Exploiting Photographic Style for Category-Level Image Classification by Generalizing the Spatial Pyramid

Gemert Jan C. Van

Exploiting Photographic Style for Category-Level Image Classification by Generalizing the Spatial Pyramid

Jan C. van Gemert*
Puzzual
Oudeschans 18
1011LA, Amsterdam, The Netherlands
jan@puzzual.com

ABSTRACT

This paper investigates the use of photographic style for category-level image classification. Specifically, we exploit the assumption that images within a category share a similar *style* defined by attributes such as colorfulness, lighting, depth of field, viewpoint and saliency. For these style attributes we create correspondences across images by a generalized spatial pyramid matching scheme. Where the spatial pyramid groups features spatially, we allow more general feature grouping and in this paper we focus on grouping images on photographic style. We evaluate our approach in an object classification task and investigate style differences between professional and amateur photographs. We show that a generalized pyramid with style-based attributes improves performance on the professional Corel and amateur Pascal VOC 2009 image datasets.

Categories and Subject Descriptors

H.3.1 [**Information Storage And Retrieval**]: Content Analysis and Indexing—*image classification, image aesthetics*

General Terms

Imaging, Classification, Retrieval, Indexing, Aesthetics

Keywords

Image classification, Photographic style, Spatial pyramid

1. INTRODUCTION

There is a relation between the composition of a photograph and its subject. Similar subjects are typically photographed in a similar style [11]. Depending on the subject, several heuristic photography rules may apply, including: object placement, the rule of thirds, a varying depth of field, *etc*. Consider figure 1, where similar images share a compositional style. A photographer's use of these styles, however, is influenced by the shape, appearance, and natural surroundings of the subject. For example, a long object is often photographed landscape, a colorful subject may be contrasted against a bland background and a small object typically has low depth of field. Such photographic compositions are often shared when subjects are of the same class.

Figure 1: An example of similar photographic subjects sharing a composition. The *Antelope* images in the top row share depth of field and positioning, whereas the *Bus* images in the bottom row have similar viewpoint and color contrast.

In this paper we investigate the hypothesis that similar photographic style within object categories can be used to improve object classification. As far as we know, this hypothesis is unexplored in the literature. The state of the art [5] in category-level object classification is the bag-of-words model and its variants [10, 13, 22, 24]. In this model, local image features are represented by discrete prototypes (visual words) describing an image as a histogram of prototype counts. Prototype-histograms are subsequently used by a classifier to separate images into object categories. Because of the state-of-the-art performance of the visual word model, we use it as our baseline and extend it with photographic style similarity matching.

The use of photographic style in an image can be described by attributes such as *colorful*, *in focus*, *well-composed*, *etc*. Other, more object-centered types of descriptive attributes have recently been used for object and scene classification [6, 7, 12, 25]. Examples of the attributes used are *striped*, *furry*, *has wheel*, *has head*, *etc*. These attributes provide a middle ground between low-level features and high-level categories.

*Work done while at the Willow group, École Normale Superiéure, Paris, France. Sponsored in part by the 2ACI DGA project and the VideoWorld ERC grant.

ICMR '11, April 17-20, Trento, Italy

By attribute sharing, only a little amount of training data is required [6]; it even allows classification with a disjoint training set [12]. In this paper, we are interested in photographic style attributes that aid in object classification. To this end we will design specific style attributes, and use these attributes in combination with the visual word model for image classification.

To incorporate photographic style similarity in the visual word model we draw inspiration from the spatial pyramid introduced by Lazebnik *et al*. [13] who in turn extend Grauman and Darrell [8]. The spatial pyramid quantizes the absolute position of visual words in an image to fixed spatial regions. For example the pyramid at level 1 has the quadrants up-left, up-right, low-left, low-right of an image. Higher levels of the pyramid are obtained by quantizing the absolute position in regions of decreasing size. The spatial pyramid creates correspondences between visual words quantized to the same equivalence class (fixed spatial region). In our case, however, instead of relating similar position, we are interested in creating correspondences between similar photographic styles. Therefore, we follow Lazebnik *et al*., by quantizing visual words in equivalence classes. However, where they use the absolute position, we are interested in compositional attributes such as colorfulness, lighting, depth of field, viewpoint and saliency. These compositional attributes may be quantized in equivalence classes based on photographic attributes such as saturation, brightness, blur-level, *etc*. Assigning visual words to these equivalence classes allows us to create correspondences between similar photographic styles.

