Detecting Bias on Aesthetic Image Datasets

Adrian Carballal, Department of Information and Communication Technologies, University of A Coruña, A Coruña, Spain

Luz Castro, Department of Information and Communication Technologies, University of A Coruña, A Coruña, Spain

Rebeca Perez, Department of Information and Communication Technologies, University of A Coruña, A Coruña, Spain

João Correia, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

ABSTRACT

In recent years, there have been attempts to discover the principles that determine the value of aesthetics in the domain of computing. Many and diverse studies have tried in some way to capture these principles through technical characteristics. To this end, helped by the ease of Internet data acquisition, datasets of images have been published which were obtained online at random from websites and photography competitions. To guarantee the validity of a system of aesthetic image classification, one must first guarantee its capacity for generalization. This paper studies how the indiscriminate selection of images can affect the generalization capacity obtained by a binary classifier.

Keywords: Entropy, Image Aesthetics, Image Retrieval, Machine Learning, Pattern Recognition

1. INTRODUCTION

In the history of humanity we have always used art as a form of expression for our inquisitiveness, thoughts and experiences. However, it is with the birth of IT and artificial intelligence that art and aesthetics have come into the sphere of computerized systems.

In recent years various attempts have been made to create computer vision systems capable of the classification and the ordering of a series of images similar to those carried out by humans. Criteria such as originality, theme, Rule of Thirds and so on are considered. Large groups of images are used to allow contrasts of information in said approximations. The majority uses similar sources (websites, photographic competitions online…) which provide information about the different forms of people's judgments.

We understand that these images are crucial for obtaining truly representative results from

DOI: 10.4018/ijcicg.2014070104

which we are able to extrapolate. But so far, no studies have been made to ascertain the validity of these datasets or whether the data they supply are truly representative.

The classification of images meets its first handicap due to a marked characteristic of human nature. An aesthetic evaluation can be influenced by a great quantity of subjective aspects which if not actually mistaken, may not be wholly universal. For this reason, images are classified according to criteria merely aesthetic or objective such as shapes, colours and composition which allow us a quantitative evaluation, leaving to one side the content.

In this paper we detail the studies of image datasets most cited in aesthetic classification experiments: two collections obtained from the photography website "Photo.net" Datta et al. (2006), and Datta et al. (2008) and another two from the photography competition "DPchallenge.com" (Ke et al., 2006), testing its suitability for this type of tasks.

Employing these sample groups simple characteristics will be detailed which will be used to classify the images in function of a series of quantitative criteria. Later, with the results obtained, an individual analysis will be carried out on each dataset separately. We will present a study showing their capacity for generalization about images obtained both from the same and different sources in such a way as to show if it is possible to extrapolate from their results.

2. STATE OF THE ART

Within the group of works orientated towards automatic aesthetic classification, some of the most cited are from Datta et al. (2006), Wong et al. (2008), Ke et al. (2006), Luo et al. (2009). Each one of these authors has supplied a different method in the search for the ideal design characteristics in relation to technical components such as luminosity, saturation, etc.

Despite being different in their aims and methods, these author's investigations have all employed the same kind of datasets, which include photographs and human evaluations.

Although these datasets may be a suitable source for study, each one comes with its own peculiarities. They consist of large groups of images created by third parties external to the investigation. Also, each photograph includes its own evaluation in the form of a rating carried out by various individuals on the basis of different criteria.

However, the conditions in which these ratings were carried out were not controlled as in an experiment attended in person. There is also a significant dearth of information regarding the participants and we cannot disregard the possibility that extraneous variables have contaminated the sample.

The greater part of these photo databases exhibit a semantic bias and a bias in terms of content. Aimed at professional photographers there is always a certain tendency towards various types of subject, framing and uses of colour.

Datta et al. (2006) employed a dataset known as "Photo.net", which has been used profusely in experiments related to aesthetic classification. This image database contains more than a million photographs evaluated and rated by its members in terms of the "originality" and "aesthetic" of each photograph. Nevertheless, these two criteria end up being very closely related forms of rating, which shows that finally the users do not differentiate between both aspects. The difference between aesthetics and originality being minimal, the distinction is not relevant.

Ke et al. (2006) use a dataset from "DPchallenge.com". This website contains a total of 16,509 images, each rated by at least one user. But by concerning itself with a photographic competition we must always take into account the aesthetics, the pre-established subjects, the different subjective components and the selective ratings among its members.

Both sets have been used by other investigators to test new methodologies and distinct characteristics. Despite this, until now nobody has studied the suitability of these image banks, nor if their results can be extrapolated. That certain characteristics might be useful in the

study of a particular group of images, does not mean that these characteristics are universal and even less that they should be applicable to new photographs. To be able to be sure of the validity of the focus we need to be able to have sufficiently ample and representative groups as a reference. Only then can be sure when we generalize on the basis of their results.

