Have a SNAK. Encoding Spatial Information with the Spatial Non-alignment Kernel

Radu Tudor Ionescu(✉) and Marius Popescu

University of Bucharest, No. 14 Academiei Street, Bucharest, Romania
{raducu.ionescu,popescunmarius}@gmail.com

Abstract. The standard bag of visual words model model ignores the spatial information contained in the image, but researchers have demonstrated that the object recognition performance can be improved by including spatial information. A state of the art approach is the spatial pyramid representation, which divides the image into spatial bins. In this paper, another general approach that encodes the spatial information in a much better and efficient way is described. The proposed approach is to embed the spatial information into a kernel function termed the Spatial Non-Alignment Kernel (SNAK). For each visual word, the average position and the standard deviation is computed based on all the occurrences of the visual word in the image. These are computed with respect to the center of the object, which is determined with the help of the objectness measure. The pairwise similarity of two images is then computed by taking into account the difference between the average positions and the difference between the standard deviations of each visual word in the two images. In other words, the SNAK kernel includes the spatial distribution of the visual words in the similarity of two images. Furthermore, various kernel functions can be plugged into the SNAK framework. Object recognition experiments are conducted to compare the SNAK framework with the spatial pyramid representation, and to assess the performance improvements for various state of the art kernels on two benchmark data sets. The empirical results indicate that SNAK significantly improves the object recognition performance of every evaluated kernel. Compared to the spatial pyramid, SNAK improves performance while consuming less space and time. In conclusion, SNAK can be considered a good candidate to replace the widely-used spatial pyramid representation.

Keywords: Kernel method · Spatial information · Bag of visual words

1 Introduction

Computer vision researchers have recently developed sophisticated methods for object class recognition, image retrieval and related tasks. Among the state of the art models are discriminative classifiers using the *bag of visual words* (BOVW) representation [18,20] and spatial pyramid matching [12], generative models [6] or part-based models [11]. The BOVW model, which represents an image as a

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part I, LNCS 9279, pp. 97–108, 2015.
DOI: 10.1007/978-3-319-23231-7_9

histogram of local features, has demonstrated impressive levels of performance for image categorization [20], image retrieval [15], or related tasks. The standard bag of words model ignores spatial relationships between image features, but researchers have demonstrated that the performance can be improved by including spatial information [12,16,19].

This work presents a novel approach to include spatial information in a simple and effective manner. The proposed approach is to embed the spatial information into a kernel function termed the *Spatial Non-Alignment Kernel*, or SNAK for short. The proposed kernel works by including the spatial distribution of the visual words in the similarity of two images. For each visual word in an image, the average position and the standard deviation is computed based on all the occurrences of the visual word in that image. These statistics are computed with respect to the center of the object, which is determined with the help of the objectness measure [1]. Then, the pairwise similarity of two images can be computed by taking into account the distance between the average positions and the distance between the standard deviations of each visual word in the two images. This simple approach has two important advantages. First of all, the feature space increases with a constant factor, which means that it uses less space than other state of the art approaches [12]. Second of all, the SNAK framework can be applied to various kernel functions, thus being a rather general approach. Object recognition experiments are conducted in order to assess the performance of different kernels based on the SNAK framework versus the spatial pyramid framework, on two benchmark data sets of images, more precisely, the Pascal VOC data set and the Birds data set. The performance of the kernels is evaluated for various vocabulary dimensions. In all the experiments, the SNAK framework shows a better recognition accuracy compared to the spatial pyramid.

The paper is organized as follows. Related work on frameworks for including spatial information is discussed in Section 2. The Spatial Non-Alignment Kernel is described in Section 3. All the experiments are presented in Section 4. Finally, the conclusions are drawn in Section 5.

