EVALUATION OF SHAPE AND COLOR DESCRIPTORS BY USING BAG OF WORD TECHNIQUES WITH ONE VS ALL CLASSIFICATION

Fidalgo E. [1], Fernández-Robles L.[2], García-Ordás M. [3], García-Olalla O. [4], Alegre E. [5]
[1,3,4,5] Department of Electrical, Systems and Automatic Engineering
[2] Department of Mechanical, IT and Aerospatial Engineering
Campus de Vegazana S/N 24071, University of León, León, Spain
{eduardo.fidalgo.fernandez[1]}@gmail.com , {l.fernandez[2], mgaro[3], ogaro[4], enrique.alegre[5]}@unileon.es

Abstract

This paper presents the evaluation of several shape and color descriptors on two different DataSets, the "Soccer" dataset and a new Dataset called "Karinas_Dataset1". To represent an image by its corresponding shape (or color) descriptor, a standard BoW creation technique has been applied. In this way, all the images in both DataSets are represented by a descriptor whose dimension depends on the visual dictionary previously created. Afterwards, a one-vs-all classifier has been trained and tested using random images from the Datasets. To avoid spurious data the tests have been repeated 10 times and an average result has been taken as a reference.

Key Words: shape descriptor, color descriptor, bag of words, one-vs-all.

1. INTRODUCTION

Object classification in digital images is one of the most challenging tasks in computer vision. Advances in the last decade have produced methods to extract and describe distinctive local features in natural images. In order to apply machine learning techniques in computer vision systems, a representation based on these features is needed.

The typical object recognition process is composed of the following steps: (i) extraction of local image features (e.g., Hue descriptors), (ii) encoding of the local features in an image descriptor (e.g., a histogram of the quantized local features), and (iii) classification of the image descriptor (e.g., by a support vector machine).

During the past few years, Bag-of-Words (BoW) approaches have allowed significant advances in image classification [1].

Local features are an efficient tool for image classification due to their robustness with respect to occlusion and geometrical transformations [14]. Among the multiple approaches that have been proposed to describe the shape of local features, the SIFT descriptor [5] has proven to be one of the best [13]. But recently people have started to enrich local image descriptors with color information [1, 9, 10, 4, 11, 12]. The main challenge in color description is to obtain robustness with respect to photometric variations, as they are common in the real world, like shadow and shading variations and changes of the light source color. In order to achieve this purpose, the color descriptors are generally based on photometric invariants [6, 10, 16], such as hue and normalized RGB.

The purpose of this article is to compare the results of applying three different descriptors (SIFT, Hue Histogram and Color Name) using two Datasets (Soccer and Karinas_Dataset1). The descriptors are initially compared by plotting the results of the one-vs-all logistic regression classification with different values of iterations and regularization factor.

A summary of the used descriptors is explained in Section 2. Section 3 comments the methods for evaluating the previous descriptors once they are obtained for each image on the Data Set. Section 4 details the Datasets used in this article. Section 5 shows and discusses the results obtained after applying the evaluation method over the different descriptors. The conclusion of our work and possible future directions are stated in Section 6.

2. DESCRIPTORS USED

2.1 SHAPE DESCRIPTORS

2.1.1 SIFT

Scale Invariant Feature Transform (SIFT) is an algorithm published by David Lowe in 1999 [3] and further explained in 2004 [5]. It consists of extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. For an object in an image, interesting points on the object can be extracted to provide a description of the object based on the features of its interesting points and then they are stored in a database. A new image is matched by individually comparing each feature from the new image to this previous database and finding candidate matching features based on measuring the distance of their feature vectors. The correct matches can be filtered from the full set of matches by identifying subsets of key points that agree on the object and its location, scale and orientation in the new image.

The features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination which is essential to perform reliable recognition. This method has proved to robustly identify objects even among clutter and under partial occlusion.

