individual pixels and to local regions in the image based on colour separation. Coniglio et al. (2017) used saliency regions to estimate if a pixel belongs to the foreground or background with the final objective of people detection. Alexe et al. (2012) introduced the objectness measure, which is an object detector for generic classes that evaluates the likeliness of a window in the image to contain an object in contrast to the background. Selective search (Uijlings et al., 2013) tries to build object regions, instead of eliminating many windows from the set of candidate regions and it has become very popular since the supervised R-CNN detector (Girshick et al., 2014) uses it. Objectness belongs to the category of window scoring methods whereas selective search is considered as a grouping method. Tang et al. (2017) uses both approaches to generate saliency maps for object detection.

Hou et al. (2012) introduced a method that has been reported as one of the best algorithms for detecting the most significant saliency regions in images. It obtains the saliency maps of images that predict human fixation points with low computational cost. In their work they recommend σ, i.e., the standard deviation of the Gaussian kernel, which is called *blurring factor*, to be proportional to the size of the object of interest. They concluded that it was possible to choose a fixed value that worked well for most of the saliency maps algorithms they evaluated. However, the use of a fixed blurring factor does not always guarantee that the resulting saliency map contains only foreground information, which is not desirable when this saliency map is utilised in the BoVW framework. No matter which blurring factor we used, it was not possible to remove all background information.

In this paper, we presented three strategies based on the Hou et al. saliency map (Hou et al., 2012) to remove this background information at the region level. We initially proposed an efficient method to select automatically the blurring factor used in the Hue et al. saliency map. We named it *Automatic Level of Attention* or *AutoBlur* and is based on the concept of human fixation process.

Then, we proposed two variants of SARF method (Semantical Attention Region Filtering), which consists of labelling regions of the image as foreground or background regions by means of the segments obtained using Mean Shift (Comaniciu and Meer, 2002). The first method, called SARF based on the **I**ntersection of **S**aliency **M**aps (henceforth named *SARF-ISM*), starts by the generation of a binary saliency map, which we called "attention seeds", from the image. Then, it labels the regions resulting of the Mean Shift segmentation as foreground if their intersection with the attention seeds is not empty. The second one, called SARF based on **K**eypoint **V**oting (henceforth named *SARF-KV*), makes the decision based on the proportion of keypoints of the region with lower distance on a foreground or a background dictionary. These dictionaries are built using attention regions obtained by means of the previously described AutoBlur method.

The rest of the paper is organised as follows. In Section 2 image signature algorithm, the influence of its blurring factor σ and the automatic level of attention method that we propose are explained. Next, Section 3 describes the two approaches of our filter method, together with the segmentation algorithm used. The experimentation settings, datasets and the results of the experiments carried out to assess the proposed method are discussed in Section 4. Finally, the conclusions and future perspectives are presented in Section 5.

2. Semantic Attention Filtering at region level

In this section, we initially explain the saliency map algorithm (Hou et al., 2012) used for the experimentation. Next, we evaluate how the blurring factor of the image signature saliency map, σ, affects the feature extraction stage of the BoVW model, in terms of the accuracy obtained in the image classification task. Once the influence of the factor is confirmed, we propose the AutoBlur rule to calculate the level of attention for each image efficiently automatically. We call "level of attention" the binary output obtained after thresholding the saliency map calculated with a fixed blurring factor. This rule guarantees in most of the cases that the objects of interest are highlighted.

2.1. Image signature Saliency Map

Hou et al. (2012) proposed an algorithm which is based on the image signature, considering an image as the sum of a foreground and a background signals.

The authors defined the image signature, I_{sig}, as follows:

$$I_{sig} = sign(DCT(I)), \tag{1}$$

where DCT is the Discrete Cosine Transform and $sign$ the sign function.

Since the foreground of an image is visually conspicuous with respect to its background, its saliency map, SM_σ, is calculated by smoothing the squared reconstructed image (see Equation 2), I_{rec}, with a Gaussian Kernel, g, with standard deviation (i.e., blurring factor) σ. A Gaussian smoothing is necessary due to the spatially sparse result obtained for I_{rec}.

$$I_{rec} = IDCT(sign(DCT(I))) \tag{2}$$

$$SM_\sigma = g(\sigma) * (I_{rec} \circ I_{rec}) \tag{3}$$

The symbol $\circ$ stands for the Hadamard product, having the resulting image the same dimensions as I_{rec}.

2.2. Evaluation of the effect of the blurring factor

Hou et al. (2012) recommended the standard deviation, σ, of the Gaussian Kernel, also called blurring factor, to be proportional to the size of the object of interest. They demonstrated that it is possible to choose a fixed value that achieves excellent results for most of the methods, regarding saliency maps evaluation.