2 Related Work

Several approaches of adding spatial information to the BOVW model have been proposed [9,10,12,16,19]. The spatial pyramid [12] is one of the most popular frameworks of using the spatial information. In this framework, the image is gradually divided into spatial bins. The frequency of each visual word is recorded in a histogram for each bin. The final feature vector for the image is a concatenation of these histograms. To reduce the dimension of the feature representation induced by the spatial pyramid, researchers have tried to encode the spatial information at a lower level [9,16]. Spatial Coordinate Coding scheme [9] applies spatial location and angular information at descriptor level. The authors of [10] model the spatial location of the image regions assigned to visual words using Mixture of Gaussians models, which is related to a soft-assign version of the spatial pyramid representation. A similar approach is proposed in [16], but the

change is made at the low level feature representation, enabling the model to be extended to other encoding methods. It is worth mentioning that in [10], the spatial mean and the variance of image regions associated with visual words are used to define a Mixture of Gaussians model. In the SNAK framework, the spatial mean and the standard deviation of visual words are also used, but in a completely different way, by embedding them into a kernel function. Another way of using spatial information is to consider the location of objects in the image, which can be determined either by using manually annotated bounding boxes [19] or by using the objectness measure [8,16].

3 Spatial Non-Alignment Kernel

A simple yet powerful framework for including spatial information into the BOVW model is presented next. This framework is termed *Spatial Non-Alignment Kernel* (SNAK) and it is based on measuring the spatial non-alignment of visual words in two images using a kernel function. In the SNAK framework, additional information for each visual word needs to be stored first in the feature representation of an image. More precisely, the average position and the standard deviation of the spatial distribution of all the descriptors that belong to a visual word are computed. These statistics are computed independently for each of the two image coordinates. The SNAK feature vector includes the average coordinates and the standard deviation of a visual word together with the frequency of the visual word, resulting in a feature space that is 5 times greater than the original feature space corresponding to the histogram representation. The size of the feature space is identical to a spatial pyramid based on two levels, but it is roughly 4 times smaller than a spatial pyramid based on three levels.

Let U represent the SNAK feature vector of an image. For each visual word at index i, U will contain 5-tuples as defined below:

$$u(i) = \left(h^u(i), m_x^u(i), m_y^u(i), s_x^u(i), s_y^u(i)\right).$$

The first component of $u(i)$ represents the visual word's frequency. The following two components ($m_x(i)$ and $m_y(i)$) represent the mean (or average) position of the i-th visual word on each of the two coordinates x and y, respectively. The last two components ($s_x(i)$ and $s_y(i)$) represent the standard deviation of the i-th visual word with respect to the two coordinates x and y. If the visual word i does not appear in the image ($h^u(i) = 0$), the last four components are undefined. In fact, these four values are not being used at all, if $h^u(i) = 0$.

Using the above notations, the SNAK kernel between two feature vectors U and V can be defined as follows:

$$k_{\text{SNAK}}(U,V) = \sum_{i=1}^{n} \exp\left(-c_1 \cdot \Delta_{mean}(u(i), v(i))\right) \cdot \exp\left(-c_2 \cdot \Delta_{std}(u(i), v(i))\right), \tag{1}$$

where n is the number of visual words, c_1 and c_2 are two parameters with positive values, $u(i)$ is the 5-tuple corresponding to the i-th visual word from U, $v(i)$ is the 5-tuple corresponding to the i-th visual word from V, and Δ_{mean} and Δ_{std} are defined as follows:

$$\Delta_{mean}(u,v) = \begin{cases} \left(m_x^u - m_x^v\right)^2 + \left(m_y^u - m_y^v\right)^2, \text{ if } h^u, h^v > 0 \\ \infty, \text{ otherwise} \end{cases}$$

$$\Delta_{std}(u,v) = \begin{cases} \left(s_x^u - s_x^v\right)^2 + \left(s_y^u - s_y^v\right)^2, \text{ if } h^u, h^v > 0 \\ \infty, \text{ otherwise} \end{cases}$$

where m_x, m_y, s_x, and s_y are components of the 5-tuples u and v. If a visual word does not appear in at least one of the two compared images, its contribution to k_{SNAK} is zero, since Δ_{mean} and Δ_{std} are infinite.