The contributions of this paper are fourfold. First, we introduce the use of photographic style for category-level object classification. The main assumption of this paper is that similar objects are photographed in a similar manner. Second, we extend the spatial pyramid to a more general version based on equivalence classes, with the spatial pyramid as a special case. Third, we propose several style attributes for creating such equivalence classes. Fourth, we investigate whether photographic style features behave differently on amateur images than on professional photographs.

In the next section we present work related to photographic style classification and automatic photo composition. In section 3 we give our approach for incorporating photographic style in object classifation. Section 4 contains the experimental validation and is followed by our conclusions.

2. RELATED WORK

Photographic style features can be used for separating professionally made photographs from amateur images. One of the first approaches to this problem is by Tong *et al*. [20] who feed global features based on blur, contrast, colorfulness and saliency in several classifiers to distinguish between amateur images and a professional image set. Further research [4, 9, 18] adds various other global image features including blur, average saturation, average brightness and several color- and texture-based features. Feature selection shows that blur is a good performing feature because it determines depth of field. The work by Luo and Tang [16] moves away from global image features. The authors detect a bounding box in the image as the main subject area by determining what is in focus by a blur detection algorithm [14]. The classification results outperform previous work by using features based on the ratio in the subject area versus the image background with features such as brightness, contrast and color. In our work, however, we are not interested in using style features to separate professional photographs from amateur ones, but in using style features to aid image classification. Instead of the compositional differences between images we exploit their compositional similarities. Nevertheless, we draw inspiration from the proposed photographic style features and use local features in the visual word model for creating correspondences between style attributes.

Photographic style is also used for automatically finding good compositions in images. Such automatic compositions provide a user with a "touch-up" button to help enhancing photographic image quality. Automatic compositions can be achieved by directly applying heuristics such as the rule of thirds and blurring of the background [2]. In [15] such heuristic rules are combined with saliency detection in an aesthetic score function. In [3], the authors combine saliency with the GIST image descriptor in a stochastic search algorithm to find a composition that matches one of several well-composed reference images. From these works we can use the heuristics and the proposed saliency. However, these papers optimize an aesthetic function to find a good composition in a single image. In contrast, we use composition similarity between images and refrain from labeling compositions as good or bad.

3. APPROACH

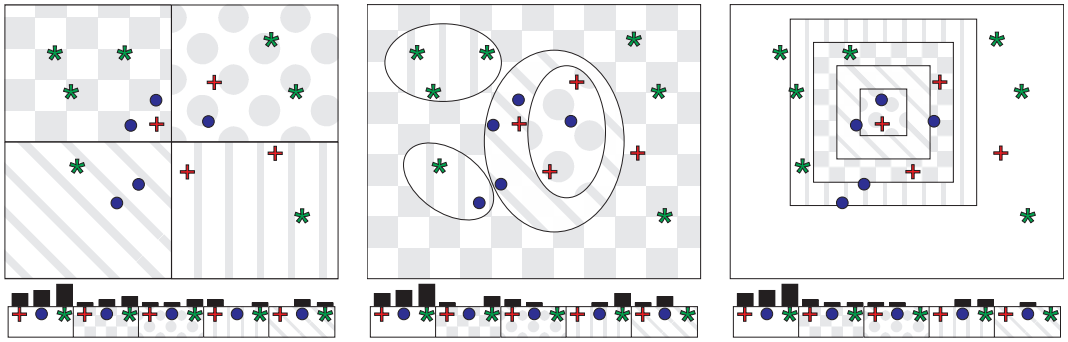
We aim to exploit similarities in photographic style for image categorization. To this end we create correspondences within the visual word model between similarly-styled image attributes. Examples of such style attributes are colorfulness, intensity and depth of field. To incorporate correspondences between these styles we use techniques inspired by spatial pyramid matching [13]. However, where the spatial pyramid uses approximate global geometric correspondences between visual words, we are interested in approximate style correspondences. In the next section we will briefly review the spatial pyramid, after which we present our generalization to a style pyramid, and the style attributes themselves.

The spatial pyramid [13] by Lazebnik *et al*. is based on the pyramid matching scheme by Grauman and Darrell [8]. Whereas Grauman and Darrel use coarse-to-fine bins in feature space, Lazebnik *et al*. create a pyramid based on the spatial image layout. The spatial pyramid repeatedly divides an image into fixed sub-regions of finer resolution, where spatial pyramid level $\ell \in \{0, \ldots, L\}$ has $R(\ell) = 2^{2\ell}$ sub-regions. In image X all features are assigned to their best visual word index v, selected from a vocabulary V. The frequency of visual word v inside sub-region i of image X is given by the histogram bin $H_X^i(v)$. Similarity between images X and Y on level ℓ of the spatial pyramid [13] is given by histogram intersection

$$I^\ell(X,Y) = \sum_{i=1}^{R(\ell)} \sum_{v=1}^{|V|} \min\left(H_X^i(v), H_Y^i(v)\right). \quad (1)$$