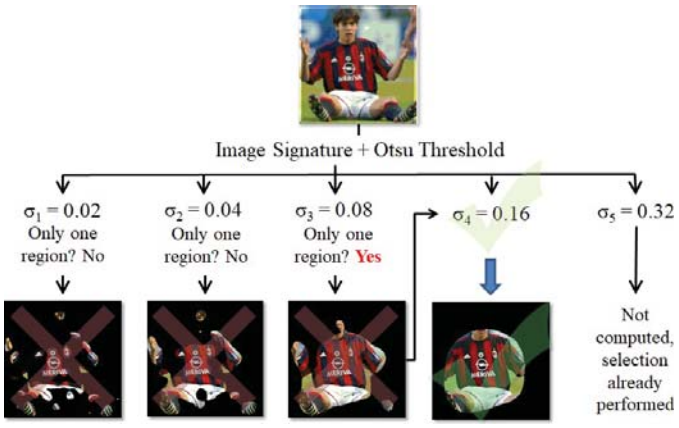

Fig. 1. AutoBlur overview. Resulting images after masking the original one with the binarized saliency maps obtained with different blurring factors. The maps show an increase in the level of details retained from the foreground information when increasing the blurring factor.

However, given the results of our experiments, we concluded that this is not valid in the context of image classification, instead of on saliency map evaluation. If the saliency map is used before image description, e.g. the BoVW model, to select where to extract the descriptors from, the classification performance does depend on the selected blurring factor.

Our objective was to evaluate how the blurring factor of the image signature algorithm affects the BoVW feature image classification when using different blurring factors. To do that, we assessed the following values of σ: 0.02, 0.04, 0.08, 0.16, 0.32 and 1.28. We selected these values following a sampling process that allows seeing the influence of different blurring factors in the image classification. Then, we binarized the resulting saliency map, Eq. (3), employing an Otsu thresholding (Otsu, 1979), so we obtained a binary image where the attention zone corresponds to white pixels. Later, this binary image is used to mask the original one.

$$SM_{bin} = Otsu(SM_\sigma) \tag{4}$$

As it is depicted in Figure 1, for the first chosen factor, $\sigma = 0.02$, the resulting saliency map selects the smallest and most salient regions whereas the last one displayed, $\sigma = 0.16$, causes the attention regions to contain most of the object of interest, and also part of the background.

If a descriptor is calculated from a pixel contained in the area of SM_{bin} (Eq. (4)), we labelled it as a foreground descriptor or a descriptor from an attention zone; otherwise, we considered it as a background descriptor. As further discussed in Section 4.3.1, the use of these foreground descriptors extracted from each resulting region affects the image classification accuracy.

2.3. Automatic selection of Blurring factor

Based on the previous demonstration about the influence of the blurring factor on the image classification results, and due to the different underlying foreground information, we tried to remove this dependency from the image signature algorithm while we keep a higher accuracy than the baseline.

We continued the analysis started in Figure 1 and we observed that the attention area of the image is less detached and more defined on the entire object of interest as the blurring factor increases. We can make an analogy between this observation and how the human brain and eyes work when they first glance at any fixed scene: first of all there are lot of small areas that draw the brain's attention, but the longer the eyes stare at the scene, the better the brain understands what exactly the scene represents. This is represented in Figure 1 when σ increases from 0.02 to 0.16.

Based on previous explanation, we propose a simple but effective rule to estimate the blurring factor for each image on a dataset. From now on we will refer to it as the *automatic level of attention* method or *AutoBlur*. First of all we compute, for each original image, the resulting saliency map (Eq. 3) with the blurring factor $\sigma = 0.02$ and binarize it. Then, if there are more than one binary attention region (Eq. 4) – i.e., more than one connected component in the image resulting of binarizing the saliency map – we repeat the process with a blurring factor of 2σ. When we find a binary saliency map with only one attention region, we select as the optimal one the double of it. We experimented with both the blurring factor for one attention region and the next factor, and we obtained better accuracy in image classification using the later because it keeps some information that is relevant for classifying each object. In the example of the Figure 1, with $\sigma = 0.08$ there is only a binary attention region, so we select $\sigma = 0.16$ as the recommended one for this particular image. In other words, we consider that

with $\sigma = 0.16$, brain and eyes attention are focused in a region that mainly contains the objects of interest.

We consider that AutoBlur works well for fine-grained image classification, where commonly there is only an object of interest per image and the object does not present slim parts that can lead to be considered as several objects. If the images contain several objects of interest, the proposed method can be applied on regions of the images comprising single objects. The different regions in each image could be located using any method of proposal regions that is effective for the images at hand, what is out of the scope of this paper.