It can be easily demonstrated that SNAK is a kernel function. Indeed, the proof that k_{SNAK} is a kernel comes out immediately from the following observation. For a given visual word i and two 5-tuples u and v, the equations below represent two RBF kernels:

$$\exp\left(-c_1 \cdot \Delta_{mean}(u(i), v(i))\right)$$
$$\exp\left(-c_2 \cdot \Delta_{std}(u(i), v(i))\right),$$

and their product is also a kernel. By summing up the RBF kernels corresponding to all the 5-tuples inside the SNAK feature vectors U and V, the k_{SNAK} function is obtained. From the additive property of kernel functions [17], it results that k_{SNAK} is also a kernel function.

An interesting remark is that k_{SNAK} can be seen as a sum of separate kernel functions, each corresponding to a visual word that appears in both images. This is a fairly simple approach, that can be easily generalized and combined with many other kernel functions. The following equation shows how to combine SNAK with another kernel k^* that takes into account the frequency of visual words:

$$k^*_{\text{SNAK}}(U,V) = \sum_{i=1}^{n} k^*(h^u(i), h^v(i)) \cdot \\ \cdot \exp\left(-c_1 \cdot \Delta_{mean}(u(i), v(i))\right) \cdot \exp\left(-c_2 \cdot \Delta_{std}(u(i), v(i))\right). \quad (2)$$

Equation (2) can be used to combine SNAK with other kernels at the visual word level, individually. Certainly, using the above equation, SNAK can be combined with kernels such as the linear kernel, the Hellinger's kernel, or the intersection kernel. Moreover, being a kernel function, SNAK can be combined with any other kernel using various approaches specific to kernel methods, such as multiple kernel learning [7].

3.1 Translation and Size Invariance

Intuitively, the SNAK kernel measures the distance between the average positions of the same visual word in two images. SNAK can be used to encode spatial information for various classification tasks, but some improvements based on

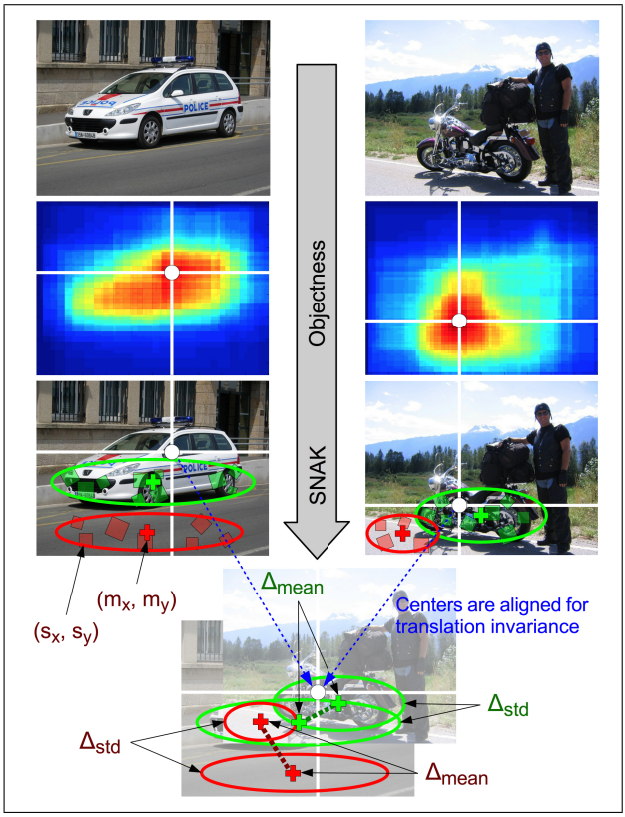

Fig. 1. The spatial similarity of two images computed with the SNAK framework. First, the center of mass is computed according to the objectness map. The average position and the standard deviation of the spatial distribution of each visual word are computed next. The images are aligned according to their centers, and the SNAK kernel is computed by summing the distances between the average positions and the standard deviations of each visual word in the two images.