2.1.1.1 Scale-space extrema detection

The first stage of keypoint detection is to identify locations and scales that can be repeatable assigned under differing views of the same object. For each octave of scale space σ, the image is convolved with Gaussian filters (which is a scale-space kernel) at different scales, and then the Difference-of-Gaussian (DoG) function of two successive Gaussian-blurred images separated by a constant multiplicative factor k are taken. A DoG image *D(x, y ,σ)* is computed by

$$D(x,y,\sigma)=L(x,y,k\sigma)-L(x,y,\sigma) \qquad (1)$$

where *L(x, y, kσ)* is the convolution of the original image *I(x, y)* with the Gaussian filter *G(x, y, kσ)* at scale *kσ*. After each octave, the image is down-sampled by a factor of 2. The value of σ is divided by 2 by taking every second pixel in each row and column. This process is schematized in Figure 1. In order to detect the local maxima and minima of *D(x, y, σ)*, each pixel is compared to its eight neighbours in the current image and nine neighbours in the scale above and below. It is selected as a keypoint only if it is lager than all of these neighbours or smaller than all of them.

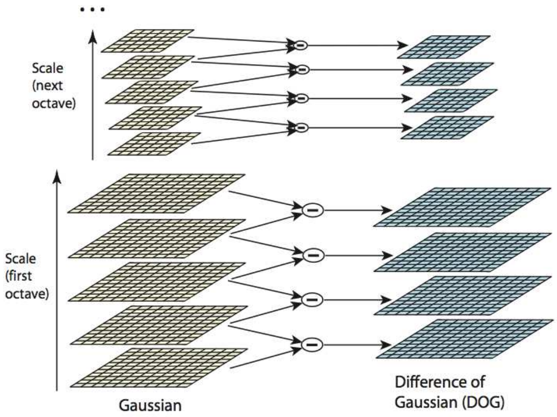


Figure 1. Scale-space extrema detection sketch.

2.1.1.2 Keypoint localization

Scale-space extreme detection produces many keypoint candidates and some of them are unstable. The next step consists of performing a detailed fit to the nearby data for location, scale and ratio of principal curvatures. This allows points with low contrast (and therefore sensitive to noise) or poorly localized along an edge to be rejected.

In order to discard low contrast key points, this approach calculates the interpolated location of the extreme by using the quadratic Taylor expansion of the DoG scale-space function, with the candidate keypoint as the origin. Afterwards, strong responses along edges are suppressed by computing the principal curvatures from a 2x2 Hessian matrix computed at the location and scale of the keypoint. A poorly defined peak in the DoG function will have a large principal curvature across the edge but a small one in the perpendicular direction.

2.1.1.3 Orientation assignment

One or more orientations are assigned to each keypoint location based on local image gradient directions. All future operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations. For an image sample at scale σ, the gradient magnitude, *m(x, y)*, and orientation, *θ(x, y)*, are pre-computed using pixel differences. An orientation histogram with 36 bins is formed, with each bin covering 10 degrees. Each sample in the neighbouring window added to a histogram bin is weighted by its gradient magnitude and by a Gaussian-weighted circular window with σ that is 1.5 times that of the scale of the keypoint. The peaks in this histogram correspond to dominant orientations. Once the histogram is filled, the orientations corresponding to the highest peak and local peaks that are within 80% of the highest peaks are assigned to the keypoint.

2.1.1.4 Keypoint descriptor

The local image gradients are measured at the selected scale in the region around each keypoint. First a set of orientation histograms is created on 4x4 pixel neighbourhoods with 8 bins each. These histograms are computed from magnitude and orientation values of samples in a 16x16 region around the keypoint such that each histogram contains samples from a 4x4 sub-region of the original neighbourhood region. Then, the magnitudes are weighted by a Gaussian function. The descriptor becomes a vector of all the values of these histograms. Since there are 4x4=16 histograms each with 8 bins the descriptor has 128 elements. A representation of this process can be seen in Figure 2. This vector is then normalized to unit length in order to enhance invariance to affine changes in illumination. Effects of non-linear illumination are avoided applying a threshold of 0.2 and then the vector is again normalized.