Apart from the sample groups, the methodology which is usually employed in this area can also have underlying problems because different classifiers are developed and validated using images from the same source. Without having studied these sample groups beforehand one cannot certify their universality. For this reason it would help achieve truly coherent results if different image databases were used in the phase of learning.

In this study we propose to study the suitability of the sample groups by means of simple classifiers which use basic features related to the statistical values of the distribution of intensity and entropy. These stated characteristics will be obtained from images belonging to the datasets detailed by Datta et al. (2006, 2008) and Ke et al. (2006).

3. DATASETS CONCERNING AESTHETICS

As we have stated, investigators such as Datta et al. (2006, 2008) and also Ke et al. (2006) have based their experiments regarding the classification of visual data on a series of image database. These datasets have been the most often referred to in this field. We will go into depth in detailing the sources from which these statistical data were gathered. Furthermore, other state of the art datasets will be addressed.

3.1. Source: PHOTO.NET

In Datta et al. (2006) is shown a dataset sourced from "Photo.net", a website which gathers more than a million images belonging to more than 400,000 users. Each image receives a two criteria rating: aesthetics and originality and is given points on a scale from 1 to 7. The information and ratings belonging to any image are of a public nature and can be seen on the same site.

Photo.net publishes part of the information about the images which constitute it, showing with each image the ratings of originality and aesthetics on an average scale and the values of their characteristics. However, this data does not include information about the commentators. The dataset includes 3,581 images each judged by at least two different people, with an average score situated within the range of (3,55 –7) with an overall average totaling 5,06 and a statistical deviation from the norm of 0,83.

That a high correlation has been detected between the originality and aesthetics scores indicates that the users may not be differentiating them sufficiently.

To our understanding, this sample group exhibits various and connected underlying problems. (i) the quantity of images may be insufficient for a result from which we can generalize and extrapolate (ii) there is no reason given for the choice of the sample size (iii) it is assumed that the average value obtained by an image is representative when there is no control over the rating and sometimes these have only been submitted by two people, which is a rather insufficient sample.

Datta at al. and equally other investigators (Wong et al., 2009) who have used this sample group, carry out a preliminary distribution with the aim of obtaining two different groups, employing the aesthetic ratings of the users: a group of high quality images named *High* and another low quality named *Low*. Those images which have obtained a score greater or equal to 5,8 are classified as high, those less than or equal to 4,2 will be classified as low, finally two groups of images are gathered, the high one of 832 and the low one of 760 items.

3.2. Source: PHOTO.NET (2008)

Subsequently, Datta et al. published a new study in which they presented other datasets which could be used in aesthetic classification tasks. Also, a second dataset from the website Photo.net was published on the authors' website.

This new dataset is composed of 20,278 images with an average rating of 12 people and with a typical deviation of 13. However, many of these photographs are not now available due to their elimination from the web, the same as occurs with some of the images of the previous dataset.

In comparison with the previous collection of images this dataset permits a more complete statistical analysis because it supplies specific data about the ratings for each image indicating the number of votes for each level on the rating scale. In this case, all the images have been rated by at least four people. The range of average ratings is situated between (2,33-6,90), with an average of 5,15 and a deviation from the norm of 0,58.

This dataset still has the same inconvenience as the first in terms of the number of ratings per image; seeing as cases exist here where an image has only been rated by four people, while others have been rated by several hundreds, thus generating an unequal sample. We do not have a record of whether scientific results pertaining to aesthetic classification have been published relating to this dataset.

3.3. Source: DPCHALLENGE.COM (2006)

The dataset published by Ke et al. (2006) is one of the most often used in aesthetic classification experiments of which, for the results obtained, Luo et al.'s (2008) is the outstanding example. It was created compiling the images of the photography website DPChallenge.com and possesses a total of 60,000 photographs. Unlike the two previous collections obtained from Photo.net, all the images have been rated by at least one hundred people. The ratings vary between 1 and 10 (the former being the lowest score possible and the latter being the highest).

We have already commented on the drawbacks related to the source DPChallenge.com: the ultimate lack of control over the ratings and the users who carry them out, with a subject aimed at a professional public who constitute a bias in terms of the content and subjective aspects which have no value for computerized systems. Furthermore the dataset does not specify in any case the rating tendency for the images and their average deviation from the norm.

With the purpose of working with more controlled collections, Ke et al. (2006) have created a two part collection: High and Low. From the total they extract 10% with the highest and lowest average rating in such a way that each subgroup is composed of 6,000 images in total. Afterwards they have carried out a division using a random dichotomy with each one of the two sub-groups, obtaining finally four groups of 3,000 each, two of high and two of low quality. Each one of these groups is employed towards the same end; two will help to develop the systems created while the other two will validate the efficacy and capacity of the two previous ones.