In Section 4 we will discuss the results after the application of this rule to the different datasets used in the experimentation.

3. SARF: Semantic Attention Region Filtering

AutoBlur serves as a reference to our two region filtering proposals: we name them SARF (Semantic Attention Region Filtering). We first introduce the Mean Shift segmentation, which will be used together with AutoBlur method in both approaches, and then we will go into details for each of the proposed SARF methods.

3.1. Mean Shift segmentation

One of the steps of our region filtering methods is the image segmentation by means of the algorithm proposed by Comaniciu and Meer (2002) based on Mean Shift (Fukunaga and Hostetler, 1975). Mean Shift is a non-parametric feature-space analysis technique for locating the maxima of a density function. Its application in the clustering domain makes possible to segment an image into regions.

3.2. SARF based on Intersection of Saliency Maps (SARF-ISM)

As we have already discussed, one of the features of the saliency map proposed by Hou et al. (2012) is the capability to predict the human fixation points on a scene. In initial tests, we have observed that the binary saliency maps (Eq. 4) obtained with $\sigma = 0.02$ can be used as "attention seeds".

For each original image, we compute (i) the Mean Shift segmentation output, (ii) the attention seeds (i.e., the mentioned saliency map with $\sigma = 0.02$) and (iii) the binary saliency map

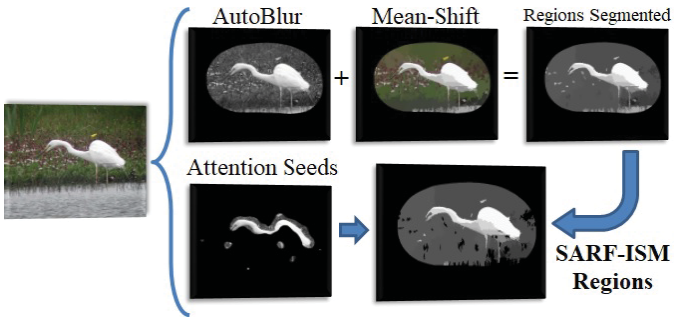

Fig. 2. SARF-ISM overview. Features are initially extracted from all the image and then labelled according to their correspondence to the overlapping areas between the attention seeds and the Mean Shift regions.

obtained using the AutoBlur method (Eq. 4). These three images are the inputs to the SARF-ISM method. Thereafter, we first mask the image segmented by Mean Shift with the binary saliency map obtained using the AutoBlur method. Then, we perform an additional masking process in which we discard the regions that do not overlap with the attention seeds in at least one pixel. The features belonging to the remaining regions, i.e. attention regions, are coded into the BoVW (see Eq. 5).

$$\forall Reg \in SM_{AB} \begin{cases} Reg \cap SM_{0.02} \neq \emptyset \rightarrow (d_f, kp_f)_{Reg} \in BoVW \\ Reg \cap SM_{0.02} = \emptyset \rightarrow (d_f, kp_f)_{Reg} \notin BoVW \end{cases}, \quad (5)$$

where Reg stands for each of the segmented regions within the binary saliency map SM_{AB} obtained from AutoBlur, and $SM_{0.02}$ is the "attention seeds", i.e., the image signature obtained with a blurring factor $\sigma = 0.02$. $(d_f, kp_f)_{Reg}$ indicates the initial foreground descriptors, and their corresponding keypoints, which belong to the analysed region.

3.3. SARF based on Keypoint Voting (SARF-KV)

The inputs to this method are the results of the Mean Shift segmentation and the binary saliency map resulting of the AutoBlur method.

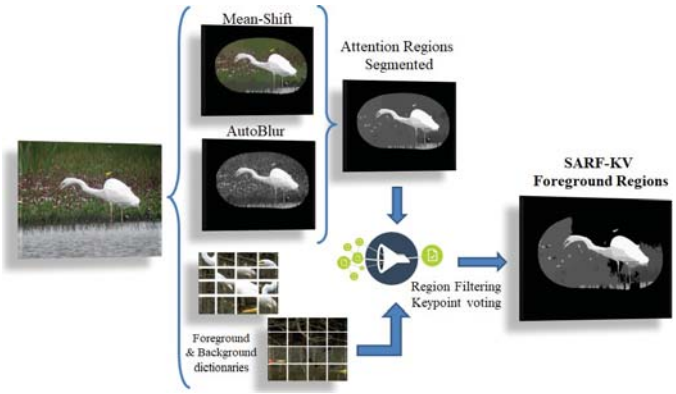

Fig. 3. SARF-KV overview. Regions are filtered after a voting system with the foreground and background labelled regions.

