Recognizing Scene Categories of Historical Postcards

Rene Grzeszick, Gernot A. Fink
{*rene.grzeszick, gernot.fink*}*@tu-dortmund.de*

Department of Computer Science, TU Dortmund

Abstract. The recognition of visual scene categories is a challenging issue in computer vision. It has many applications like organizing and tagging private or public photo collections. While most approaches are focused on web image collections, some of the largest unorganized image collections are historical images from archives and museums. In this paper the problem of recognizing categories in historical images is considered. More specifically, a new dataset is presented that addresses the analysis of a challenging collection of postcards from the period of World War I delivered by the German military postal service. The categorization of these postcards is of greater interest for historians in order to gain insights about the society during these years. For computer vision research the postcards pose various new challenges such as high degradations, varying visual domains like sketches, photographs or colorization and incorrect orientations due to an image in the image problem. The incorrect orientation is addressed by a pre-processing step that classifies the images into portrait or landscapes. In order to cope with the different visual domains an ensemble that incorporates global feature representations and features that are derived from detection results is used. The experiments on a development set and a large unexplored test set show that the proposed methods allow for improving the recognition on the historical postcards compared to a Bag-of-Features based scene categorization.

1 Introduction

Archives, museums and libraries have access to tremendous amounts of historical documents and images, which provide valuable insight into the past not only for historians but also for the greater public. Most of these samples are only interesting in the context of a whole data collection, referred to as a mass source. Despite the availability of such mass sources, they are very difficult to interpret manually. The amount of data makes their digitization difficult, but a crucial part of the work is assigning machine readable labels to the data samples.

In this paper the automatic recognition of categories within a challenging collection of postcards from the period of World War I is considered. The postcards were delivered by the German military postal service, so called *feldpost* [2]. More specifically, thematic categories on the images of the postcards, e.g.

Fig. 1. Example of German feldpost postcards from World War I. From top left to bottom right showing the following categories: *cartoon, destruction, frontline, landscape, love & poem, patriotism, portrait* and *weapons*. The samples show degradations, destructions of the images, different color schemes and different orientations. The images are taken from [1].

published in: German Conference on Pattern Recognition (GCPR) 2014

landscapes, portraits, love or *patriotic* themed images, are recognized. The images in some of these categories provide insights about the society by expressing political opinions or representing the everyday life or typical role models and the changes of these over time (cf. [2]). Examples of the postcards and their categories are shown in Fig. 1. For a detailed analysis an automatic recognition of certain categories is desirable. Here, scene recognition methods are applied that use global image representations like Bag-of-Features and spatial tiling, GIST or color histograms in order to classify natural scene images [19]. In contrast to scene recognition problems on modern benchmarks that are often based on web images some additional challenges arise on the historical postcard images.

The first challenge is their visual appearance. The postcards are often degraded or the images are faded. In addition, some of the postcards are black and white photographs, others are painted color images or colorized images and some of them are only drawn sketches. These color schemes create different visual domains and thus increased intra class variations. An ambitious approach for dealing with these visual domains is cross-domain image matching, as described in [14]. Features are re-weighted based on their uniqueness. In order to learn the uniqueness of features, the features of a query are compared with features from thousands of images. Thus, the unique features in a visual domain are found and used for matching a query with a labeled dataset. However, the application is still quite impractical since the uniqueness of features cannot be efficiently computed. Also problems might arise from the fact that common images are typically obtained from web databases which may not reflect the domain of the historical images appropriately. A more common approach toward these variations is the adaption between different domains using SVM margin based approaches for adapting from a training set to a test set in a different domain, as described in [6]. However, in contrast to [6] the visual domains are not known beforehand. This makes adaption between them impractical without further manual effort. Therefore, multiple visual cues are used in order to capture information from different domains. Global image feature representations and features that are derived from detections are combined in an ensemble. The idea is to infer information about a category based on entities that occur in the image. Useful detections could be buildings, persons or even text. In this paper the focus is on faces which are the most promising entity for distinguishing the categories and can be computed in an unsupervised manner. A face detector can be pre-trained on existing data and then be adapted to the historical image domain. The second and less obvious challenge is the orientation of the images. Natural scene images are typically in the proper orientation, which allows for context constraints that are widely used in scene recognition. For example, the tiling in the Spatial Pyramid approach implicitly learns that the sky is on the top and people are rarely upside down (cf. [9, 19]). In contrast, the images of historical postcards can best be described as an image in an image whose orientation is not necessarily known beforehand.

published in: German Conference on Pattern Recognition (GCPR) 2014

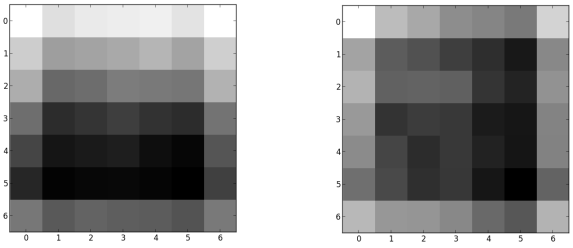

Fig. 2. Example of the average 7×7 tiny image for landscape (left) and portrait (right) images. The landscape images show a strong horizontal structure (e.g. buildings), whereas the portrait images show a similar but vertical structure.

