This is a preliminary version of an article published by
S. Kopf, M. Zrianina, B. Guthier, L. Weiland, P. Schaber, S. Ponzetto, W. Effelsberg
Enhancing Bag of Visual Words with Color Information for Iconic Image Classification
20th International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV), 2016

Enhancing Bag of Visual Words with Color Information for Iconic Image Classification

Stephan Kopf[1], Mariia Zrianina[1], Benjamin Guthier[1], Lydia Weiland[2],
Philipp Schaber[1], Simone Ponzetto[2], Wolfgang Effelsberg[1]
[1] Department of Computer Science IV, [2] Data and Web Science Group
University of Mannheim, Germany
Email: {kopf, zrianina, guthier, lydia, schaber, simone, effelsberg}@informatik.uni-mannheim.de

***Abstract*—Iconic images represent an abstract topic and use a presentation that is intuitively understood within a certain cultural context. For example, the abstract topic "global warming" may be represented by a polar bear standing alone on an ice floe. This paper presents a system for the classification of iconic images. It uses a variation of the Bag of Visual Words approach with enhanced feature descriptors. Our novel color pyramids feature incorporates color information into the classification scheme. It improves the average F1 measure of the classification by 0.118.**

Keywords: semantic image search, iconic images

I. Introduction

When searching for multimedia content, image search engines like Google Images or Flickr find a large number of pictures. Most commercial search engines rely heavily on a textual description that surrounds the content, for example on a Web page or that was added manually. These search engines work well, if the topic to be searched for can be labeled with a brief and meaningful description. However, it is not always easy for users to find suitable keywords to be used in the search. This becomes even more challenging when searching for abstract topics like "climate change". Such a search request may be answered by a large variety of multimedia content. Images may show reasons for climate change, e.g., "air pollution" or possible solutions like "wind turbines". Figure 1 shows two example results of such a search query.

In this paper, our aim is to search for *iconic images*. In an iconic image, the visualized objects are not relevant on their own, but the complete scene represents a larger, more abstract topic which is understood intuitively within a certain cultural context. An example would be the picture of a polar bear standing on an ice floe. In many western countries, such an iconic picture represents global warming, and it has been used in this context for years. Both photographs in Figure 1 are typical examples of iconic images as well. Smokestacks can be considered iconic if they are associated with the topic of "air pollution", which in turn represents the larger theme "climate change". A lot of previous work (e.g., [1, 2]) focuses on identifying *canonical* (representative) views of similar scenes. Some authors label these canonical images as *iconic*. We do not follow this definition of iconic images.

Fig. 1. Two iconic images of climate change. Smokestacks (causal attribution of climate change) and wind turbines (proposed solution) [3].

Our long-term goal is the automatic classification of rich semantic concepts in images. This paper goes one step towards our goal and presents a complete system that allows the automatic classification of iconic images. We use the bag of visual words (BoVW) method as a starting point. Our color pyramid scheme improves the basic algorithm and increases its accuracy for iconic image search.

The rest of this paper is structured as follows. Section II discusses methods that extend the basic BoVW algorithm and other approaches of classifying iconic images. The used algorithm as well as the developed color pyramid feature are described in Section III. Sections IV and V present our dataset and the experimental results, respectively. Section VI summarizes the paper.

II. Related Work

Much work has been done in the area of image classification and content-based image retrieval, most of which focuses on the retrieval of specific objects [4, 5, 6] or scene categories [7, 8]. One of the most successful techniques is the Bag of Visual Words (BoVW) approach which has been widely studied in the literature [9, 10].