The binary saliency map (Eq. 4) obtained with the σ calculated with AutoBlur yields an estimation of the foreground and background of the image. It is possible to compute a dictionary for the foreground and another one for the background, which are used in a later stage of the method.

For each region in the image where the Mean Shift segmentation and AutoBlur outputs are combined, the distances between the descriptors, i.e. dense SIFT, of its points and (a) the foreground and (b) background dictionaries are computed. If the majority of the descriptors are closer to the foreground, the region is considered as a foreground region , i.e., an *attention region*. Otherwise, it is discarded.

Finally a BoVW is computed from the remaining foreground regions and the experiments are carried out in the same way as the one explained for SARF-ISM.

4. Results discussion

4.1. Datasets used in the experiments

In Figure 4, some samples of each of the datasets described in Table 1 are shown.

Fig. 4. In columns, Birds (1^{st}); Flowers (2^{nd}); Soccer (3^{rd}); ImageNet-7Arthropods (4^{th}) and ImageNet-6Weapons (5^{th}).

In Birds (Lazebnik et al., 2005) and Flowers (Nilsback and Zisserman, 2006), it can be observed how they are surrounded by different types of environment, which can confuse the attention areas because sometimes the object of interest shares some colours with the background.

Table 1. Features of the dataset used in the experimentation. Soccer training and test sets are chosen as in (van de Weijer and Schmid, 2006).

DATASETS	Images	Classes	Training/Test proportion
Birds	600	6	75/25
Flowers	1360	17	75/25
Soccer	280	7	62.5/37.5
ImageNet-7Arthropods	1400	7	75/25
ImageNet-6Weapons	4500	6	75/25

In the case of the Soccer dataset (van de Weijer and Schmid, 2006), despite being the smallest one, it is quite challenging because the teams of the soccer players are sometimes mixed between them and with the public or the referee.

We also wanted to assess these methods with bigger fine-grained datasets, so we generated ImageNet-7Arthropods and ImageNet-6Weapons from different synsets from the ImageNet collection. The first set reproduces a similar but wider environment than Birds and Flowers. In the second, ImageNet-6Weapons, we generated a noisy and unbalanced dataset whose categories are considered a challenging and exciting topic on Tor Darknet (Wesam et al., 2017). The groups represent the most frequent types of weapons sold in Tor, and they were selected after a visual inspection of the TOr Image Categories (TOIC) dataset content proposed by Fidalgo et al. (2017)

4.2. Experimental Setup

We have made publicly available the source code made of our proposed method and the experiments using BoVW, Sc-SPM frameworks[1] and CNN features[2].

For all our experiments we have used MATLAB and Python 3, and as a reference for comparison, i.e. baseline implementation, the BoVW model (Csurka et al., 2004) with dense SIFT descriptors (Lowe, 2004) extracted from the whole image.

Dense SIFT descriptors have been computed using the VLFeat library (Vedaldi and Fulkerson, 2010), with step and size set to 7. They have been clustered in a dictionary of 2048 words using K-means with the approximate nearest neighbour algorithm (Lloyd, 1982) . A hard assignment approach (Csurka et al., 2004) has been used to build the BoVW feature vectors that represent each image.

To carry out the Mean Shift segmentation on MATLAB, we have used the EDISON code[3] implementation of Comaniciu's method (Comaniciu and Meer, 2002) with a MEX wrapper from Shai Bagon[4]. We fixed for our experiments the spatial bandwidth $h_s = 30$, the ranging bandwidth $h_r = 5$ and the minimum size of the final output regions in pixels $M = 40$. We set these parameters to get small segmented regions in the object of interest that allow us to test the effectiveness of our strategy.

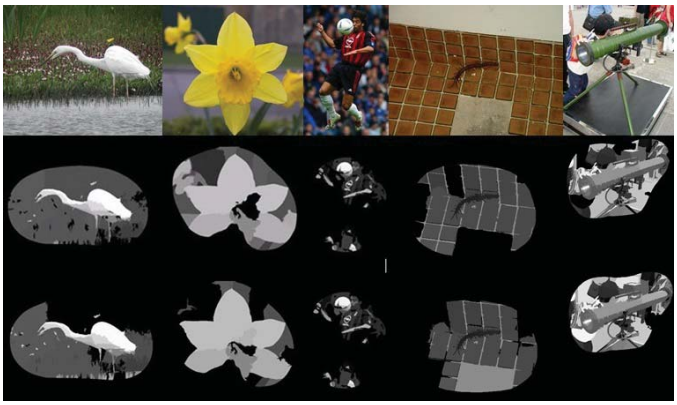

Fig. 5. Original images (1^{st} row), SARF-ISM (2^{nd} row) and SARF-KV (3^{rd} row) results in Birds, Flowers, Soccer, ImageNet-7Arthropods and ImageNet-6Weapons