2 Method

For identifying categories in the historic postcards a two stage approach is proposed. Before classifying the categories the postcards are automatically put in the proper orientation by a pre-processing step that is described in section 2.1. Then, an ensemble approach is used that addresses the high intra class variations by using different representations including global image feature representations and a face detection based feature, which are discussed in section 2.2 and 2.3.

2.1 Postcard orientation

Most feature representations, e.g. Spatial Pyramids [9], GIST [11], Tiny Images [16], build upon the assumption that the image is correctly orientated. For example, the Spatial Pyramid subdivides the image in tiles and SIFT descriptors for a Bag-of-Features representation are computed on a grid with a fixed orientation. Thus, when rotating an image, it will have a completely different representation in the feature space. Typically, the assumption of a correct spatial configuration holds. Web images, which are often used for creating image datasets, are mostly provided in the proper orientation and photographs from private collections are typically rotated by the user or simply by using the EXIF data from the camera.

However, for images of postcards this assumption does not hold. While the text side of a postcard is always landscape the front side image can either be a landscape or a portrait image. Therefore, we propose a pre-processing step in which the orientation of the postcards is determined automatically. Even if putting the images in the proper orientation does not improve the recognition rates, it is desirable for archives to do so without further manual effort. A 7×7 pixel gray scale thumbnail is computed that covers the typical structures appearing in landscape or portrait images. The average images for both categories are shown in Fig. 2. In addition, a gradient histogram with 3×3 tiles and 8 orientations, as well as the mean and standard deviation of the gradients along the X and Y-axis are computed. This results in a 125 dimensional feature representation. An SVM is trained in order to classify images as portrait or landscape oriented. For further processing, the postcards in an incorrect format are rotated.

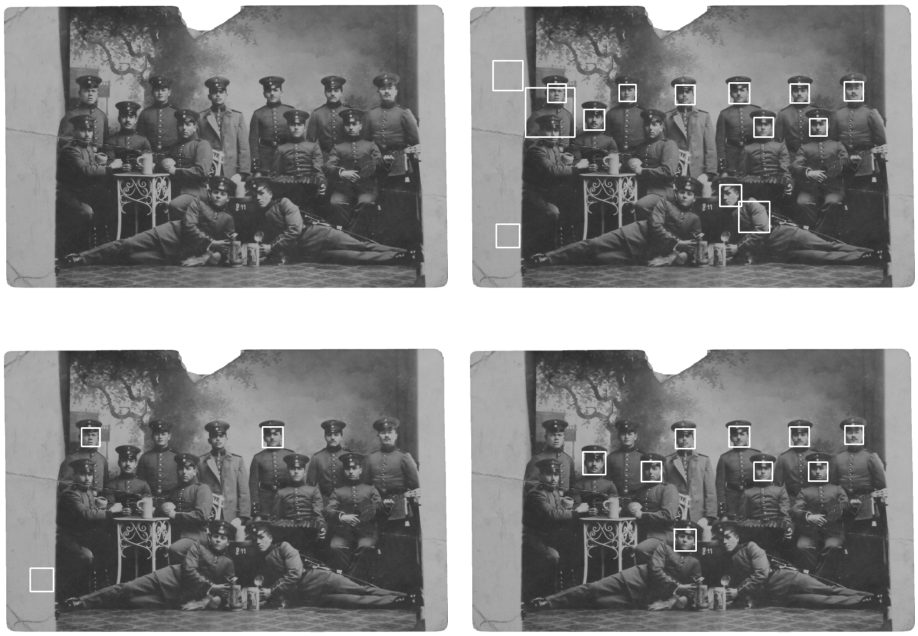

Fig. 3. Qualitative example showing the difficulty to recognize faces in the postcard images. Top left: an example of a group portrait with degradations and destructions of the image. Top right: detection results using a haar cascade [18]. A few false positives are shown due to the sensitivity to noise. Lower left: detection after applying a global threshold on the detections. Due to the large amount of clutter in the whole collection most positive detections are removed. Lower right: using a face classification trained on the best detections instead of thresholding. Most faces are recognized and the noise is removed.