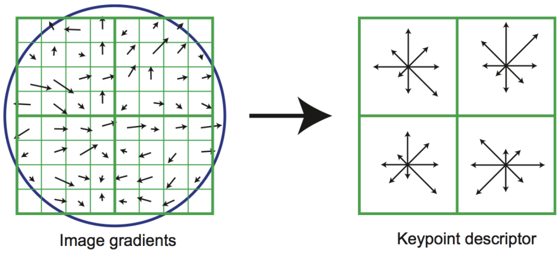


Figure 2. Keypoint descriptor composition.

2.2 COLOR DESCRIPTORS

2.2.1 Hue

Hue descriptor is composed by 36 features.

In the HSV color space, it is known that the hue becomes unstable near the gray axis. To this end, van de Weijer et al. [7] applied an error propagation analysis to the hue transformation. The analysis shows that the certainty of the hue is inversely proportional to the saturation. Therefore, the hue histogram becomes more robust by weighing each sample of the hue by its saturation. The H color model is scale-invariant and shift-invariant with respect to light intensity [7].

That is to say, for hue descriptor, it is computed the hue and saturation at each position; these can also be represented as a vector, where the hue is the angle and the saturation the length. It is computed the hue histogram of the patch where the strength of the update is equal to the saturation of the measurement. This ensures that pixels with low saturation (black-grey-white), where the hue is undefined, have no influence on the final color descriptor.

2.2.2 Color Name

The set of color names used in English is huge and includes several labels like ”white”, ”green”, ”pastel”, and ”light blue”. For this descriptor, the 11 basic color terms of the English language: black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow are used. The basic color terms were defined in the influential work on color naming of Berlin and Kay [2]. A basic color term of a language is defined as being not subsumable to other basic color terms (e.g. turquoise can be said to be a light greenish blue, and therefore it is not a basic color term). The color descriptor K is described as the vector containing the probability of the color names given an image region R

$$K = \{ p(n_1 \mid R), \ldots, p(n_{11} \mid R) \} \qquad (2)$$

With

$$p(n_1 \mid R) = \frac{1}{N} \sum_{x \in R} p(n_{i1} \mid f(x)) \qquad (3)$$

where n_i is the i-th color name, x are the spatial coordinates of the N pixels in region R, f = {L*, a*, b*} and p (n_i | f) is the probability of a color name given a pixel value. The probabilities p (n_i | f) are computed from a set of manually annotated images.

3. BOW PROCESS

Bag of Words method requires the input format shown in Figure 3 so a normalization process is carried out with the image points descriptors obtained with SIFT.

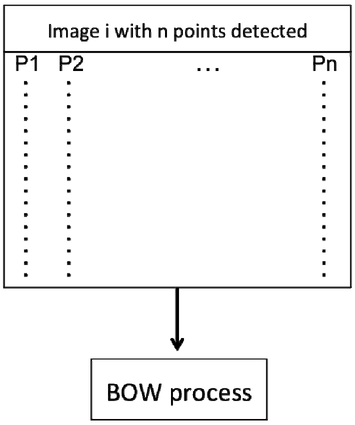


Figure 3. Input format required by BOW.

The Bag of Words method consists mainly on two steps:
First, n centres are obtained taking into account all the image points using a clustering algorithm. This process is called *CalculatedDictionary*. In our case, we have evaluated Bag of Words with k-means as clustering method.

Secondly, each point is assigned to its corresponding centre taking in consideration the distance between points and centres. This process is called *Assignment*. Both processes are shown in Figure 4.