To apply our filtering techniques on Spatial Pyramid Matching (SPM) and with Convolutional Neural Networks (CNN) features, we have used the following settings. In SPM we used the code released by Yang et al. (2009), and we included our filtering strategies, i.e. AutoBlur, SARF-ISM and SARF-KV, at the feature extraction step. We adapted the descriptor step, size and dictionary size to 7, 7 and 2048 to have a baseline in SPM similar to the one obtained in BoVW (Fidalgo et al., 2016), and we did not modify the rest of parameters. In our experimentation with CNN features, we used the VGG16 model pre-trained on the ImageNet challenge Russakovsky et al. (2015), which is publicly available in Keras[5].

We have used the five different image datasets indicated in Table 1 and a Support Vector Machine (SVM) (Cherkassky, 1997; Suykens and Vandewalle, 1999) with a linear kernel. Figure 5 visually presents the output from both SARF methods for

[1] https://es.mathworks.com/matlabcentral/fileexchange/67369-autoblur-and-sarf-filtering-techniques-applied-to-bag-of-visual-words-and-spatial-pyramid-frameworks

[2] https://github.com/efidalgo/AutoBlur_CNN_Features

[3] http://coewww.rutgers.edu/riul/research/code/EDISON/

[4] http://www.wisdom.weizmann.ac.il/ bagon/matlab.html

[5] https://keras.io/applications/

the same image. In all three frameworks, i.e. BoVW, SPM and CNN features, we have splitted each dataset randomly into disjoint training and test sets with the proportions indicated in Table 1. This process has been repeated five times to avoid the possible effects of random sub-sampling, and the presented results are the average results of the previous experiments.

4.3. Results discussion

In the next subsections, we discuss the results obtained with different values of the blurring factor on the image signature and the BoVW framework for image classification. Then we explain in detail the performance of the three filtering techniques applied to BoVW, and finally, we extend our discussion to the SPM and the initial experimentation made on CNN.

4.3.1. Results obtained with different blurring factors

As it is depicted in Table 2, the accuracy strongly depends on the σ value used to obtain the binary saliency map — Eq. (4).

Table 2 shows that the best accuracy is achieved with the following blurring parameters, σ, on each dataset: Soccer: σ = 0.04, Birds: σ = 0.04, Flowers: σ = 0.16, ImageNet-6Weapons: σ = 0.16 and ImageNet-7Arthropods: σ = 0.08. Hou et al. (2012) indicated that the optimal blurring factor σ is quite stable across different saliency maps algorithms. That assertion was only true because they were focused on a saliency map evaluation, but not on an image classification task, as we demonstrated in our experiments with the different blurring factors. Therefore, the affirmation made by Hou et al. (2012) does not apply in the context of this work.

4.3.2. AutoBlur and SARF performance on BoVW framework

In Table 3 it is shown how the accuracy obtained with the AutoBlur method for these datasets boosts the baseline one, i.e. from 48.00±3.66% to 51.81±4.79%, 67.47±2.96% to 82.40±0.89%, 62.24±3.53% to 67.82±1.05%, 38.29±1.63% to 46.80±2.16% and 55.06±2.21% to 56.52±1.37% in Soccer, Birds, Flowers, ImageNet-7Arthropods and ImageNet-6Weapons respectively. Baseline results were computed in Fidalgo et al. (2016) with the configuration described in 4.2.

The selection of an adequate blurring factor allows to discard mainly background descriptors, resulting in a richer dictionary that retains more foreground information about the classes.

Fig. 6. Samples of Autoblur on Birds (up) and Soccer (down). Attention zones are not mainly focused on a bird, neither in a single team on an image.

Regarding both SARF approaches, they also boost the Baseline results in the five datasets, except the SARF-KV strategy on ImageNet-6Weapons. SARF-ISM method obtains the highest growth over the baseline in Soccer and ImageNet-6Weapons, with 6.10 and 0.30 points of improvement, while SARF-KV gets 14.40 and 6.05 points on Birds and Flowers datasets. In Soccer and ImageNet-6Weapons datasets, even with a small value of the blurring factor, the attention seeds points the relevant features of the main objects, that is why SARF-ISM obtains slightly better results than SARF-KV.

In Table 3 it can be checked how SARF-KV does not outperform the baseline results in ImageNet-6Weapons.