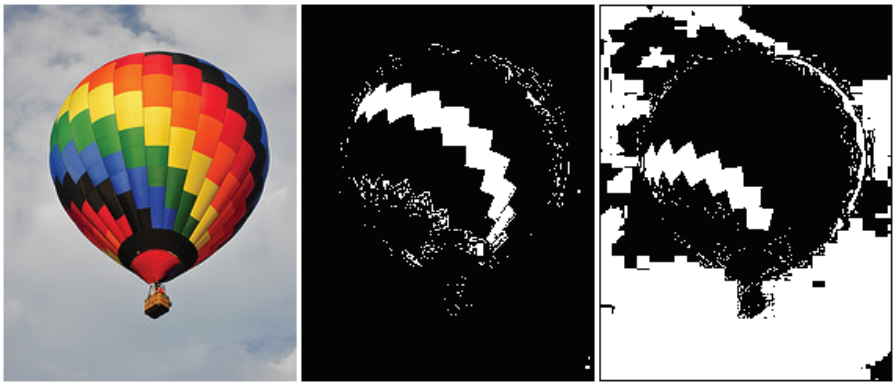

Fig. 2. An example of color masks. Left: original image. Center: color mask with hue value 30 (orange). Right: color mask with hue value 110 (blue). Color masks are computed with a range of 10. The dilation filter has not been applied in this example.

Several improvements for the general BoVW approach have been proposed. Lazebnik et al. [11] additionally include the location of visual words and use it alongside the visual histograms that represent the frequency of visual words for each image. Sharma [12] extended this method by adding saliency information. The computed features are weighted with their corresponding value in the saliency map.

Using the definition of [3], an iconic image concisely represents an entity that refers to a larger topic, and that is widely used in public communication. Such a topic is identified easily by media users and can trigger a substantial affective, cognitive and/or behavioral reaction. The only work in the context of image retrieval that considers iconic images was proposed by Ponzetto et al. [3]. The approach starts with a human-selected basic set of iconic images along with their caption. This basic set is enlarged by using a query-by-text approach from the images' descriptions. Outliers are filtered out in the final step.

III. Classification System

In the following subsection, the basic BoVW approach is briefly discussed, followed by our proposed extension that uses color pyramids. Implementation details are given at the end of this Section.

A. Bag of Visual Words Approach

Building a vocabulary of visual words includes the detection of keypoints, the computation of descriptors, and clustering. We use the SIFT and the SURF detectors and descriptors in our system. The computationally efficient GRID detector was also added. It divides the image into equally sized cells where the center of each cell is considered as a point of interest. The descriptors that are obtained from a training set of images are clustered using the k-means algorithm. The set of computed centroids then defines the vocabulary of visual words. Our system allows the user to manually define the vocabulary size.

The next step is the training of a classifier on the labeled training dataset. We implemented the two classifiers: Support Vector Machines (SVM) and Normal Bayes Classifier (NBC). To predict a label for a new, unknown image, a visual word histogram for this image is calculated and passed to the trained classifier.

B. Color Pyramids Feature

Feature descriptors like SIFT or SURF use the local contrast but do not consider color information. We propose *color pyramids* as a novel feature to enhance the basic BoVW method with color information. The idea of color pyramids was motivated by the concept of spatial pyramids as presented by Lazebnik et al. [11]. Instead of dividing an image into spatial sub-regions, it is divided into color sub-regions (e.g., assigning a color to each pixel). If coarse to fine color intervals are used, a hierarchy is created that is similar to spatial pyramids. E.g., the colors red, yellow, green, or blue may be used on the coarsest level, and the next level splits each color into several subcolors.

In a first step, keypoints and descriptors are calculated in the original input image. To compute the color pyramids feature, the input image is converted into the HSV color space, and all channels but hue are discarded. L evenly spaced values c_k from the hue channel ($k = 1 \dots L$) are selected and L color masks M_k are calculated that contain a range r of colors around each value. The color mask M_k is defined as

$$M_k(x,y) = \begin{cases} 255, & I(x,y) \in [c_k - r, c_k + r), \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where I denotes the source image and (x,y) is the pixel coordinate. Optionally, color masks are smoothed