3.4. Other Published Datasets

As we have already commented, Datta et al. (2008) proposed four new datasets for conducting experiments. The images were extracted from websites like "Terragalleria.com" and "Alipr.com" and new compilations in "Photo.net" and "DPChallenge.com". In all of these one could carry out an evaluation on the basis of a rating for each photo. These datasets were presented as a source for carrying out experiments on the prediction of aesthetic results, prediction of an aesthetic class and even the prediction of emotion. Furthermore, Murray et al. (2012) introduce a dataset of images on a grand scale called AVA.

The dataset obtained by DPChallenge.com is composed of 16,509 images with an average evaluation for each image of 205 and a deviation from the norm of 53. All the images have been rated by at least one user.

Terragaleria.com employs images created by the photographer QuangTuan-Luong

throughout his travels. Despite being one of the most important collections for the US National Park, it only includes photos taken by him although all of them have been evaluated by third parties. The scale of the scoring is from 1 (the lowest) to 10 (the highest). The average rating for an image is 22 with an average deviation of 23.

The dataset which employs images from Aplipr.com includes more than 13,010 images rated by users according to their emotions. However, it does not possess statistical information relating to the evaluation carried out and furthermore some of the photos are repeated.

We do not have record of any of these datasets being used in aesthetic classification experiment. Nevertheless, the three exhibit the same deficiencies as those before in terms of their control of ratings, sample sizes and characteristics acquired by the development and evaluation of their own collections.

Murray et al. (2012) introduce a new large-scale database for conducting Aesthetic Visual Analysis: AVA. It contains over 250,000 images covering a wide variety of subjects on 963 challenges along with a rich variety of meta-data, including a large number of aesthetic scores for each image, semantic labels for over 60 categories as well as labels related to photographic style.

AVA provide three types of annotations: i) aesthetics: Each image is associated with a distribution of scores which correspond to individual votes. The number of votes per image ranges from 78 to 549, with an average of 210 votes generated by hundreds of amateur and professional photographers with a practiced eye. They provide three types of annotations: ii) semantic: providing 66 textual tags describing the semantics of the images (over 150,000 images contain at least two tags); iii) photographic style: selecting 72 challenges corresponding to photographic styles identified 14 resulting photographic styles: Complementary Colors, Duotones, High Dynamic Range, Image Grain, Light on White, Long Exposure, Macro, Motion Blur, Negative Image, Rule of Thirds, Shallow DOF, Silhouettes, Soft Focus, Vanishing Point.

4. STUDY OF STATE OF THE ART SAMPLE GROUPS

In this section we explain the elements necessary on the experimental side to attempt a test of whether these sample groups can, on average, aesthetically classify other photos despite not being obtained from the same source.

We will use a binary classifier based on Support Vector Machines. As entry data for this said classifier we will use statistics and entropy estimators which characterize each one of the images of the sample group. In our case we will develop distinct classifiers using images from Photo.net and those of DPChallenge.com in a separate way. The generalizing capacity of the distinct systems will be tested using images from the development and the validation; both those obtained from the same source and by employing different images. In the case of the photos taken from DPChallenge.com it has been possible to use two image groups employed by Ke et al. which are utilized to validate their system of 6,000 images each. However, it has been impossible to recover the complete group of photographs that figure in (Datta et al., 2006) constituted by 3,581 images and those detailed in (Datta et al., 2008) which consists of 20,278. In actual fact, it has only been possible to obtain a total of 3,247 images from the first and 18,105 from the second due to some not now being available on "Photo.net". The same problem has also hampered other investigators (Wong et al., 2009).

In Table 1 we detail the four image groups which will be used, indicating their origin, those works which have been shown for the first time, the number of original images which it has been able to access and finally, we shall also give it a name.

In continuation we specify the metrical statistics and those relative to the entropy employed.

4.1. Basic Features Used

The majority of experiments in the field of computing concerned with aesthetics follow the

Table 1. Information relative to the dataset used in the experiment part

Source	Publication	#Original Set	#Available Set	Name
Photo.net	Datta et al 2006	3,581 images	3,247 images	PN06
Photo.net	Datta et al 2008	20,278 images	18,105 images	PN08
DPChallenge.com	Ke et al 2006	6,000 images	6,000 images	DPCt
DPChallenge.com	Ke et al 2006	6,000 images	6,000 images	DPCv

same pattern. Different computing classifiers are used, normally binary, fed by a combination of data which represent values attributed to the image samples. In general, values related to technical aspects are used such as brightness and saturation for the purpose of finding those which achieve the best results possible in image classification tasks when relating to aesthetic criteria.