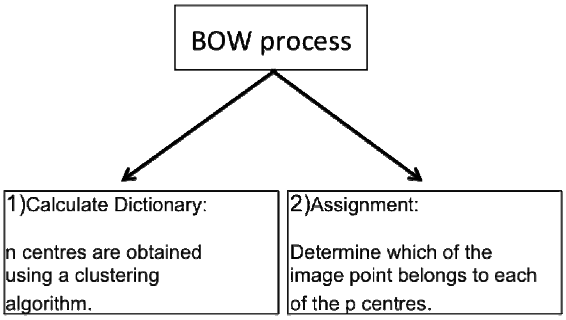
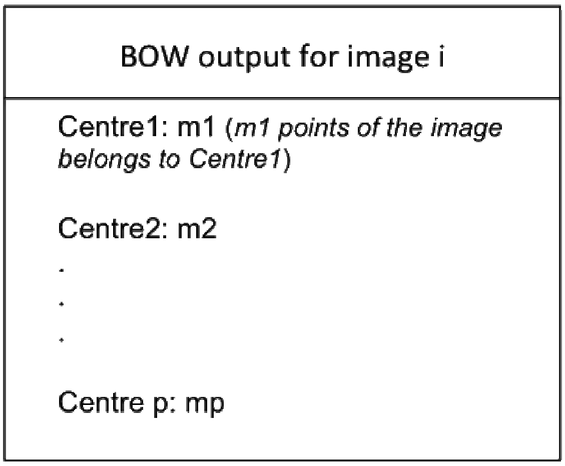

BOW process
1)Calculate Dictionary:
n centres are obtained using a clustering algorithm.
2)Assignment:
Determine which of the image point belongs to each of the p centres.

Figure 4. Bow process steps.

Both methods are carried out for each image (See Figure 5) so at the end of the process, a matrix like the one shown in Figure 6 is obtained.



BOW output for image i

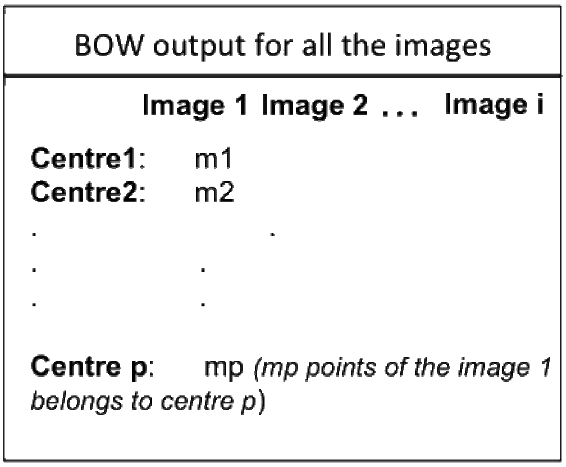
Centre1: m1 (*m1 points of the image belongs to Centre1*)

Centre2: m2

.
.
.

Centre p: mp

Figure 5. For each image, BOW returns the number of image points belonging to each centre.



BOW output for all the images

Image 1 Image 2 . . . Image i

Centre1: m1
Centre2: m2
. .
. .
. .

Centre p: mp (*mp points of the image 1 belongs to centre p*)

Figure 6. Matrix with the number of image points belonging to each centre for all the images.

4. ONE-VS-ALL EVALUATION METHOD

A one versus all classification based on logistic regression have been carried out in order to evaluate these descriptors. The logistic regression classification has been performed for each class, considering one class as positive and all the rest as negatives. Once all the logistic classifiers have been trained, each element is assigned to the class that gives the best response. In order to find the best configuration, several values for the regularization parameter (λ) have been tested just to avoid over fitting (for high values of λ) and under fitting (small values of λ) problems. λ=0 means that no regularization is applied. Small values of λ are referred to values close to λ=0.

5. DATASETS USED

Two datasets were used for the evaluation, the well-known Soccer Dataset (color dominant) and a new dataset created by VARP Group called "Karinas_Dataset1".

5.1 SOCCER

This data set [17] contains images from 7 soccer teams (see Figure 7) taken from the web, comprehending 40 images per class, divided into 25 training and 15 testing images per class. Although, players of other teams were allowed to appear in the images, no players being a member of the other classes in the data set were accepted.