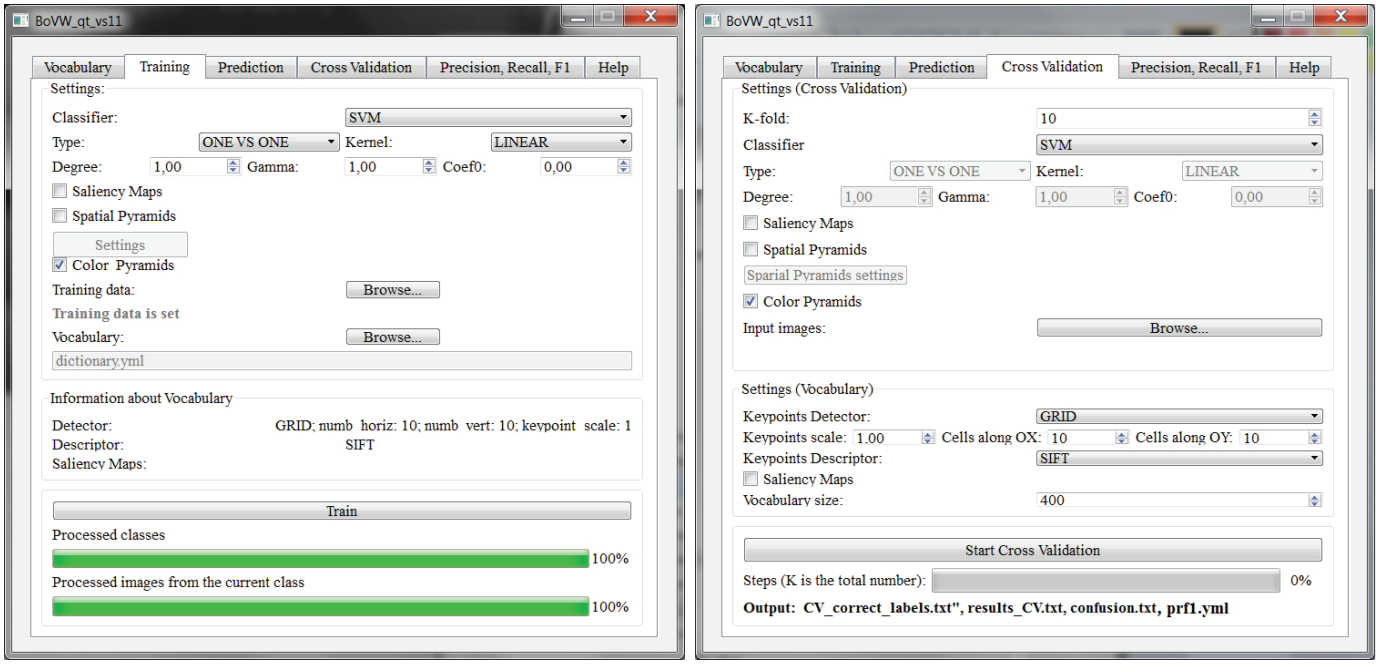

Fig. 3. GUI of the iconic image classification system. Training the classifier (left) and validation (right).

to reduce noise. Figure 2 shows an example of two computed color masks. White means that a pixel's color is within the color range of the mask, and black means that it is not.

Next, a complete histogram of visual word vectors is computed by using all keypoints along with their descriptors. For each created color mask $M_k, k \in 1, ..., L$, all keypoints that lie on black pixels in the mask are filtered out. By using only the remaining keypoints and their descriptors, a histogram of visual word vectors v_k is computed that is specific to the considered color mask M_k. All vectors v_k are then concatenated into one large visual word histogram vector $(v_1, v_2, ..., v_L)$ that represents an image and also captures its color information. This vector may now contain duplicates of visual words due to the partially overlapping color masks. The vector is then used as feature in the BoVW method.

C. Implementation

We use the OpenCV library for C++ and the QT framework for the implementation of our system. The SVM implementation is based on the LibSVM library. Our system supports a simple graphical user interface where the functionality is divided into six categories (see the tabs in Figure 3). The source code[1] is available under the GNU public license.

[1]The source code of the system is available at: http://ls.wim.uni-mannheim.de/de/pi4/research/projects/iconicimages/

IV. Dataset

Each global topic of iconic images includes several more narrow sub-categories. For example, the categories mushroom, reef, or summer forest are included in the larger topic of biodiversity. By classifying images from such sub-categories, it is possible to distinguish iconic images from the global topics.