The goal of this work is not to study the level of improvement which the choice of values can afford; we are more interested in testing if the choice of images results is trivial in terms of tests of this type. It is for this reason that we will use two sets of features in our experiments. One will be obtained on the basis of the average and the typical deviation from the value of the pixels which constitute each image and the other one will be use features which attempt to estimate the entropy relative to an image on the basis of the value of one pixel in respect to its neighbouring pixels. Further on we will explain both sets in more detail.

Before calculating each one of these features we have undertaken the transformation of each image. First we re-dimensioned to 256x256 pixels and after we transformed it to an RGB model with a depth of 8-bit per channel scaled on a range of (0,255) with this we have got all images to share dimensions and format. Finally, we have processed each image with a colour HSV model which will be fundamental for the calculation of the first used set of metrics. Some steps towards this transformation, as in the case of the change in the relation of the 1:1 ratio of appearance, constitute a loss of information and hence a deformation of the image. But in previous experiments, in other fields, it has been proven that such a transformation does not affect the ability of this type of system to manage classifications of this type (Romero et al., 2011, 2012).

Within the first set we have employed two statistical measures for the colour of each image: the average and the standard deviation. Both are calculated on the basis of the value of the intensity of the pixels in its effect on different channels of the colour HSV model, with the exception of channel H (hue). Given that channel H is circular, the average is calculated based on the norm and the angle of the hue values. With all of this, we obtain a total of seven statistical values: four for the average and three for the typical deviation. We will refer to this set of values from now on as *AvgStd*.

The second set of values used corresponds to image entropy estimators. The entropy measures the degree of disorder existing in the system independently of its own nature. According to Arnheim (1974), "order is a necessary condition for everything the human mind is to understand" (p. 1). Using this as their principle, there exist various works in this field that relate entropy, or the degree of complexity of an image, with the associated aesthetic beauty (Machado and Cardoso, 1998) (Machado et al, 2007). Many entropy estimators exist although for this experiment we have opted to employ Zipf´s law.

Zipf´s law (Zipf, 1949) is based on the observation of phenomena generated by auto-adaptive organisms, such as human beings. It is commonly known as "the principle of least effort". Once a phenomena or event has been selected for a study, one examines the contribution of each particular case in respect to the

whole and determines the range of importance or predominance. Informally the smallest events tend to occur with greater frequency while the bigger events tend to occur with less frequency. A variant within Zipf´s law uses the size of the phenomenon instead of its range, generating a distribution of size-frequency. We will use this formula in our experiment.

The calculation of the characteristics referring to Zipf´s size-frequency are carried out by obtaining the difference of a pixel to each of its neighbours and the counting of the total number of times the said differential value occurs. After, we will organize these values in descending order according to the number of times they occur and will represent them in a Cartesian axis according to their value and their frequency. From this same graph we will use as entropy estimates the inclination (M) and the lineal correlation (R2) of that line of tendency. As happens with the average and the typical deviation from the statistical norm, we obtain these two values for the three HSV colour model channels. In this way we speak of a set formed by six values or entropy estimators and we will refer to those from now on as *SizeFreq*.

4.2. Model of Classification Employed

It is very common for work derived from the classification of images concerning aesthetic criteria to use a type of classifier called SVM or Support Vector Machines (Vatnik, 1997). These allow classifications to be carried out among sets of data on the basis of a maximum margin of separation which exists among them. An SVM represents sample data which constitutes a decision surface and such data which are not separated linearly are converted by a function or a kernel to a space of characteristics with a larger dimension. Once this is done the system determines a decision frontier which separates the points of the sample into distinct classes. This frontier function, which represents a hyper-plane, permits a distinction among the data which belongs to which class.

There exist a great number of applications which allow the use of this classifier, in a way both simple and intuitive for experts and participants alike. One of the most used is WEKA (Wiakato Environment for Knowledge Analysis) (Witten and Frank, 2002) which will be used in these experiments.

In the studies mentioned previously (Datta et al. 2006) (Ke et al., 2006) (Luo et al., 2008) (Wong et al., 2009) they use this type of classifier both with the default parameters and with some specific empirically determined ones. In the experiments carried out we have used a function or linear kernel provide by the package LibSVM (Chang and Lin, 2001) with their default parameters (γ=3.7, ν=0.5, ε=0.1, without the normalization of the entry data).

5. EXPERIMENTS CONDUCTED

In this section we carry out two independent studies: i) the generalization obtained training and validating our classifiers with distinct datasets but obtained from the same source, and ii) the generalization obtained from using distinct datasets coming from different sources. For this we will use four sets of images belonging to the state of the art seen in previous sections.

We have conducted two experiments with the same values, the same classifiers and the same parameters of learning, only interchanging the data of training and validation. This allows us to draw two sets of conclusions which will serve to evaluate the generalization capacity obtained with both sets of sample images.