After reweighting larger sub-regions on a lower pyramid level, the final spatial pyramid becomes

$$\kappa^L(X,Y) = \frac{1}{2^L} I^0(X,Y) + \sum_{\ell=1}^{L} \frac{1}{2^{L-\ell+1}} I^\ell(X,Y). \quad (2)$$



(a) (b) (c)

Figure 2: (a) Spatial pyramid, (b) Style pyramid, (c) Local co-occurrence pyramid. The star, dot and plus denote three visual word types. The striped, checkered and polka dot patterns represent four equivalence classes. Below the example image are the histograms of the visual words, grouped by equivalence class. Note that the first equivalence class histogram (in white) represents the whole image.

See figure 2(a) for an example of the spatial pyramid.

The spatial pyramid builds correspondences by quantizing the absolute position of an image feature in disjoint equivalence classes. For photographic style correspondences we propose to use the same approach, only for more general equivalence classes based on style. As a running example, let us assume that color saturation is a measure of colorfulness. Then, we propose to quantize the amount of saturation in disjoint equivalence classes, and create correspondences between these equivalence classes. If color saturation ranges between $[0, \ldots, 1)$ then an example of two disjoint equivalence classes is given by $\left\{[0, \ldots, \frac{1}{2}), [\frac{1}{2}, \ldots, 1)\right\}$. Each image feature can be assigned to an equivalence class depending on its average amount of saturation. A pyramid may be obtained by creating multiple levels of equivalence classes. More formally, for a visual word v and a scalar style attribute function $S(v)$ with a value range $[a, \ldots, b)$, we quantize visual words on level ℓ of the pyramid in $R(\ell) = 2^\ell$ equivalence classes, where the visual words for in each equivalence class $\{0, ., i, ., R(\ell) - 1\}$ are given by

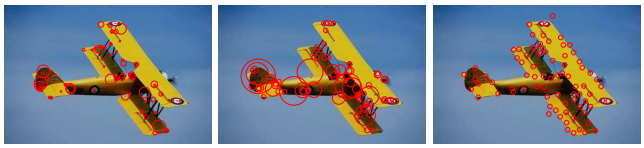
$$E(i) = \left\{ v \in X \mid a + i\frac{b-a}{R(\ell)} \;\leq\; S(v) \;<\; a + (i+1)\frac{b-a}{R(\ell)} \right\}. \tag{3}$$

A style pyramid can be created by replacing the image similarity for the spatial pyramid in equation 1 with the similarity between images X and Y on level ℓ of the style pyramid by

$$I^\ell(X, Y) = \sum_{i=1}^{R(\ell)} \sum_{v=1}^{|V|} \min\left(H_X^{E(i)}(v), H_Y^{E(i)}(v)\right), \tag{4}$$

where $H_X^{E(i)}(v)$ denotes the frequency of visual word v in equivalence class $E(i)$ for image X. Note that the spatial pyramid is a special 2-dimensional case of this approach, where the equivalence classes are based on the x and y position. See figure 2(b) for an example of a style pyramid.

Creating correspondences between photographic styles is now a matter of matching a photographic style to a function $S(v)$. By quantizing this style function $S(v)$, we match similar styled visual words to each other. By relating styles, we aim to relate images that share a similar photographic composition.



(a) (b) (c)

Figure 3: Example of interest point detection. (a) Harris-Laplacian (b) Hessian-Laplacian. (c) Color boosted Hessian. Note that we only show 80 points for clarity. Best viewed in color.

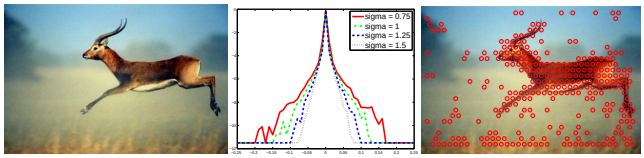
3.1 Style Attributes

Inspired by the related work in section 2, we selected several style attributes and their corresponding style function $S(v)$. A visual word v represents a patch of pixels in the image. The style values for these pixels are Gaussian weighted and averaged to obtain a single value S(v), unless stated otherwise. These style attributes give us the equivalence classes to match similarly styled images.