task-specific information are possible. One such example is object class recognition. If the objects appear in roughly the same locations, the SNAK approach would work fine. However, this restriction may be often violated in practice. Any object can appear in any part of the image, and a visual word describing some part of the object can therefore appear in a different location in each image. Due to this fact, SNAK is not invariant to translations of the object. If the object's location in each image is known a priori, the average position of the visual word can be computed with respect to the object's location, by translating the origin of the coordinate system over the center of the object. The exact location of the object is not known in practice, but it can be approximated using the *objectness* measure [1]. This measure quantifies how likely it is for an image window to contain an object. By sampling a reasonable number of windows and by accumulating their probabilities, a pixelwise objectness map of the image can be

produced. The objectness map provides a meaningful distribution of the (interesting) image regions that indicate locations of objects. Furthermore, the center of mass of the objectness map provides a good indication of where the center of the object might be. The SNAK framework employs the objectness measure to determine the object's center in order to use it as the origin of the coordinate system of the image. The range of the coordinate system is normalized by dividing the x-axis coordinates by the width of the image and the y-axis coordinates by the height of the image. For each image, the coordinate system has a range from -1 to 1 on each axis. Normalizing the coordinates ensures that the average position or the standard deviation of a visual word do not depend on the image size, and it is a necessary step to reduce the effect of size variation in a set of images. The SNAK framework is illustrated in Figure 1.

4 Experiments

4.1 Data Sets Description

The first data set used in the experiments is the Pascal VOC 2007 data set [5], which consists of 9963 images that are divided into 20 classes. The training and validation sets have roughly 2500 images each, while the test set has about 5000 images. This data set was also used in other works that present methods to encode spatial information [10,16], thus becoming a de facto benchmark.

The second data set was collected from the Web by the authors of [11] and consists of 600 images of 6 different classes of birds: egrets, mandarin ducks, snowy owls, puffins, toucans, and wood ducks. The training set consists of 300 images and the test set consists of another 300 images. The purpose of using this data set is to assess the behavior of the SNAK framework in the context of fine-grained object recognition. The Birds data set is available at http://www-cvr.ai.uiuc.edu/ponce_grp/data/.

4.2 Implementation and Evaluation Procedure

In the BOVW model used in this work, features are detected using a regular grid across the input image. At each interest point, a SIFT feature [14] is computed. This approach is known as dense SIFT [3,4]. Next, SIFT descriptors are vector quantized into visual words and a vocabulary (or codebook) of visual words is obtained. The vector quantization process is done by k-means clustering [13], and visual words are stored in a randomized forest of k-d trees [15] to reduce search cost. The frequency of each visual word is then recorded in a histogram which represents the final feature vector of the image. A kernel method is used for training.

Three kernels are proposed for evaluation, namely the L_2-normalized linear kernel, the L_1-normalized Hellinger's kernel, and the L_1-normalized intersection kernel. The norms of the kernels are chosen such that the γ-homogeneous kernels are L_γ-normalized. It is worth mentioning that all these kernels are used in

the dual form, that implies using the *kernel trick* to directly build kernel matrices of pairwise similarities between samples. An important remark is that the intersection kernel was particularly chosen because it yields very good results in combination with the spatial pyramid, and it might work equally well in the SNAK framework. The kernels proposed for evaluation are based on four different representations, three of which include spatial information. The goal of the experiments is to compare the bag of words representation with a spatial pyramid based on two levels, a spatial pyramid based on three levels, and the SNAK feature vectors. The spatial pyramid based on two levels combines the full image with 2×2 bins, and the spatial pyramid based on three levels combines the full image with 2×2 and 4×4 bins. In the SNAK framework, the linear kernel, the Hellinger's kernel, and the intersection kernel are used in turn as k^* in Equation (2). Note that SNAK can also be indirectly compared with the approach described in [10], since the results reported in [10] are very similar to the spatial pyramid based on three levels.

The training is always done using Support Vector Machines (SVM). In the second experiment, the SVM classifier based on the *one versus all* scheme is used for the multi-class task. The objectness measure is trained on 50 images that are neither from the Pascal VOC data set nor from the Birds data set. The objectness map is obtained by sampling 1000 windows using the NMS sampling procedure [2].