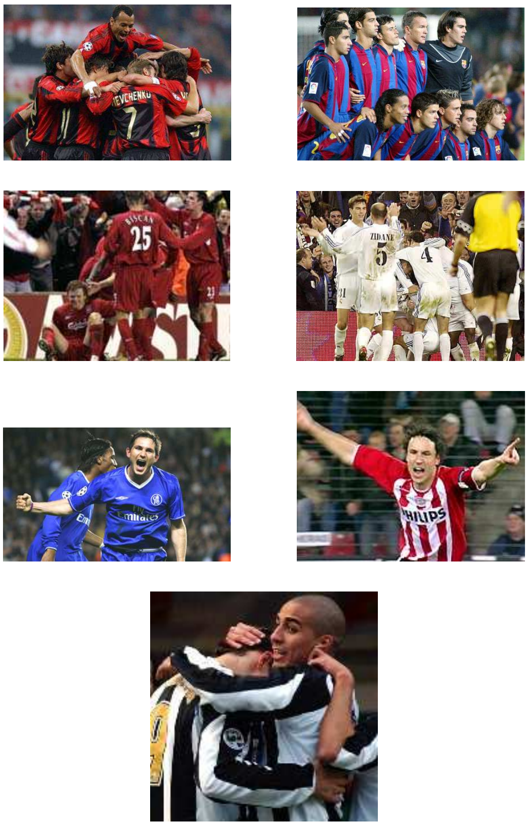


Figure 7: Soccer Dataset: One sample per class.

5.2 KARINAS_DATASET1

VARP Group of the University of León has created Karinas_Dataset1 [15], composed by 614 frames of 640x480 pixels that come from 3 videos recorded under different conditions. All videos were recorded in different bedrooms and with different distributions, illumination, textures, etc., making the object retrieval a challenging task. Nevertheless some objects are present in all videos such as two toy cars, some clothespins, a stuffed bee, some pens, some cups or a child book together with a big doll. The doll is usually the principal actor in the videos and helps us to simulate partial occlusions of the objects and a more realistic scenario. Although these objects are present in every video, they do not appear in every frame. For image classification in this article has been used the book, the blue and yellow car, and the pink, blue and green clothespin shown in Figure 8. The total number of query objects present among the 614 frames that we used in this paper can be found in Table 1.

Table 1: Karinas_Dataset1 Dataset information.

Object	Number
Book	115
Blue car	102
Green car	138
Pink clothespin	125
Blue clothespin	92
Green clothespin	42

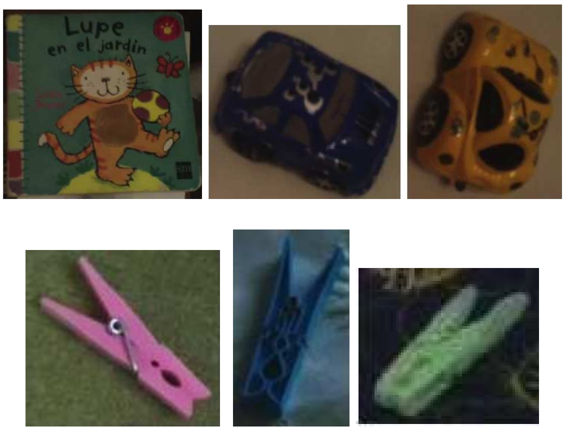


Figure 8: Karinas_Dataset1: One sample per class.

6. RESULTS

6.1 EXPERIMENT ARRANGEMENT

In the experiments, the BoW creation explained in chapter 3 has been used, in which the key points are extracted through Harris-Laplace keypoint detector. The visual dictionaries are constructed using the K-means clustering algorithm. Feature vectors (bag-of-word histograms) are built for each cue separately by using the corresponding visual dictionary. In this series of experiments, two visual cues were used: the shape and the color. Shape information is extracted from SIFT descriptors, and color information is extracted from Hue-histograms [9] and Color Name (CN) [8] descriptors.

The visual dictionary size of the different descriptors remains as follows:

- SIFT: 400 visual words.
- Hue: 300 visual words.
- Color Name: 300 visual words.

All the reported results are obtained using a one-vs-all classifier learned from the feature vectors of the different methods. As an evaluation criterion the accuracy (%) has been used.

The following table represents a sample of the tests performed. Each descriptor has been tested with different values of regularization parameter ($\lambda = 0$ means no regularization at all) and different number of iterations. Since training and test images for each experiment are randomly chosen (train 63,5% and test 37,5%), the experiment for each descriptor, with a fixed value of λ and a fixed number of iterations is repeated ten times and the mean value and the standard deviation of the obtained results are computed.