Fig. 7. AutoBlur output in ImageNet-6Weapons (1^{st} Row) and ImageNet-7Arthropods (2^{nd} row), five samples per Dataset

In the first example of Figure 7, i.e. red squared, the object of interest is entirely highlighted due to the absence of distracting background. But in the rest, the objects of interest contained in the images do not represent a significant part of the image itself and the introduction of other potential saliency regions, e.g. persons and environment, causing SARF-KV fails in boosting the Baseline results on ImageNet-6Weapons.

AutoBlur selects a blurring factor for each individual image and does not need a training set to compute it. Despite it does not achieve the best results in all the experiments on the datasets analysed, it guarantees a higher accuracy than when the BoVW descriptors are calculated upon the whole image, getting rid of the need of using several blurring factors.

Nevertheless, we have observed that for some images, the AutoBlur rule does not always manage to focus the soccer player, bird or flower into the main attention area, as it can be noticed in Figure 6.

4.3.3. AutoBlur and SARF performance on SPM and CNN framework

Once we validated our filtering techniques in the BoVW framework, we evaluated their effectiveness in the Spatial Pyramid Matching framework, a BoVW extension that uses the spatial order of local descriptors. After confirming the good performance in SPM, we also included in this paper initial experimentation in the field of CNN. We selected the resulting bounding box over the region depicted by AutoBlur, Eq.(4), and we fed the CNN with the resulting image after cropping the original image with this bounding box. CNN features are extracted from the last layer of the VGG16 pre-trained network and used together with the SVM classifier with a lineal kernel. Both SPM and CNN results are presented in Tables 4 and 5.

From Table 4 we can observe how the SPM framework outperforms the Baseline results obtained with the BoVW (see Table 3) in all the datasets, except the Soccer one. Regardless its small size, Soccer dataset represents a challenging situation

Table 2. Summary of the results obtained in the five datasets using image signature saliency map with different blurring factors. Baseline results (⋆) from Fidalgo et al. (2016). In bold, the best accuracy per dataset

DATASETS	Baseline	σ = 0.02	σ = 0.04	σ = 0.08	σ = 0.16	σ = 0.32	σ = 1.28
Soccer	(⋆)48.00±3.66%	55.23±6.56%	**58.48±3.60%**	53.90±1.59%	52.19±2.64%	51.62±5.96%	49.52±4.52%
Birds	(⋆)67.47±2.96%	81.47±0.87%	**81.87±1.66%**	79.07±4.21%	80.4±1.46%	75.87±0.73%	72.93±2.52%
Flowers	(⋆)62.24±3.53%	58.29±4.52%	60.94±3.65%	64.65±2.8%	**68.00±2.06%**	68.35±2.09%	67.65±2.08%
ImageNet-6Weapons	55.06±2.21%	52.42±1.31%	55.29±1.61%	56.86±1.76%	**58.60±1.63%**	56.64±2.44%	56.38±2.05%
ImageNet-7Arthropods	38.29±1.63%	43.94±1.15%	46.34±1.28%	**49.20±2.90%**	48.17±2.07%	47.31±1.95%	44.91±2.91%

Table 3. Summary of the results obtained in Bag of Visual Words framework with AutoBlur and both SARF approaches . Baseline results (⋆) from Fidalgo et al. (2016)

DATASETS	Baseline	AutoBlur	SARF-ISM	SARF-KV
Soccer	(⋆)48.00±3.66%	**51.81±4.79%**	**54.10±4.92%**	**51.81±4.29%**
Birds	(⋆)67.47±2.96%	**82.40±0.89%**	**81.73±3.42%**	**81.87±1.79%**
Flowers	(⋆)62.24±3.53%	**67.82±1.05%**	**67.06±3.74%**	**68.29±3.03%**
ImageNet-7Arthropods	38.29±1.63%	**46.80±2.16%**	**46.29±1.95%**	**47.71±1.46%**
ImageNet-6Weapons	55.06±2.21%	**56.52±1.37%**	**55.36±1.84%**	54.07±1.90%

Table 4. Results obtained in Spatial Pyramid Matching framework. Baseline represents the use of the descriptors from all the image.

DATASETS	Baseline	AutoBlur	SARF-ISM	SARF-KV
Soccer	45.91±%	**54.28±3.08%**	**56.00±4.38%**	**54.85±4.12%**
Birds	91.20±%	**93.60±2.43%**	**94.13±3.10%**	**93.60±3.41%**
Flowers	74.88±%	**75.53±2.19%**	**75.88±2.18%**	**75.88±2.42%**
ImageNet-6Weapons	68.68±%	67.00±2.35%	67.61±2.31%	67.88±2.20%
ImageNet-7Arthropods	66.11±%	**66.74±3.02%**	66.05±1.33%	**66.45±2.70%**

for SIFT descriptors, since most of the information to identify the team is present into the colour of the clothes rather than the shape or spatial distribution of the vectors. AutoBlur and SARF filtering strategies outperform the baseline results, i.e. all descriptors used, in four out of five of the datasets analysed. On the one hand, SARF-ISM obtains the best performance in Soccer, Birds and Flowers datasets, boosting the baseline results in 10.09, 2.93 and 1.00 points respectively. On the other hand, AutoBlur obtains the best performance in ImageNet-7Arthropods with 0.73 points higher than the baseline results. Sadly, our filtering strategies fail to improve the ImageNet-6Weapons baseline, maybe due to the behaviour of the automatic blurring factor selection which retained information not belonging to the object of interest, as Figure 7 depicts.