The experiments are conducted using 500, 1000, and 3000 visual words, respectively. The evaluation procedure for the first experiment follows the Pascal VOC benchmark. The qualitative performance of the learning model is measured by using the classifier score to rank all the test images. In order to represent the retrieval performance by a single number, the mean average precision (mAP) is often computed. The mean average precision as defined by TREC is used in the Pascal VOC experiment. This is the average of the precision observed each time a new positive sample is recalled. For the second experiment, the classification accuracy is used to evaluate the various kernels and spatial representations.

4.3 Parameter Tuning

The SNAK framework takes both the average position and the standard deviation of each visual word into account. In a set of preliminary experiments performed on the Birds data set, the two statistics were used independently to determine which one brings a more significant improvement. The empirical results demonstrated that they roughly achieve similar accuracy improvements, having an almost equal contribution to the proposed framework. Consequently, a decision was made to use the same value for the two constants c_1 and c_2 from Equation (1). Only five values in the range 1 to 100 were chosen for preliminary evaluation. The best results were obtained with $c_1 = c_2 = 10$, while choices like 5 or 50 were only $2-3\%$ behind. Finally, a decision was made to use $c_1 = c_2 = 10$ in the experiments reported next, but it is very likely that better results can be obtained by fine-tuning the parameters c_1 and c_2 on each data set. An important remark is that c_1 and c_2 were tuned on the Birds data set, but the same choice

Table 1. Mean AP on Pascal VOC 2007 data set for different representations that encode spatial information into the BOVW model. For each representation, results are reported using several kernels and vocabulary dimensions. The best AP for each vocabulary dimension and each kernel is highlighted in bold.

Representation	**Vocabulary**	**Linear L_2**	**Hellinger's L_1**	**Intersection L_1**
Histogram	500 words	28.59%	39.06%	39.11%
Histogram	1000 words	28.71%	42.28%	42.99%
Histogram	3000 words	28.96%	45.23%	46.97%
Spatial Pyramid (2 levels)	500 words	31.17%	44.21%	45.17%
Spatial Pyramid (2 levels)	1000 words	31.38%	46.94%	48.27%
Spatial Pyramid (2 levels)	3000 words	31.85%	49.21%	50.78%
Spatial Pyramid (3 levels)	500 words	38.49%	45.20%	47.66%
Spatial Pyramid (3 levels)	1000 words	39.59%	47.87%	49.85%
Spatial Pyramid (3 levels)	3000 words	40.97%	50.37%	51.87%
SNAK	500 words	**42.56%**	**47.39%**	**49.75%**
SNAK	1000 words	**44.69%**	**49.54%**	**51.99%**
SNAK	3000 words	**45.95%**	**52.49%**	**54.05%**

was used on the Pascal VOC data set, without testing other values. Good results on Pascal VOC might indicate that c_1 and c_2 do not necessarily depend on the data set, but rather on the normalization procedure used for the spatial coordinate system. It is interesting to note that the two coordinates are independently normalized according to Section 3.1, resulting in small distortions along the axes. Two other methods of size-normalizing the coordinate space without introducing distortions were also evaluated. One is based on dividing both coordinates by the diagonal of the image, and the other by the mean of the width and height of the image. Perhaps surprisingly, these have produced lower average precision scores on a subset of the Pascal VOC data set. For instance, size-normalizing by the mean of the width and height gives a mAP score that is roughly 0.5% lower than normalizing each axis independently by the width and height.

In the Pascal VOC experiment, the validation set is used to validate the regularization parameter C of the SVM algorithm. In the Birds experiment, the parameter C was adjusted such that it brings as much regularization as possible, while giving just enough room to learn the entire training set with 100% accuracy.

4.4 Pascal VOC Experiment

The first experiment is on the Pascal VOC 2007 data set. For each of the 20 classes, the data set provides a training set, a validation set and a test set. After validating the regularization parameter of the SVM algorithm on the validation set, the classifier is trained one more time on both the training and the validation sets, that have roughly 5000 images together.