Table 2: Experiments Table Sample.

	λ = 0 Iter = 100		**λ = 0 Iter = 1000**	
	time (s)	**% Accu**	**time (s)**	**% Accu**
Test 1	6,06	35,29	55,20	28,43
Test 2	5,25	31,05	52,66	33,66
Test 3	5,12	35,95	52,88	30,72
Test 4	5,10	31,70	52,68	32,03
Test 5	5,22	35,29	53,28	31,70
Test 6	5,27	34,31	51,63	31,70
Test 7	5,42	31,05	53,64	30,72
Test 8	5,16	34,64	52,72	31,05
Test 9	5,28	33,99	52,83	31,05
Test 10	5,12	33,01	54,55	30,07
Mean	**5,30**	**33,63**	**53,21**	**31,11**
STD	0,28	1,83	1,03	1,36

To avoid an excessive number of tables (8 tables per descriptor) we created two plots, one per Dataset, to represent the mean values obtained under every condition (values of λ and Iterations). One example of the results with λ equal 0 and Iterations equal to 100 and 1000 is shown in Table 2.

6.2 SOCCER RESULTS

The results obtained for this dataset are showed in Figure 9.

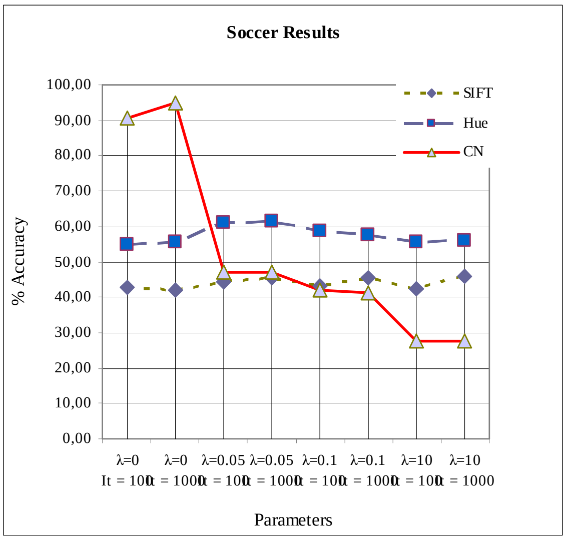


Figure 9: Soccer Dataset Plot Results.

On the one hand, as expected, the SIFT descriptors are less relevant than the Color ones (Hue and CN) for classifying soccer images.

On the other hand, the obtained results are different when using or not using regularization depending on the descriptor. It can be easily seen that with no regularization (λ=0) the results are quite better than even with a small value of lambda for the case of Color Name descriptors. However, for SIFT and Hue, when adding this regularization factor the results remains practically the same despite the value of lambda used in the case. More precisely, they are slightly better with a small factor of regularization.

Moreover, an increase on the iterations used to calculate the outputs of the LR model implies a slightly increase in the accuracy of the classification. Nevertheless, in some cases it is not worth it since the rise of time invested due to the increase of iterations needed is not proportional related with the improvement on the accuracy results.

6.3 KARINA_DATASET1 RESULTS

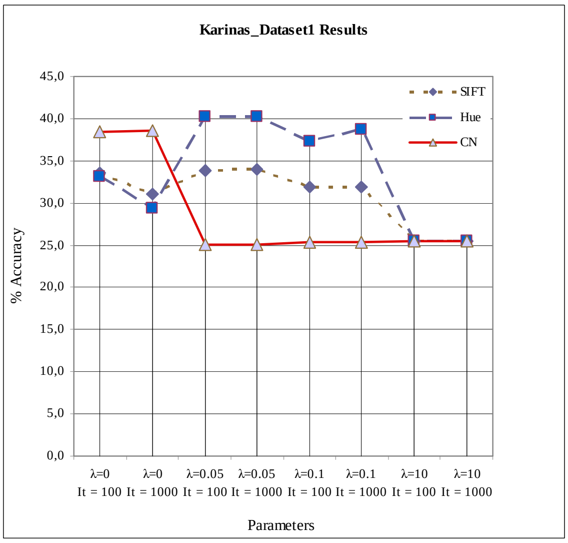


Figure 10: Karinas_Dataset1 Dataset Plot Results.