The iconic image dataset was generated based on the pipeline described in [3]. The seed images along with their keywords for each topic were chosen based on the results of Google image search, restricted to the National Geographic and Wikipedia encyclopedias. Afterwards, based on the requests with the collected keywords which represents names for the used categories, dataset images were gathered from Flickr. Topics that have less than 100 pictures or with licenses other than Creative Commons were not selected. The remaining images were filtered manually based on their correspondence to their topic.

The created dataset consists of fifteen categories with 100 images in each. Each of the categories belongs to one of the five global topics. Figure 4 shows one example image for each of the categories.

V. Experimental Results

This section presents the classification results of our proposed technique that uses color pyramids and compares it to the basic BoVW approach. The color pyramids technique is orthogonal to other existing methods that improve the BoVW approach and can be combined with

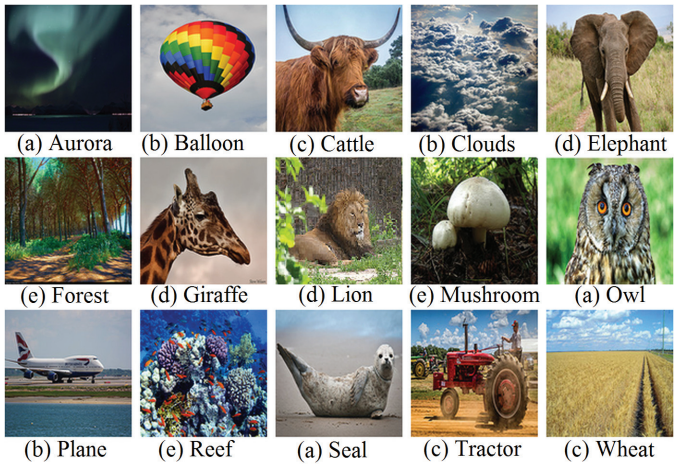

Fig. 4. Image examples from the used dataset. The name of the class label is specified for each category. The global topics are: (a) North Nature, (b) Air, (c) Agriculture, (d) Africa Nature, (e) Biodiversity.

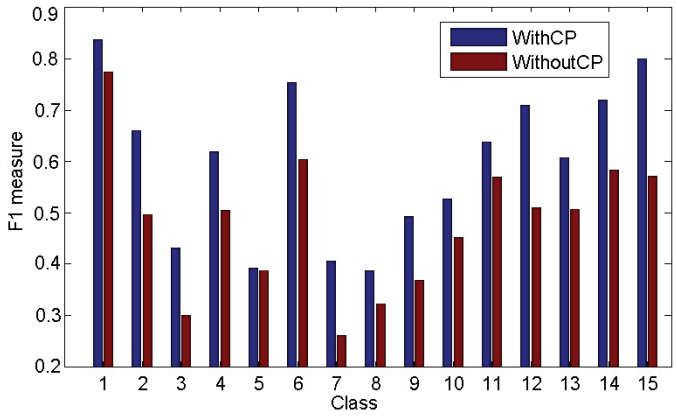

Fig. 5. Achieved F1 measure for each class with and without using color pyramids.

them. For example, it can be used together with spatial pyramids or features weighted by saliency maps.

We made detailed preliminary tests to identify suitable and robust parameters for the BoVW approach. For the comparison presented here, the following parameters are used: Keypoints are selected with the GRID method (10×10 regions) in combination with the SIFT descriptor. To achieve multi-class classification, an SVM with a linear kernel and a "one against one" approach was used. One SVM is trained for each pair of classes, and a label for an entity is assigned in a maximum voting process. To implement the color pyramids method, ten different color masks were computed with the hue values equally distributed between 0 and 180. The range was set to 10 to create slightly overlapping color masks. Each mask was then smoothed with a Dilation filter of size 9. The accuracy is computed with tenfold cross validation. If a category that is assigned to an image is incorrect, even

Fig. 6. Examples of false negative classification without using color pyramids. Under each image, its actual class label is written along with the assigned wrong class label in parentheses.

though the larger topic is correct, it is still considered as an error.