Table 1 presents the mean AP of various BOVW models obtained on the test set, by combining different spatial representations, vocabulary dimensions, and kernels. For each model, the reported mAP represents the average score on all the 20 classes of the Pascal VOC data set. The results presented in Table 1 clearly indicate that spatial information significantly improves the performance of the

BOVW model. This observation holds for every kernel and every vocabulary dimension. Indeed, the spatial pyramid based on two levels shows a performance increase that ranges between 3% (for the linear kernel) and 6% (for intersection kernel). As expected, the spatial pyramid based on three levels further improves the performance, especially for the linear kernel. When the 4×4 bins are added into the spatial pyramid, the mAP of the linear kernel grows by roughly $7-8\%$, while the mAP scores of the other two kernels increase by $1-2\%$. Among the three kernels based on spatial pyramids, the best mAP scores are obtained by the intersection kernel, which was previously reported to work best in combination with the spatial pyramid [12].

The best results on the Pascal VOC data set are obtained by the SNAK framework. Indeed, the results are even better than the spatial pyramid based on three levels, which uses a representation that is more than 4 times greater than the SNAK representation. The mAP scores of the Hellinger's and the intersection kernels based on SNAK are roughly 2% better than the mAP scores of the same kernels combined with the spatial pyramid based on three levels. On the other hand, a $4-5\%$ growth of the mAP score can be observed in case of the linear kernel. Among the three kernels, the best results are obtained by the intersection kernel. When the intersection kernel is combined with SNAK, the best overall mAP score is obtained, that is 54.05%. This is 2.18% better than the intersection kernel combined with the spatial pyramid based on three levels.

Overall, the empirical results indicate that the SNAK approach is significantly better than the state of the art spatial pyramid framework, in terms of recognition accuracy. Perhaps this comes as a surprising result given that the images from the Pascal VOC data set usually contain multiple objects, and that SNAK implicitly assumes that there is a single relevant object in the scene, due to the use of the objecteness measure. The SNAK framework also provides a more compact representation, which brings improvements in terms of space and time over a spatial pyramid based on three levels, for example.

4.5 Birds Experiment

The second experiment is on the Birds data set. Table 2 presents the classification accuracy of the BOVW model based on various representations that include spatial information. The results are reported on the test set, by combining different vocabulary dimensions and kernels.

The results of the SNAK framework on this data set are consistent with the results reported in the previous experiment, in that the SNAK framework outperforms again the spatial pyramid representation. The spatial pyramid based on two levels improves the classification accuracy of the standard BOVW model by $3-4\%$. On top of this, the spatial pyramid based on three levels further improves the performance. Significant improvements can be observed for the linear kernel and for the intersection kernel.

The spatial pyramid based on two levels shows little improvements over the histogram representation for the vocabulary of 3000 words, and more significant improvements for the vocabulary of 500 words. The certain fact is that the

Table 2. Classification accuracy on the Birds data set for different representations that encode spatial information into the BOVW model. For each representation, results are reported using several kernels and vocabulary dimensions. The best accuracy for each vocabulary dimension and each kernel is highlighted in bold.

Representation	**Vocabulary**	**Linear L_2**	**Hellinger's L_1**	**Intersection L_1**
Histogram	500 words	59.67%	72.00%	70.00%
Histogram	1000 words	64.67%	78.33%	71.00%
Histogram	3000 words	69.33%	80.33%	74.67%
Spatial Pyramid (2 levels)	500 words	62.67%	75.67%	74.00%
Spatial Pyramid (2 levels)	1000 words	66.67%	79.33%	74.33%
Spatial Pyramid (2 levels)	3000 words	69.67%	81.00%	77.00%
Spatial Pyramid (3 levels)	500 words	68.33%	76.67%	76.00%
Spatial Pyramid (3 levels)	1000 words	70.33%	**80.67**%	78.00%
Spatial Pyramid (3 levels)	3000 words	**73.00**%	82.67%	79.67%
SNAK	500 words	**69.33**%	**79.00**%	**76.33**%
SNAK	1000 words	**71.67**%	80.33%	**78.67**%
SNAK	3000 words	72.33%	**83.67**%	**81.33**%