Salient Points. The visual word model treats an image as a bag of local features and uses the occurrence counts of feature prototypes for classification. The use of local features has also proved beneficial for classifying professional versus amateur photographs [16]. Therefore, we base our work solely on local features (standard SIFT). We use dense sampling of image features, since this has proven to give good classification results [13, 22]. Besides dense sampling, local features can be detected based on interesting, or salient image structures such as corners and blobs. Since salient points represent interesting image structure, their occurrence, or lack of occurrence, influences the photographic style. Consider for example figure 3, where all salient points are found on the subject, which is framed by featureless sky. To capture such style similarities we create correspondences between various types of interest point detectors. We use the Harris-Laplace and Hessian-Laplace detectors [17] and a detector based on color interest points [23]. We put each visual word that originates from the same detector in its own equivalence class. In figure 3 we show examples of the used salient point detectors.

(a) (b) (c)

Figure 4: Example of saliency maps. (a) Original image. (b) Frequency-tuned saliency map. (c) Isocentric saliency map. The saliency level is given by the brightness.



(a) (b) (c)

Figure 5: Example of blur detection. (a) Original image (b) Log histograms of the image derivatives for various width of the kernel (sigma). (c) For densely sampled points, the non-blurred points are given in red. Best viewed in color.

Saliency Maps. Professional photographs typically have a clear subject-background separation, whereas amateur ones may have various distracting elements. Such distracting, or salient, regions may be automatically detected. We use two recent salient region detection algorithms. One of these detectors is based on shape and color [21]. The other detector is based on frequency, color and luminance and outperforms other approaches in an object segmentation task [1]. We create equivalence classes based on the saliency values. Note that most work on object classification with saliency keep only the salient regions. In contrast, we create equivalence classes based on saliency, keeping non-salient regions in a class of their own. Hence, our approach retains features that are non-salient. Moreover, if they are consistently non-salient within an object category, such features will still aid in object classification. In figure 4 we show an example of the two feature maps. For our experiments we use $\sigma = 5$ for both methods, and curvedness = 5 for [21].

Viewpoint. We model object viewpoint with a local co-occurrence pyramid of visual words. Consider for example the *bus* images in figure 1. Similarity in viewpoint may be inferred from the co-occurrence of the wheels at a certain distance. Thus, we use a local co-occurrence pyramid to model the viewpoint. Specifically, we build on the work of Savarese *et al.* [19] in their efficient approach to compute visual word correlograms. A correlogram is a binary triangle matrix of size $\frac{|V|*|V+1|}{2}$ that expresses the co-occurrence of a pair of visual words at a certain distance d. We extend Savarese *et al.* [19] by our pyramid of equivalence classes in equation 4. *I.e.*, we calculate the co-occurrence of a pair of visual words (i, j) as $d_1 \leq \text{mindist(i,j)} < d_2$, where mindist(i,j) is the minimum distance along the x or y axis between visual word i and j in pixels. In figure 2(c) we illustrate the local co-occurrence pyramid.

Depth of Field. The important parts of an image are typically in focus. This is probably the reason why depth of field by blur detection is a good-performing feature for separating professional images from amateur photographs [4, 9, 16]. To create correspondences for depth of field, we use the degree of blur as a style function. We implement the method by [16], who extend the horizontal blur detection method in [14] with vertical blur detection. Where they use a uniform kernel we use a Gaussian kernel to extend beyond discrete blur levels. The blur detection approach is based on natural image statistics of derivative filters. Derivative filters measure edge strength which intuitively is inversely related to the blur level. The blur level of a local window in the image is found by the maximum likelihood over a range of image derivative kernel sizes. Specifically, let the derivative distribution $p^\sigma(i)$ of the pixels i in image I be given by $p^\sigma(i) \propto \text{hist}(d^\sigma_{xy} * I)$, where d^σ_{xy} is a 2D Gaussian derivative kernel with kernel size σ. Then the log-likelihood of a blur level σ for a window of pixels W in image I is $l^\sigma(W) = \sum_{i \in W} \log p^\sigma(i)$. In our case, each visual word corresponds to a set of pixels in the image, we use these pixels for W. The blur level k over a range of blur levels K of window W is given by the maximum likelihood $k = \arg\max_{\sigma \in K} l^\sigma(W)$. We show a blur detection example in figure 5.

Rule of thirds. One of the basic rules of thumb in photographic style is the rule of thirds [11]. This rule states that the subject should be located at one of the intersections of the three equally spaced horizontal lines with three equally spaced vertical lines. Effectively, the image has four of these intersections, located in each quadrant. These quadrants are also the equivalence classes for level 1 of the spatial pyramid. Therefore, we will use the standard spatial pyramid to take the rule of thirds into account. In figure 2(a) we illustrate the spatial pyramid. Note that the four equivalence classes each capture an intersecting line from the rule of thirds.