In our case, we use a training procedure named 5-fold cross-validation (Browne, 2000), which consists in dividing the training patterns into 5 sets without any being of the same size. The process of learning is carried out five times in total. In each case one of the five sets is used as a test set and the other four to train. For this reason, all of the patterns are used once for the test and four times for the training (this phase we call *Train*). After, we will present another set of distinct images to each training classifier.

After, to each independently trained classifier, we will present a distinct set of images using the previous stage. The reported results in this experiment refer to those of this external validation (this phase we will call *Validation* or *Valid*).

5.1. Experiment 1: Distinct Datasets Belonging to the Same Sources

Given that we have two sources, we will start with the dataset belonging to "Photo.net" which we previously called PN06 and PN08. In Figure 1 we show the results for the validation obtained for each set of data, (PN06 on the left and PN08 on the right) training in turn with the same set and then with the other. The data obtained show that the set of training images which are in this case different, does not significantly affect the classification. In the case of the set of features *AvgStd* results plus *SizeFreq*, the difference is always inferior to 2% accuracy. For this reason, we could conclude that the generalization reached is satisfactory.

On the other hand, in Figure 2 are shown the results obtained from those two datasets originating in "DPChallenge.com" (DPCt and DPCv). Unlike what was observed in the previous case there exists a clear difference between validating one set and another. On validating the set DPCt, with whichever of the two we obtain similar results. On the contrary, on validating the set DPCv the results are very different in two ways: (i) when training with distinct sets the difference is superior 5% and (ii) the difference validating both sets by themselves is close to 30%. These results could indicate various situations: either DPCv is a subset of images with characteristics clearly differentiable by means of the information obtained by the set DPCt or the images which conform to the set DPCv have some intrinsic bias which permits a clear distinction even with basic features.

If we not only attend to the metrical statistics and entropy, but also the colour channels, we then obtain the results shown in Figure 3. There exists a clear internal difference within

Figure 1. Results obtained training and validating with each set on an individual basis and interchanging those sets in a process of validation with images from photo.net

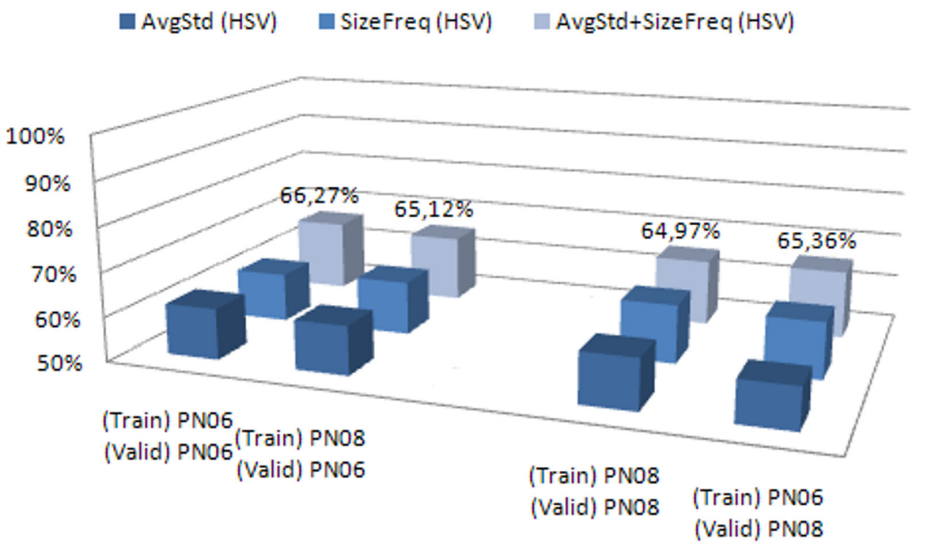


Figure 2. Results obtained training with each set and on an individual basis and interchanging those sets in a process of validation with images from dpchallenge.com