spatial information helps to improve the classification accuracy on this data set, but the best approach seems to be the SNAK framework. With only two exceptions, the SNAK framework gives better results than the spatial pyramid based on three levels. Compared to the spatial pyramid based on two levels, which has the same number of features, the SNAK approach is roughly $3-5\%$ better. An interesting observation is that the intersection kernel does not yield the best overall results as in the previous experiment, but it seems to gain a lot from the spatial information. For instance, the accuracy of the intersection kernel grows from 71.00% with histograms to 78.67% with SNAK, when the underlying vocabulary has 1000 words. The best accuracy (83.67%) is obtained by the Hellinger's kernel combined with SNAK, using a vocabulary of 3000 visual words. When it comes to fine-grained object class recognition, the overall empirical results on the Birds data set indicate that the SNAK framework is more accurate than the spatial pyramid approach.

5 Conclusion and Future Work

This paper described an approach to improve the BOVW model by encoding spatial information in a more efficient way than spatial pyramids, by using a kernel function. More precisely, SNAK includes the spatial distribution of the visual words in the similarity of two images. Object recognition experiments were conducted to compare the SNAK approach with the spatial pyramid framework, which is the most popular approach to include spatial information into the BOVW model. The empirical results presented in this paper indicate that the SNAK framework can improve the object recognition accuracy over the spatial pyramid representation. Considering that SNAK uses a more compact representation, the results become even more impressive. In conclusion, SNAK has all the ingredients to become a viable alternative to the spatial pyramid approach.

In this work, the objectness measure was used to add some level of translation invariance into the SNAK framework. In future work, the SNAK framework can be further improved by including ways of obtaining scale and rotation invariance. Ground truth information about an object's scale can be obtained from manually annotated bounding boxes. A first step would be to use such bounding boxes to determine if it helps to compare objects at the same scale with the SNAK kernel. Another direction, is to extend the SNAK framework to use the valuable information offered by objectness [1], which is only barely used in the current framework.

Acknowledgments. The work of Radu Tudor Ionescu was supported from the European Social Fund under Grant POSDRU/159/1.5/S/137750.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: What is an object?. In: Proceedings of CVPR, pp. 73–80 (June 2010)
2. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2189–2202 (2012)
3. Bosch, A., Zisserman, A., Munoz, X.: Image Classification using random forests and ferns. In: Proceedings of ICCV, pp. 1–8 (2007)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of CVPR, vol. 1, pp. 886–893 (2005)
5. Everingham, M., van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
6. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding **106**(1), 59–70 (2007)
7. Gonen, M., Alpaydin, E.: Multiple Kernel Learning Algorithms. Journal of Machine Learning Research **12**, 2211–2268 (2011)
8. Ionescu, R.T., Popescu, M.: Objectness to improve the bag of visual words model. In: Proceedings of ICIP, pp. 3238–3242 (2014)
9. Koniusz, P., Mikolajczyk, K.: Spatial coordinate coding to reduce histogram representations, dominant angle and colour pyramid match. In: Proceedings of ICIP, pp. 661–664 (2011)
10. Krapac, J., Verbeek, J., Jurie, F.: Modeling spatial layout with fisher vectors for image categorization. In: Proceedings of ICCV, pp. 1487–1494 (November 2011)
11. Lazebnik, S., Schmid, C., Ponce, J.: A maximum entropy framework for part-based texture and object recognition. In: Proceedings of ICCV, vol. 1, pp. 832–838 (2005)
12. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of CVPR, vol. 2, pp. 2169–2178 (2006)
13. Leung, T., Malik, J.: Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. International Journal of Computer Vision **43**(1), 29–44 (2001)
14. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of ICCV, vol. 2, pp. 1150–1157 (1999)

15. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of CVPR, pp. 1–8 (2007)
16. Sánchez, J., Perronnin, F., de Campos, T.: Modeling the spatial layout of images beyond spatial pyramids. Pattern Recognition Letters **33**(16), 2216–2223 (2012)
17. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)
18. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: Proceedings of ICCV, pp. 370–377 (2005)
19. Uijlings, J., Smeulders, A., Scha, R.: What is the spatial extent of an object?. In: Proceedings of CVPR, pp. 770–777 (2009)
20. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. International Journal of Computer Vision **73**(2), 213–238 (2007)