Colorfulness. Color is a powerful cue for contrasting the subject from the background. The background may be less colorful to make the subject stand out, as for example the Cat of the Corel collection in figure 6, or alternatively, the background may be more colorful as for example the Airplane of the Corel collection in figure 6. We do not focus on only the colorful regions, rather we provide the machinery for features to match to approximately the same level of colorfulness. For measuring colorfulness we use the saturation from the HSV color space. Note that the saturation is different from the hue, or from a RGB color histogram. The hue or the histogram would create correspondences between similar colors, whereas the saturation creates correspondences between similar colorful areas irrespective of the color itself. Note that the SIFT descriptor we use, only takes the intensity channel into account. In figure 7 we show an example of local features split by saturation level.

Lighting. Similar to colorfulness, the lighting of an image region may be related to its importance in the photograph. Hence, for similar objects, the lighting level may be the same. Therefore we use the brightness by the Value channel from the HSV color space as a measure of lighting.

Figure 6: Example images of the amateur set (Pascal VOC 2009) and the matching categories of the professional set (Corel).

Figure 7: Example of equivalence classes with densely sampled points for saturation and brightness. Each style is split in four equivalence classes, where low to high values are given by the yellow circle, green plus, blue star and red cross respectively. Best viewed in color.

Note that the SIFT descriptors themselves are invariant to brightness changes. However, by creating correspondences between brightness levels we re-introduce some sensitivity to brightness. In figure 7 we show an example of local features split by brightness level.

4. EXPERIMENTS

We experimentally validate if style features benefit category-level object classification. Moreover, we expect that photographic style is more pronounced in professional photographs than in amateur images. To investigate this hypothesis we compare classification performance between a professional and an amateur photograph collection. For the amateur collection we use the Pascal VOC set [5]. This is a well-known set for image categorization and consists of 20 categories of *Flickr* images. *Flickr* images are typically amateur images, uploaded by random users on the internet. For the professional set we use similar categories from the Corel collection. We tried to match the categories of the Pascal VOC set in the Corel collection. In figure 6 we show an example per category of the two sets.

4.1 Experimental Setup

The Corel set has 100 images per category, with a total of 2000 images. For the Pascal VOC set we combine the 3473 images in the specified train set with the 3581 in the validation set to a single set of 7054 images. We split the Corel and the VOC set in 10 random train and test sets, distributing the positive examples equally over train and test set. Instead of the standard single test set given for the VOC we use 10 repetitions to compute average and standard deviations which allow significance testing. As the performance measure we adopt Average Precision (AP), as it is commonly used for the Pascal VOC. For features we only use standard SIFT. When evaluating salient points we add a minimum baseline of densely sampled SIFT features at every 10 pixels with a Gaussian weighted window size of $\sigma = 2$. For the remaining style attributes we use all available salient point detectors. Our visual word vocabulary size $|V| = 2000$, and is created per training set for each of the 10 splits with K-means on 200.000 randomly sampled SIFT features. For classification, we use libsvm and use its built in cross-validation method to set the best C parameter. For the levels of the style pyramid we use two levels, with 2 and 4 equivalence classes. For the viewpoint feature that uses local co-occurrence, we group all local features from 0-60 pixels with steps of 15 pixels. Thus, this pyramid also includes the levels for 0-30, and 30-60 pixels.

4.2 Results

In figure 8 we show results in Average Precision (AP) for all style attributes per category for the amateur Pascal VOC set and professional Corel set. Note that the lesser amount of within-class variation in Corel causes much better overall performance than the Pascal VOC. Some classes in Corel are hard to improve because they are close to perfect (1.0) AP (Engine, Drinks, Buses, Cats, Door, Bonsai). For the professional set there are 21 style features which significantly increase results, whereas the amateur set has 49 significant increases. One reason for this are the six close to perfect categories in Corel that are hard to improve. Furthermore, the standard deviations for the Corel set is larger and therefore harder to significantly improve, which is due to the smaller size of the set. The *spatial pyramid* and *viewpoint* generally always improve results for Pascal VOC. The