Finally, Table 5 presents the results with CNN features, outperforming the BoVW and SPM frameworks, as expected. However, the purpose of the filtering techniques presented in this paper is the selection of relevant information for image classification. Despite CNN usually produces the state-of-the-art in feature extraction for image classification, the application of AutoBlur has a positive impact, increasing the accuracy in two of the datasets analysed, Soccer and Flowers with 1.76 and 2.45 points respectively.

5. Conclusions and future work

Saliency maps are still useful tools to highlight the relevant parts of images. They are especially relevant in cases in which the image collection is not big enough to ensure that Deep Learning techniques can be properly trained.

Table 5. Results obtained with CNN features. Results obtained with CNN features. Baseline: full image fed into the CNN. AutoBlurCNN: combination of CNN and AutoBlur.

DATASETS	Baseline	AutoBlurCNN
Soccer	58.29±%	**60.05±4.24%**
Birds	95.23±%	94.17±1.67%
Flowers	87.06±%	**89.51±1.61%**
ImageNet-7Arthropods	80.59±%	79.48±1.07%
ImageNet-6Weapons	84.39±%	83.74±1.34%

The output of the saliency map used in this work, i.e. saliency signature, depends on a blurring factor which can be modified. First, we demonstrated, on the five datasets analysed, that if we use the features from the image signature in the BoVW framework with different blurring parameters, the differences in accuracy are remarkable. To overcome the dependence on this factor, we proposed a basic but effective automatic selection of attention level that we named AutoBlur. We compared its results against the baseline results, i.e. standard BoVW framework over five datasets, yielding higher accuracies in all cases.

The previous filtering strategy highlights the importance of extracting suitable descriptors from the images for dictionary construction. Later on, we pushed our proposal a step further, and we introduced Semantic Attention Region Filtering (SARF) with two variants. The first one, SARF based on the Intersection of Saliency Maps (SARF-ISM), overlaps an image segmented using Mean Shift with the binary saliency map obtained with a blurring factor 0.02, what we called the attention seeds. Each region that shares a common area with the attention seeds remains as foreground or attention regions. Otherwise, it is discarded and labelled as a background region. The second version of our method, SARF based on Keypoint Voting (SARF-KV) analyses the distance between the descriptors of the key points belonging to each segmented region and both foreground-background dictionaries. These dictionaries are previously constructed with the resulting attention regions obtained with the automatic level of attention method. When the distance of a descriptor to the foreground dictionary is lower than the distance to the background one, a positive vote is counted. The number of positive and negative votes determines if the region belongs to the foreground or background, respectively. Both methods have been tested using five datasets.

Once the effectiveness of SARF and AutoBlur has been validated on the BoVW framework, we also applied them on the Spatial Pyramid Matching (SPM), and we made an initial approach about how they could be applied together with Convolutional Neural Network (CNN) features. In SPM, the filter-

ing strategies outperformed the baseline results in four out of the five datasets analysed, confirming that our strategies can be applied successfully to this framework. Finally, after cropping from the original image a bounding box equivalent to the resulting AutoBlur region, we demonstrated that our AutoBlur strategy could be used to improve the results obtained with CNN features extracted on the whole images in two datasets.

SARF and AutoBlur methods guarantee better accuracy than the baseline in the BoVW framework, and in several scenarios on SPM and CNN features, and they do not require optimisation of the saliency map parameters. For this reason, these filtering strategies are robust options to improve image classification results with BoVW, SPM and CNN features.

Further investigations will be focused on improving the blurring factor selection method and deal with situations like the one presented in ImageNet-6Weapons, i.e. objects of interest that do not represent a significant part of the image. The images from the ImageNet-6Weapons dataset will be used to train a model in such specific content that will be able to detect Weapon domains or Marketplaces in the Tor Network. The fact that AutoBlur strategy combined with CNN features, AutoBlur-CNN, outperformed the established Baseline in two datasets, together with the close results between Baseline and AutoBlur-CNN results ($\leqslant$ 1.1 points), encourages us to explore in future works how to improve the combination of AutoBlur with the CNN feature extraction, together with both SARF approaches.