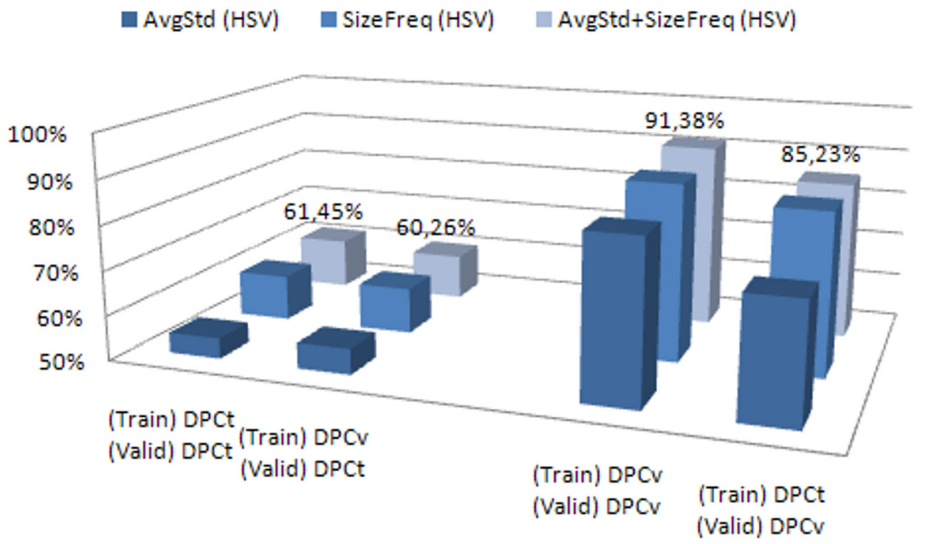
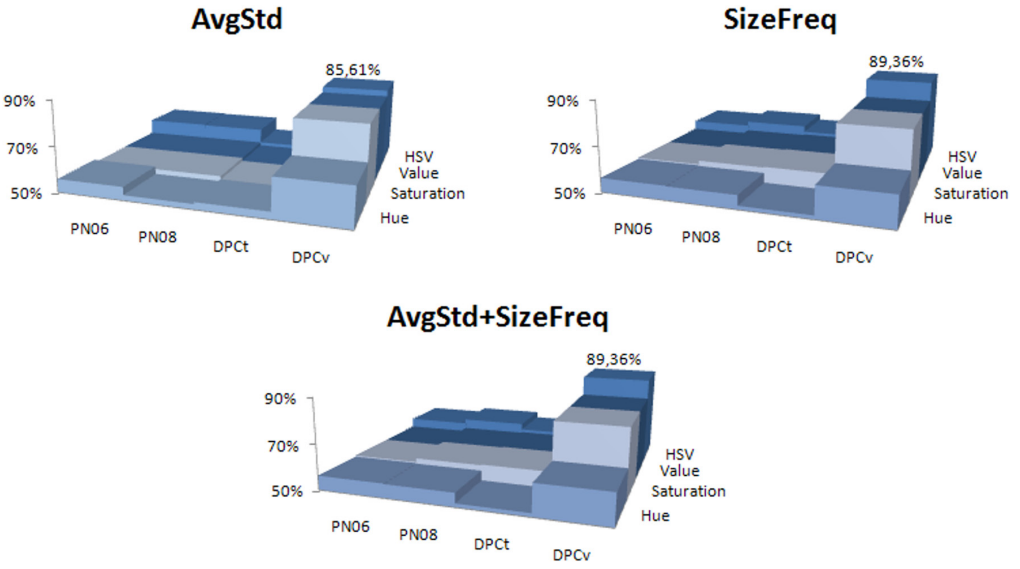


Figure 3. Results obtained according to each colour channel individually and all trained and validated with each one of the presented data sets



set DPCv which becomes more evident in the channels value and saturation. Directly observing the values obtained for those said channels in the samples, we detected that those ranges, when values are moved for the high and low sets, hardly overlapped. This allows us to differentiate with greater ease.

5.2. Experiment 2: Distinct Datasets Belonging to Different Sources

Once the capacity for generalization has been studied among the datasets coming from the same source of information, we precede to observe the behaviour of the classifiers previously created when presenting images originating from a different source.

The behaviour of the three sets of metrics is similar (see Figure 4) in the case of developing and validating the two sets of images obtained from Photo.net we observe that both sets offer similar results. If we compare the difference with those extracted from the DPCt set of images, although being superior, it seems to follow the same principles. The rate of accuracy varies within a range of [61.54%, 52.84%] for *AvgStd* [63.17%, 57.02%] for *SizeFreq* and [63.27%, 58.05%] for *AvgStd+SizeFreq*.

The fact that these results generally come close to 50% accuracy is logical seeing as we have to remember we are using simple metrics which only attend to variability of the pixels that conform to an image. We are going to come across the problem again in the set DPCv as we observe in the three cases training and validating with the same set we obtain results much superior in comparison with the rest of the experiments, with an accuracy capacity of 91,38% using *AvgStd+SizeFreq*. Even training with DPCt and validating with DPCv we obtain an accuracy rate of 80.23% also with *AvgStd+SizeFreq*.

We must remember that the best state of art results (training with DPCt and validating with DPCv) have been those by Luo et al. (2008) with 93% accuracy using characteristics based on clarity, brightness, simplicity, geometrical composition, colour harmony and the plane (identifying the background and foreground).

If we study the performance of the colour channels with these statistical features, the difference of accuracy becomes clear in the case of validation with the DPCv set itself. If we attend specifically to the colour channel saturation we obtain an accuracy rate of 83,85% which is more than a 20% difference in respect to the other combinations (see Figure 5).