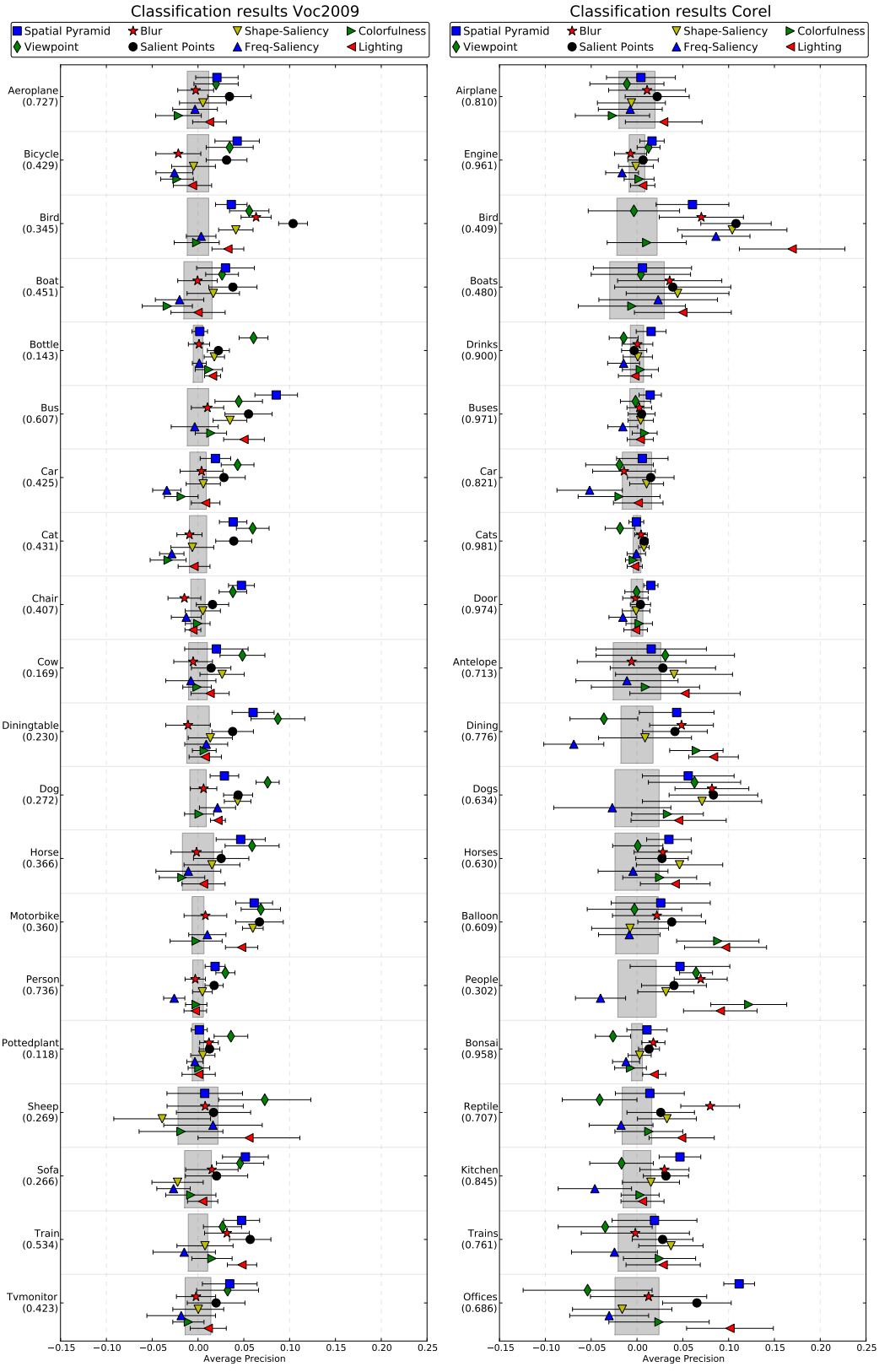


Figure 8: Results on Pascal VOC and Corel for the style pyramid for all style attributes per category. The baseline (no style) is in gray where the width of the gray box denotes standard deviation. The x-axis denotes deviation from the baseline where the baseline mean is given on the y-axis in brackets behind the category name. Corresponding categories between the Pascal VOC and Corel are grouped as in fig 6.

Table 1: Results in Mean Average Precision$*100$ for single and full pyramid levels. The best results are given in bold and a significant increase is underlined.