Acknowledgments

This research is supported by the INCIBE grant "INCIBEC-2015-02493" corresponding to the "Ayudas para la Excelencia de los Equipos de Investigación avanzada en ciberseguridad" and also by the framework agreement between the University of León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 22. We gratefully acknowledge the support of Nvidia Corporation for their kind donation of GPUs (GeForce GTX Titan Xp and K-40) that were used in this work.

References

Alexe, B., Deselaers, T., Ferrari, V., 2012. Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 2189–2202. doi:10.1109/TPAMI.2012.28.

Borji, A., Itti, L., 2013. State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 185–207.

Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M., 2015. Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 569–582.

Cherkassky, V., 1997. The Nature Of Statistical Learning Theory˜. volume 8. Springer New York, New York, NY.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619.

Coniglio, C., Meurie, C., Lézoray, O., Berbineau, M., 2017. People silhouette extraction from people detection bounding boxes in images. Pattern Recognition Letters 93, 182–191.

Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. IN WORKSHOP ON STATISTICAL LEARNING IN COMPUTER VISION, ECCV , 1—-22.

Fidalgo, E., Alegre, E., González-Castro, V., Fernández-Robles, L., 2016. Compass radius estimation for improved image classification using Edge-SIFT. Neurocomputing 197, 119–135.

Fidalgo, E., Alegre, E., González-Castro, V., Fernández-Robles, L., 2017. Illegal activity categorisation in DarkNet based on image classification using CREIC method. 10th International Conference on Computational Intelligence in Security for Information Systems I, 600–609.

Fukunaga, K., Hostetler, L., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory 21, 32–40.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. doi:10.1109/CVPR.2014.81.

Hou, X., Harel, J., Koch, C., 2012. Image signature: Highlighting sparse salient regions. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 194–201.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems, Curran Associates Inc., USA. pp. 1097–1105.

Lazebnik, S., Schmid, C., Ponce, J., 2005. A maximum entropy framework for part-based texture and object recognition. Proceedings of the IEEE International Conference on Computer Vision I, 832–838.

Lloyd, S.P., 1982. Least Squares Quantization in PCM. IEEE Transactions on Information Theory 28, 129–137.

Lowe, D.G., 2004. Distinctive image features from scale invariant keypoints. Int'l Journal of Computer Vision 60, 91–11020042.

Mathe, S., Sminchisescu, C., 2015. Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 1408–1424.

Nilsback, M.E., Zisserman, A., 2006. A visual vocabulary for flower classification, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1447–1454.

Otsu, N., 1979. A threshold selection method from Gray-level. IEEE Transactions on Systems, Man, and Cybernetics SMC-9, 62–66.

Paulin, M., Douze, M., Harchaoui, Z., Mairal, J., Perronin, F., Schmid, C., 2015. Local convolutional features with unsupervised training for image retrieval, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 91–99. doi:10.1109/ICCV.2015.19.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 211–252.

Suykens, J.A.K., Vandewalle, J., 1999. Least Squares Support Vector Machine Classifiers. Neural Processing Letters 9 9, 293–300.

song Tang, X., Hao, K., Wei, H., Ding, Y., 2017. Using line segments to train multi-stream stacked autoencoders for image classification. Pattern Recognition Letters 94, 55–61.

Tang, Y., Wang, X., Dellandréa, E., Chen, L., 2017. Weakly supervised learning of deformable part-based models for object detection via region proposals. IEEE Transactions on Multimedia 19, 393–407. doi:10.1109/TMM.2016.2614862.

Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M., 2013. Selective search for object recognition. International Journal of Computer Vision 104, 154–171.

Vedaldi, A., Fulkerson, B., 2010. Vlfeat. Proceedings of the international conference on Multimedia - MM '10 3, 1469.

Vig, E., Dorr, M., Cox, D., 2012. Space-variant descriptor sampling for action recognition based on saliency and eye movements. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7578 LNCS, 84–97.

Wang, J., Borji, A., Kuo, C.C.J., Itti, L., 2016. Learning a Combined Model of Visual Saliency for Fixation Prediction. IEEE Transactions on Image Processing 25, 1566–1579.

van de Weijer, J., Schmid, C., 2006. Coloring local feature extraction, in: Leonardis, A., Bischof, H., Pinz, A. (Eds.), Computer Vision – ECCV 2006, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 334–348.

Wesam, M., Nabki, A., Fidalgo, E., Alegre, E., De Paz, I., 2017. Classifying Illegal Activities on Tor Network Based on Web Textual Contents. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 1, 35–43.

Yang, J., Yu, K., Gong, Y., Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1794–1801.