2.2 Face detection

It is known that scene level information can be leveraged for improving detection tasks in images (cf. [3]). Here, this relation is reversed by deriving features from detection results. In general such relations could be learned from training data (cf. [20]). In case of the historical postcards considered in this paper, basic knowledge about the problem domain can be applied. Useful detections could be buildings, persons, their faces or texts like poems (cf. Fig. 1). From these detections, faces appear to be the most promising ones. On postcards, historical or today's, the appearance of persons can be a very meaningful feature for several categories. For example, some images show countrysides or buildings while others show close-up or group portraits.

The most common method for face detection uses haar cascades as first introduced in [18]. The detections are computed in different scales. Detections at

the same position at different scales are then returned as one result and assigned a weight w_i based on the number of detections at this position. Typically, a threshold ϵ is used so that only results with a weight $w_i > \epsilon$ are kept and the others removed in order to reduce false positive detections. However, there are additional challenges that arise on historical images. On modern images the detection is often enhanced by skin color estimation in order to reduce false positives (cf. [8]) which is not possible on gray scale images. Also, the postcards are degraded and show noises toward which the detector is very sensitive. Especially geometric normalizations that mostly build on top of eye detection [15] are almost impossible. Thus, the detection results contain a lot of clutter and simple thresholding does not yield satisfactory results.

Therefore, the face detector is adapted to the domain of historical postcard images. Note that no ground truth annotations are available that could be used for training a detector on historical images. First, faces are detected in the training set without any thresholding and the detections are ordered by their weights w_i. A gradient orientation feature with three scales and $8, 8$ and 4 orientations is computed on each of the detected regions. An SVM is trained on the k best detections and the $3k$ worst detections. This allows for learning typical faces and typical background clutter. It also covers the characteristics of the persons in that time, e.g. facial hair or hats that can only rarely be found in modern images, which is why it is not possible to train a classifier on large available datasets like FERET [12] or LFW [7]. Figure 3 illustrates the issues and shows a qualitative result for the proposed detection approach. The detected faces are then used for deriving features that help identifying the image category. A quantitative evaluation of the proposed face detection is shown in section 3.3.

2.3 Classification ensemble

In order to address the variations in the data an ensemble of different features is computed. For each feature representation j a random forest classifier is trained. Then, an ensemble decision is made by combining the probabilities:

$$\Omega_{max} = \underset{i}{argmax} \prod_j P(\Omega_i | c_j) \tag{1}$$

Random Forests are used in the ensemble since they outperform SVMs on this task and are able to represent classes with only a few training samples well. Also the computation of class-wise probabilities is not obvious for SVMs. Typical scene recognition representations are Bag-of-Features, LBP histograms and GIST [19]. Color histograms are also widely used, but not as powerful on diverse image datasets. Here, the color histogram may account for the bias of various categories toward different visual domains, e.g. *portraits* are mostly black and white photograph, while *cartoons* are mostly drawn color images (cf. [13]). In addition faces are detected, as described in 2.2, and features are derived from the detections. In the following the features are discussed in detail:

Bag-of-Features: SIFT features are extracted on a dense grid with a step size of 5 pixels and bin sizes of $4, 6, 8$ and 10 pixels. In order to incorporate spatial

information the concept of Spatial Visual Words is applied [5]. Here, a spatial tiling of 2×2 areas is considered. The descriptors are quantized into a codebook of $1,000$ Spatial Visual Words resulting in a Bag-of-Features representation. The histogram is then represented by square rooted frequencies [17].
LBP: A histogram of rotation invariant Local Binary Patterns is computed [10]. At each pixel 12 comparison points are chosen on a circle with a radius of one. The interpolated pixel values are compared with the center pixel in order to derive the rotation invariant LBP code. The compressed rotation invariant code has 352 dimensions. A pyramid scheme is built that computes an LBP histogram for 3×3 subregions and a histogram of the complete image is derived using max pooling.
GIST: For a global image description the GIST of a scene is computed. Here, the Spatial Envelope representation is computed as introduced in [11] using the color GIST implementation described in [4]. Three scales with 12, 12 and 4 orientations are computed, which gives a 1344 dimensional feature representation.
Color Histogram: For each color channel of an RGB image 16 equally sized bins are computed. Additionally for each color channel a mean and standard deviation is computed, resulting in a 54 dimensional color feature.
Faces: Faces are detected as described in section 2.2. Here, the $k = 30$ best detections, which can safely be assumed as a lower bound for the number of faces that can be found in the dataset, are used for learning typical faces. A seven dimensional feature is computed that consists of the number of faces and the mean and standard deviation of the position (x, y) and size of the faces.

3 Evaluation

The proposed method has been evaluated on two sets of historical postcards that are described in more detail in section 3.1. The experiments concerning the orientation and categorization are described in sections 3.2 and 3.3.