Unlike the Soccer dataset that is color dominant, Karinas_Dataset1 is considered both shape and color dominant due the complexity of the images extracted. Results obtained with no regularization confirm the previous statement in the case of no regularization classification, due the fact that the three descriptors obtained similar results (all in a range of 30-40%). These poor accuracy results may derive from the fact that the dataset is too small compared with the objects to be identified (6 query objects and 614 images for training and test).

The use of a small factor of regularization causes a slight increase in the SIFT accuracy and it also causes a noticeable improvement in the case of Hue descriptors results. But the Color Name descriptors are severely penalized by the used of the regularization term.

An increase on the iterations used to calculate the outputs of the LR model implies a slight increase in the accuracy of the classification, but again the time invested is also increased.

Like in the Soccer dataset, increase the regularization factor causes the poorest results, but Color Name keeps practically equal the accuracy with even a small value of regularization factor.

7. CONCLUSIONS

This paper presents an evaluation among three different descriptors applied on two different Datasets (Soccer and Karinas_Dataset1). Values obtained after image classification have been commented according to the dataset, descriptor and parameters used for the classification. It has been proved that colour descriptors achieved better results on a dataset that is colour dominant (Soccer), and poorer results have been obtained for a complex dataset (Karinas_Dataset1) because of the small amount of images compared with the number of objects that have been tried to identify.

Future Works will include the increase of the size of the Karina Dataset, the use of additional descriptors for image classification and SVM as the way to classify the images of the datasets. Since Karina Dataset proves to be very challenging for the task of image classification, the fusion of the previous descriptors will be tested as a way to obtain better results.

Acknowledgements

This work has been supported by the ASASEC (www.asasec.eu) European Project with reference HOME/2010/ISEC/AG/043 and by the grant DPI2012-36166 from the Spanish Government.

References

[1] A. Bosch, A. Zisserman, and X. Munoz, “Scene classification via pLSA,”, *ProcP of the European Conference on Computer Vision*, 2006.

[2] B. Berlin and P. Kay, “Basic color terms: their universality and evolution”, *Berkeley: University of California*, 1969.

[3] D. G. Lowe. “Object recognition from local scale invariant features”. *ICCV*, volume 2, pages 1150–1157 vol.2. IEEE Computer Society, August 1999.

[4] D. Vigo, F. Khan, J. van deWeijer, and T. Gevers. “The impact of color on bag-of-words based object recognition”. *ICPR*, pages 1549 – 1553, 2010.

[5] D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] J. van de Weijer and C. Schmid, “Coloring local feature extraction,” *Proc. of the European Conference on Computer Vision*, Graz, Austria, 2006, vol. 2, pp. 334–348.

[7] J. van de Weijer, T. Gevers, and A. Bagdanov, “Boosting Color Saliency in Image Feature Detection,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 150-156, Jan. 2006.

[8] J. van deWeijer and C. Schmid. “Applying color names to image description”. *ICIP*, pages 493–496, 2007.

[9] J. van deWeijer and C. Schmid. “Coloring local feature extraction”, *ECCV*, volume 3952, pages 334–348. Springer, 2006.

[10] J.M. Geusebroek, “Compact object descriptors from local colour invariant histograms”, *British Machine Vision Conference*, 2006.

[11] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. “The devil is in the details: an evaluation of recent feature encoding methods”. *BMVC*, 2011

[12] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. “Evaluating color descriptors for object and scene recognition”. *PAMI*, 32(9):1582–1596, 2010.

[13] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[14] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 62–86, 2004.

[15] Karinas_Dataset1 http://pitia.unileon.es/varp/galleries

[16] Schmid, C., and R. Mohr, “Local grayvalue invariants for image retrieval,” *IEEE PAMI*, 19, 5 (1997), pp. 530–534.

[17] Soccer Dataset http://lear.inrialpes.fr/people/vandeweijer/data.html