	Pascal VOC 2009 Baseline (Level 0): 38.5 ± 0.5				Corel Baseline (Level 0): 74.6±0.7			
	Level 1		Level 2		Level 1		Level 2	
Style	Single	Pyramid	Single	Pyramid	Single	Pyramid	Single	Pyramid
Lighting	38.3±0.5	39.5±0.5	38.0±0.3	40.2±0.3	77.1±0.9	78.0±0.8	76.9±0.9	**79.0**±0.8
Colorfulness	35.4±0.5	38.2±0.6	34.6±0.4	37.8±0.6	74.1±0.8	76.6±0.8	73.6±0.8	76.5±0.8
Blur	38.9±0.3	39.9±0.4	35.6±0.5	39.0±0.4	77.1±0.7	77.6±0.6	73.8±0.8	77.1±0.8
Freq. Saliency	36.3±0.4	38.5±0.4	34.9±0.5	37.7±0.5	71.4±0.8	74.4±0.8	69.3±0.8	73.1±0.7
Shape Saliency	37.3±0.5	39.1±0.4	37.7±0.6	39.7±0.4	74.3±0.8	76.0±0.9	75.0±0.9	76.8±0.8
Salient Points		42.0±0.8				77.8±0.6		
Spatial Pyramid	40.7±0.6	42.0±0.4			76.1±0.7	77.5±0.6		
Viewpoint	42.7±0.4	**43.6**±0.4	42.4±0.4	**43.4**±0.4	73.3±1.1	75.5±0.9	72.2±0.9	74.1±0.9

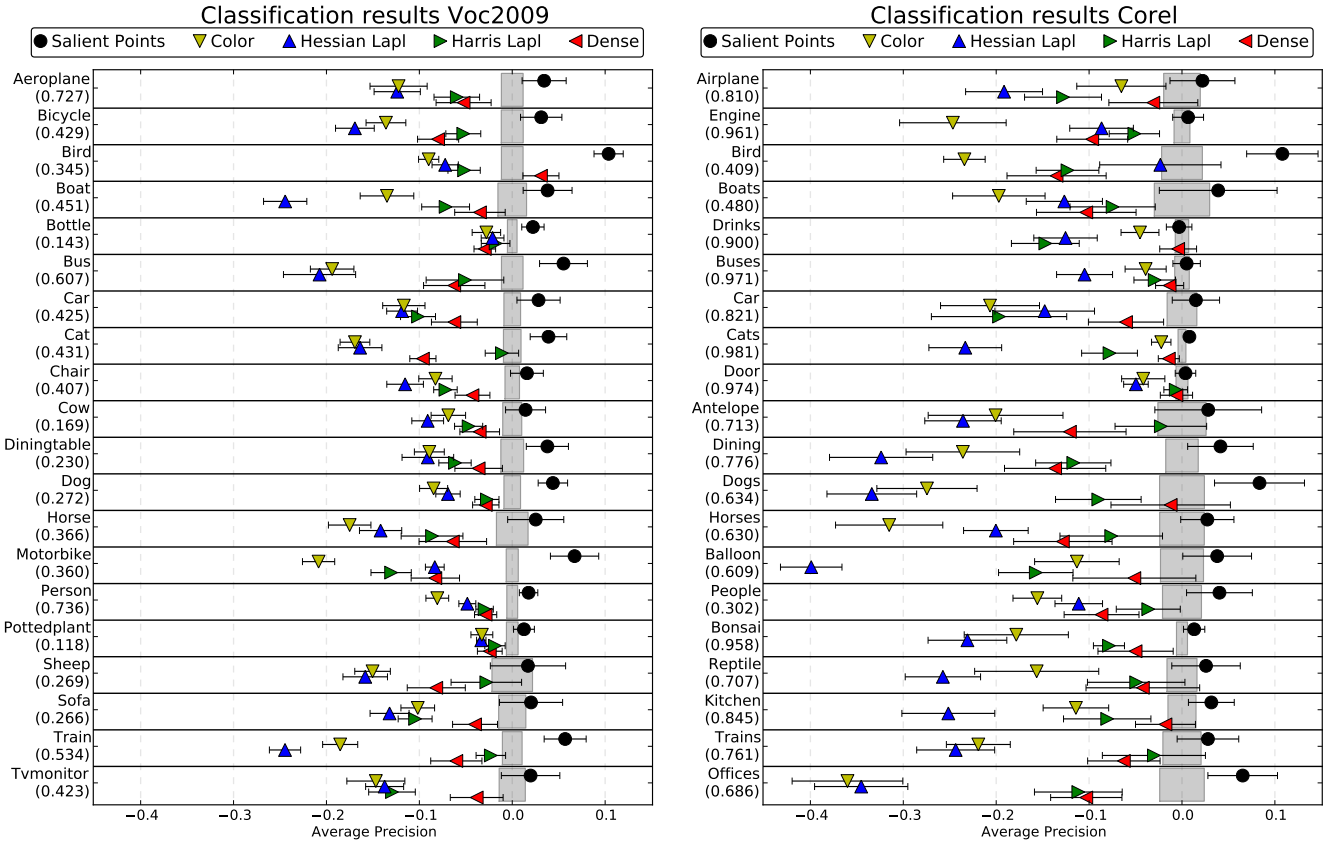


Figure 9: Results on Pascal VOC and Corel for individual salient point detectors and the salient points style pyramid. The baseline (addition of all detectors) is in gray where the width of the gray box denotes standard deviation. The x-axis denotes deviation from the baseline where the baseline mean is given on the y-axis in brackets behind the category name. Corresponding categories between the Pascal VOC and Corel are grouped as in fig 6.

spatial pyramid is better in global shape (bus, train) whereas *viewpoint* is better for local appearance (bottle, dog, pottedplant, sheep). For the Corel set, *viewpoint* often performs under the baseline, and the *spatial pyramid* is only really helpful for obvious contextual scene categories (bird, door, kitchen, offices).