Figure 5 shows the F1 measure for each category. Using color pyramids as features leads to better results. The biggest increase in F1 measure is 0.144 for the category *giraffe* (class 7). There is no significant benefit (0.0039) in the case of *elephants* (class 5) which are either gray or dark brown. On average, the use of color pyramids improves the F1 measure by 0.118.

When comparing the confusion matrices (see Figure 7), a decrease of type 1 and 2 errors can be seen when using color pyramids. For instance, the category aurora is partially misclassified as clouds in the original approach. This error drops significantly when colors are considered, because the sky in the aurora images usually contains green colors that are unlike the blue sky in the cloud pictures. *Wheat* pictures where the dominant color is beige, are no longer misclassified as *cattle* (green grass), *forest* (green color), *mushroom* (beige, green, yellow, red colors) or *owl* (brown, white, green colors) as often. Figure 6 shows examples of false classifications of the original BoVW approach.

To compare the run time of the standard and the advanced BoVW methods, the dataset was limited to 1000 pictures in total. The same test settings are used as before. We used a laptop[2] for this evaluation that is comparable to standard workplace computers. All time measurements were carried out five times on each machine. Comparing the run time of the different descriptors when training the vocabulary, the SIFT descriptor requires 28 minutes whereas SURF requires 6 minutes only. This is an expected result since SURF descriptors are designed to be computationally efficient. When changing the vocabulary size, the computation time increases with larger vocabulary sizes and varies between 33 minutes (size of 100) to 51 minutes (size of 500). The computation time of the color pyramids depends on the number of color masks used. The run

[2]Intel Core i7-2670QM (2.20 GHz), 4GB RAM, Windows 7.

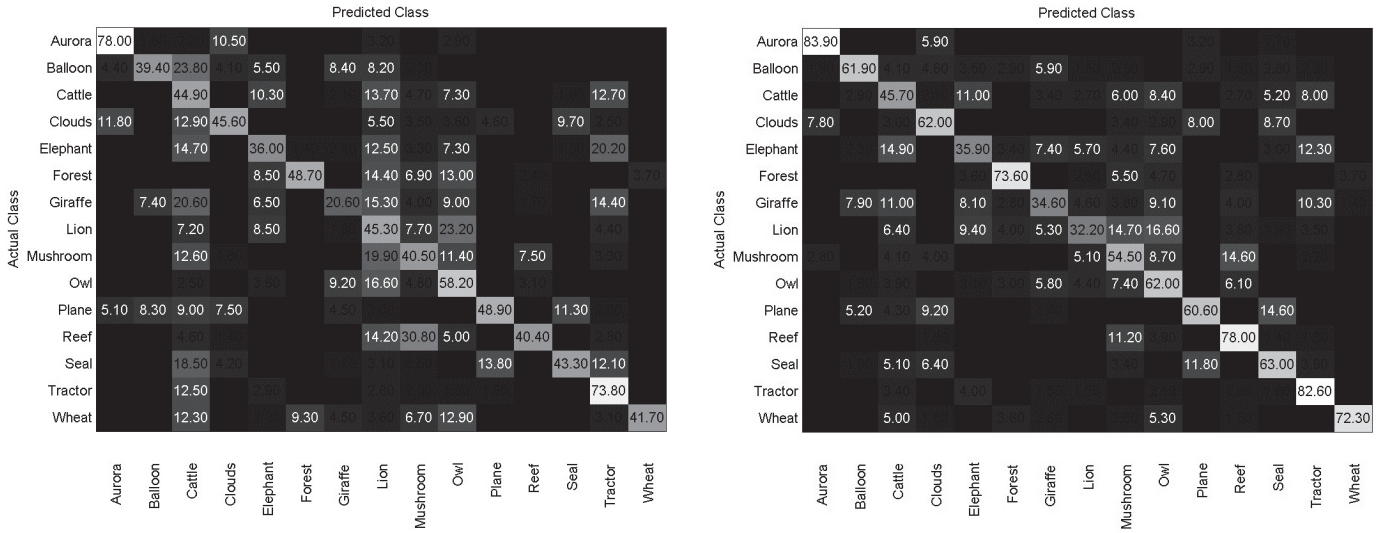

Fig. 7. Confusion matrices comparing the standard BoVW method with the SIFT descriptor (left) and the advanced method using color pyramids (right).

time increases up to a factor of 6, but drops significantly if less than 10 masks are used or if the overlap between the color masks is small.