This leads us to think that a great difference exists in the saturation between the images categorised as High and the images categorised as Low in the set DPCv. The opposite is the

Figure 4. Results obtained in all of the "train" and "valid" possible within the four datasets used

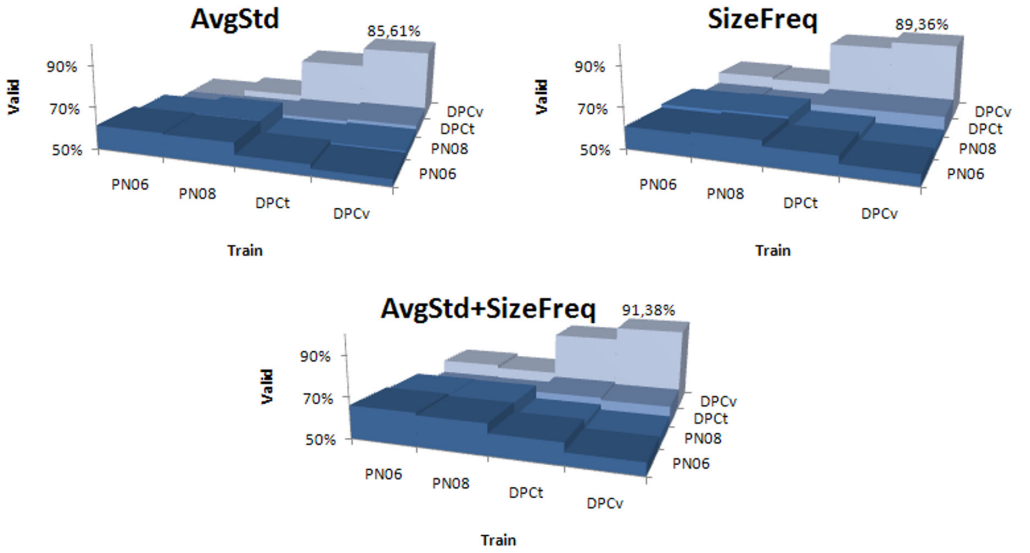
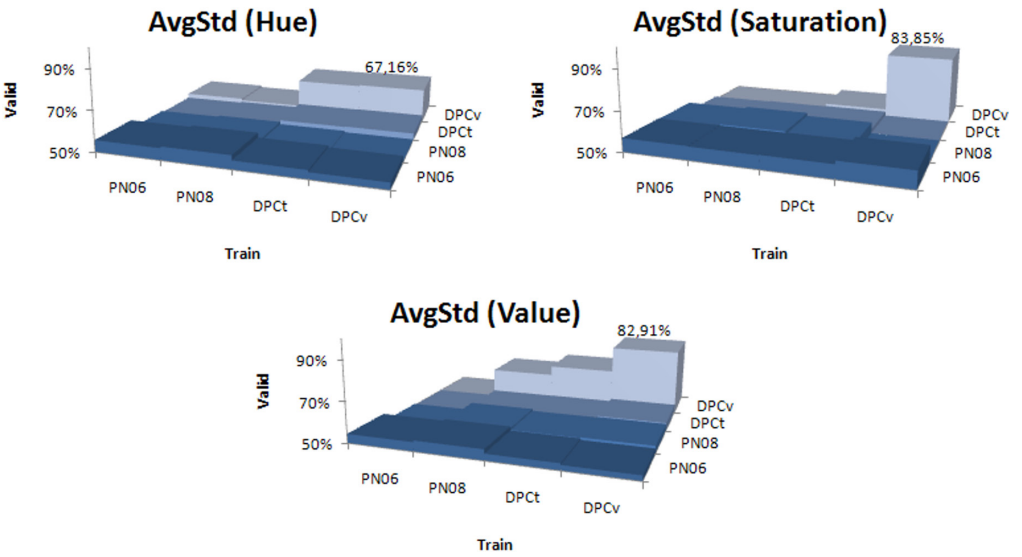


Figure 5. Results obtained accord to each colour channel using the statistical estimation in all possible combinations among the four datasets used



case in the combinations of the set DCPt with the sets PN06 and PN08, which are similar on the level of metrical statistics, entropy and the distinct colour channels.

6. DISCUSSION

In this work we have studied whether the choice of images in training can be carried out in a random way without significantly affecting the degree of generalization obtained in tasks of aesthetic classification. To demonstrate its importance we have used datasets of images employed by other investigators in different experiments. These photographs are categorised in terms of their aesthetic quality which has been determined by the evaluations of the specific specialist websites.

For this study we have created four binary classifiers trained through basic characteristics obtained for each one of the datasets. Subsequently, we have attempted to validate them with photographs coming from other datasets and not employed in the training phase of the classifiers. The results obtained suggest that the systems trained with a specific dataset may not achieve an acceptable level of generalization when in relation to the images which come from a different source. This indicates that the incorrect choice of the training set may derive from bias errors related to characteristics intrinsic to the images themselves, such as saturation.