In table 1 we summarize the classification performance for Pascal VOC and Corel over the 10 random draws of the data, measured in mean average precision for the full pyramid, and for single pyramid levels (without the underlying levels). Results for both image sets show no clear preference for a higher level pyramid, but a pyramid is always better than a single level, confirming results of [13, 19]. Most style features improve results, however the least helpful attributes are the saliency maps, where the *frequency saliency map* never helps (except for bird in Corel) and *shape saliency*

only rarely improves performance. This suggests that automatic detected image saliency is not very consistent on the category level. For the professional images in Corel, the best performing style feature is *lighting*. This illustrates that simply grouping pixels on brightness level with our generalized pyramid can already outperform the well-known spatial pyramid. For the other style attributes, only the *frequency saliency map* and the *viewpoint* do not significantly improve results. For the amateur images in Pascal VOC the *viewpoint* is the best performing attribute, and *colorfulness* and *frequency saliency map* do not significantly improve results. Note that the best performing style attribute on the amateur images (*viewpoint*) does not help for the professional images. This may be the case because the professional images are the center of attention and typically fill the entire image. Hence, local viewpoint differences capture only part of the object. In contrast, the objects in the amateur images are more cluttered which makes objects share a local context from the photographer's viewpoint. In a similar vain, the best performing attribute on the professional set (*lighting*) does not help for the amateur images. This confirms the hypothesis that objects in professional images share similar lighting whereas amateur images do not. A similar, but less strong case, can be made for *blur* and *colorfulness*, where results improve for the professional set however do not help for the amateur set.

In figure 9 we show results for each salient point detector for adding them all together, and for the salient point pyramid which puts each detector type in a equivalence class. Generally speaking, adding all detections (baseline) outperforms each individual detector. Only for Bird, in the VOC set, dense sampling outperforms the baseline. This may be explained because dense sampling mostly takes the important context (sky) into account. Putting each detector type in a equivalence class performs even better since context and subject information are both present but not wrongly matched against each other as in the baseline. What is more, the salient point pyramid always improves results for all categories for both image sets.

5. CONCLUSIONS

We presented a method to exploit style correspondences between images for object-level image classification. We achieved this by a generalization of the spatial pyramid by creating equivalence classes between approximately similar style attributes. We experimentally evaluated our method on a professional set and an amateur set of images. The results show that several style-based attributes improve performance over the baseline for both sets. Grouping salient point types and global composition benefits both sets. For amateur images, the local object configuration is most important whereas for professional images the colorfulness, the depth of field and the lighting is most beneficial.

6. REFERENCES

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned Salient Region Detection. In *CVPR*, 2009.

[2] S. Banerjee and B. Evans. In-camera automation of photographic composition rules. *Trans. Image Processing*, 16(7), 2007.

[3] Y.-Y. Chang and H.-T. Chen. Finding good composition in panoramic scenes. In *ICCV*, 2009.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 Results, 2009.

[6] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[7] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007.

[8] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

[9] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

[10] F. S. Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *ICCV*, 2009.

[11] B. P. Krages. *Photography : the art of composition*. Allworth Press, 2005.

[12] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[14] A. Levin. Blind motion deblurring using image statistics. In *NIPS*, 2006.

[15] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. *Proceedings of Eurographics*, 29(2), 2010.

[16] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008.

[17] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

[18] A. K. Moorthy, P. Obrador, and N. Oliver. Towards computational models of the visual aesthetic appeal of consumer videos. In *ECCV*, 2010.

[19] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR*, 2006.

[20] H. Tong, M. Li, H. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *In Proceedings of Pacific Rim Conference on Multimedia*, 2004.

[21] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *ICCV*, 2009.

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

[23] J. van de Weijer, T. Gevers, and A. Bagdanov. Boosting color saliency in image feature detection. *TPAMI*, pages 150–156, 2006.

[24] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *TPAMI*, 32(7):1271–1283, 2010.

[25] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *Int. J. Comput. Vision*, 72(2):133–157, 2007.