VI. Conclusion

We presented a system for the classification of iconic images, which is a highly challenging task due to the rich semantics included in such images. As a novel feature, we proposed color pyramids that enhance the standard BoVW method with color information. This makes it possible to distinguish between similar textures like grass or wheat by considering their colors. Using this feature increases the average F1 measure over all categories of iconic images by 0.117. The source code of the system is available for download.

The decision whether an image is iconic or not is still mainly made by a human observer. Familiarity with the global topic and the image context play an important role here. As future work, we would like to develop algorithms that generally answer the question about iconicity in multimedia documents. To achieve this goal, a combined analysis of text and image search will be required.

VII. Acknowledgments

The sample images shown in this paper have been provided by Flickr users under Creative Commons licenses. Thanks to Billy Idle (Aurora), Len Radin (Balloon), Martin Liebermann (Cattle), Farrukh (Clouds), Brittany H. (Elephant), Moyan Brenn (Forest), Steve Wilson (Giraffe), Finn Frode (Lion), Ryan Steele (Mushroom), Edgar Barany C (Owl), Bill Damon (Plane), thinkpanama (Reef), Northwest Power and Conservation Council (Seal), Mobilus In Mobili (Tractor), and Rae Allen (Wheat) for providing the pictures.

References

[1] T. L. Berg and A. C. Berg, "Finding iconic images," in *IEEE CVPR Workshops*, June 2009, pp. 1–8.

[2] R. Raguram and S. Lazebnik, "Computing iconic summaries of general visual concepts," in *IEEE CVPR Workshops*, June 2008, pp. 1–8.

[3] S. P. Ponzetto, H. Wessler, L. Weiland, S. Kopf, W. Effelsberg, and H. Stuckenschmidt, "Automatic classification of iconic images based on a multimodal model," in *Bridging the Gap between Here and There - Combining Multimodal Analysis from International Perspectives, Interdisciplinary Conference on*, 2014.

[4] S. Kopf, T. Haenselmann, and W. Effelsberg, "Shape-based posture and gesture recognition in videos," in *Proc. SPIE 5682, Storage and Retrieval Methods and Applications for Multimedia*, 2005, pp. 114–124.

[5] S. Wilk, S. Kopf, and W. Effelsberg, "Robust tracking for interactive social video," in *IEEE Applications of Computer Vision (WACV)*, 2012, pp. 105–110.

[6] S. Richter, G. Kuehne, and O. Schuster, "Contour-based classification of video objects," in *Proc. SPIE 4315, Storage and Retrieval for Media Databases*, 2001, pp. 608–618.

[7] U. Altintakan and A. Yazici, "Towards effective image classification using class-specific codebooks and distinctive local features," *Multimedia, IEEE Transactions on*, vol. 17, no. 3, pp. 323–332, March 2015.

[8] S. Suchitra and S. Chitrakala, "A survey on scalable image indexing and searching," in *Computing, Communications and Networking Technologies (ICCCNT), Fourth International Conference on*, 2013, pp. 1–5.

[9] J. Mukherjee, J. Mukhopadhyay, and P. Mitra, "A survey on image retrieval performance of different bag of visual words indexing techniques," in *Students' Technology Symposium (TechSym), IEEE*, 2014, pp. 99–104.

[10] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, "Generating descriptive visual words and visual phrases for large-scale image applications," *Image Processing, IEEE Trans. on*, vol. 20, no. 9, pp. 2664–2677, 2011.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE CVPR*, vol. 2, 2006, pp. 2169–2178.

[12] G. Sharma, "Discriminative spatial saliency for image classification," in *IEEE CVPR*. IEEE, 2012, pp. 3506–3513.