It would be interesting to search for, or create beforehand, a set of images which could be adopted as a general standard. However, to our understanding, such a task is extremely complicated. It would have to take into account certain elements to attempt to remove errors such as those noted throughout this study.

In the case of the entry data we could opt for the use of images with a high number of evaluations, attempting to minimalize the bias of extreme scores. Also, we would recommend using distinct sets in the training and validation of the proposed systems. In this we refer, for example, to a set of images for completely training a classifier (including the phases of training, validation and testing) and a second set of images, different in terms of content and origin, to confirm the data and show the capacity

for generalization. We should likewise, to the degree possible, avoid the existence of basic component differences in the sample images (as has been seen with the colour channels). In this way we could look forward to simplifying the classification problem which we have undertaken.

ACKNOWLEDGMENT

This research was partially funded by: Xunta de Galicia, research project XUGA-PGIDIT-10TIC105008-PR; Spanish Ministry for Science and Technology, research project TIN2008-06562/TIN; Portuguese Foundation for Science and Technology in the scope of project SBIRC (PTDC/EIA–EIA/115667/2009).

REFERENCES

Arnheim, R. (1954). *Art and visual perception: A psychology of the creative eye*. University of California Press.

Arnheim, R. (1974). *Entropy and art: An essay on disorder and order*. University of California Press.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. doi:10.1006/jmps.1999.1279 PMID:10733860

Chang, C.-C., & Lin, C.-J. (2001). *LIBSVM: A library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In Proceedings of Computer Vision ECCV 2006 (pp. 288–301). Springer. doi:10.1007/11744078_23

Datta, R., Li, J., & Wang, J. Z. (2008). Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *Proceedings of Image Processing*. IEEE.

Ke, Y., Tang, X., & Jing, F. (2006). The design of high-level features for photo quality assessment. In *Proceedings of Computer Vision and Pattern Recognition*, (vol. 1, pp. 419–426). IEEE.

Luo, Y., & Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. In Proceedings of Computer Vision (ECCV 2008), (pp. 386–399). Springer.

Machado, P., & Cardoso, A. (1998). Computing aesthetics. In *Advances in artificial intelligence* (pp. 219–228). Springer. doi:10.1007/10692710_23

Machado, P., Romero, J., & Manaris, B. (2008). Experiments in computational aesthetics. In *The art of artificial evolution* (pp. 381–415). Springer. doi:10.1007/978-3-540-72877-1_18

Murray, N., Marchesotti, L., & Perronnin, F. (2012). Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of Computer Vision and Pattern Recognition* (CVPR), (pp. 2408–2415). IEEE. doi:10.1109/CVPR.2012.6247954

Romero, J., Machado, P., Carballal, A., & Osorio, O. (2011). Aesthetic classification and sorting based on image compression. In *Applications of evolutionary computation* (pp. 394–403). Springer. doi:10.1007/978-3-642-20520-0_40

Romero, J., Machado, P., Carballal, A., & Santos, A. (2012). Using complexity estimates in aesthetic image classification. *Journal of Mathematics and the Arts*, *6*(2-3), 125–136. doi:10.1080/17513472.2012.679514

Vapnik, V. N. (1997). The support vector method. In Proceedings of Artificial Neural Networks ICANN'97 (pp. 261–271). Springer.

Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with java implementations. *SIGMOD Record*, *31*(1), 76–77. doi:10.1145/507338.507355

Wong, L.-K., & Low, K.-L. (2009). Saliency-enhanced image aesthetics class prediction. In *Proceedings of Image Processing* (ICIP), (pp. 997–1000). IEEE.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.

Adrian Carballal holds a BSc and a PhD in Computer Science from the University of A Coruna (Spain) where he works as post-doctoral research associate at the Department of Information Technologies and Communications and as Part-Time lecturer. Has authored over 10 articles and edited 3 journals. He has also participated as researcher in 5 funded research proposals. His main research interests include Image Processing and Computer Graphics.

Luz Castro Pena holds a B.Sc. in Computer Engineering by the Universidade da Coruña where she currently lectures on Multimedia and Web Development. She is also co-founder and CEO of imaxin|software. Rebeca Perez is an advertising agent and a PHD student in A Coruña (Spain). Her professional skills include graphic and web design, marketing management, sales and photography.

Rebeca Perez is working as external assistant on Image Processing and Computer Graphics researchs at the Department of Information Technologies and Communications in A Coruña.

João Correia is a PhD. student of the Doctoral Program in Information Science and Technology at the University of Coimbra. He also holds a BS and MSc in Informatics Engineering from the University of Coimbra. His main research interests include Evolutionary Computation, Machine Learning, Computational Creativity, Pattern Recognition and Computer Vision.