3.1 Dataset of Historical Postcards

The dataset of German feldpost postcards considered in this work is part of a private collection of postcards from World War I [1]. In total 1,346 postcards from nine different acquisition campaigns have already been digitized [1]. All postcards were photographed at approximately 600 dpi with a total resolution of 4288×2484 pixels. The photographs were taken in front of a red background that is removed by color space thresholding. In the following, images from the postcards that are approximately 800×525 pixels in size are considered. 256 of these postcards have been annotated. From the typical categories described in [2] the following eight could clearly be identified: *cartoon, destruction, frontline, landscape, love & poem, patriotism, portrait* and *weapons*. Images that could not

[1] The dataset will be made available for research purposes in the resources section at http://patrec.cs.tu-dortmund.de

Features	Dim.	Correctly oriented images	Landscape	Portrait	Class average
Nothing	-	65.8%	100%	0%	50%
Tiny Image	125	$83.5 \pm 1.0\%$	$81.1 \pm 6.8\%$	$88.2 \pm 7.1\%$	$84.6 \pm 4.1\%$
GIST	960	$71.2 \pm 5.4\%$	$69.2 \pm 13.3\%$	$75.2 \pm 10.7\%$	$72.2 \pm 2.0\%$
Bag-of-Features	100	$81.2 \pm 2.7\%$	$79.6 \pm 6.5\%$	$84.5 \pm 8.3\%$	$82.0 \pm 2.5\%$
Bag-of-Features	1,000	$84.8 \pm 2.8\%$	$82.9 \pm 5.6\%$	$88.5 \pm 3.7\%$	$85.7 \pm 1.6\%$

Table 1. Number of correctly oriented postcards after the orientation classification. Note that 65.8% of the images in the validation set are landscapes and thus already correctly oriented without any processing.

be clearly associated with one of these categories are assigned to a background class. In addition to these annotations a few sub categories (e.g. portrait and group portraits) and the correct orientation of the images have been annotated. This small 256 postcard dataset is used for validating the proposed method. The remaining images are a large, so far unexplored, collection from three different acquisitions that can be used as a realistic test set. For these images the recognition quality can only be estimated as no ground truth is available.

For all of the following experiments the 256 postcard dataset was randomly split into a training and a validation set. A major issue is that the dataset is highly unbalanced. Therefore, for each category 50% of the images, but no more than 30, were used for training and the remaining ones for validation. All experiments were repeated five times using different training samples.

3.2 Orientation

For the orientation classification the 256 postcard dataset was split into a training and validation set of portraits and landscapes. The results are shown in Table 1. The proposed approach is compared to global image representations: a GIST and a Bag-of-Features representation using SIFT descriptors. Note that without any classification already 65.8% of the images are correctly oriented since all postcards were digitized as landscapes. The GIST representation yields a classification rate of 71.2%. The proposed approach can be computed more efficiently and also allows for correctly estimating the orientation of 83.5% of the images. The Bag-of-Features representation achieves competitive recognition rates in comparison with the proposed approach but at much higher computational costs due to the clustering and quantization of the local image features.

In the following the effect of rotating the images correctly with respect to the classification performance has been evaluated. A categorization experiment was performed using a single feature type. The postcards were used without changing the orientation and after changing it by using the proposed orientation classification approach. In addition, an oracle experiment has been performed on the ground truth orientation as an upper baseline. Table 2 shows the results for all feature representations described in section 2.3. Please note that the color histogram is rotation invariant and that the face features do not work

Features	Unoriented	Oriented*	Oriented (GT)
GIST	$43 \pm 2.2\%$	$46.3 \pm 1.3\%$	$46.5 \pm 1.1\%$
Bag-of-Features	$59.1 \pm 0.9\%$	$59.2 \pm 1.5\%$	$60.5 \pm 1.1\%$
LBP	$53.6 \pm 2.9\%$	$54.2 \pm 2.0\%$	$56.6 \pm 1.6\%$
Color Histogram	$45.9 \pm 3.1\%$	-	-
Faces	-	$51.3 \pm 1.2\%$	$53.2 \pm 1.5\%$

Table 2. Classification results of a random forest on the 256 postcard dataset. Left: the postcards were used as photographed. Middle: the orientation was corrected using the classification as proposed in section 2.1 (*). Right: the images are correctly oriented using the ground truth annotations (GT).

properly on incorrectly oriented images. Using the ground truth orientations improves the classification rate significantly in all three cases which clearly states the usefulness of having the images in the proper orientation. Correcting the orientation using the proposed method also yields an improvement on the classification rate. The Bag-of-Features representation shows the best recognition results and appears to be relatively robust against changes in the orientation. Most likely because it covers finer structures that are typical for the categories instead of global orientation information.

3.3 Ensemble recognition

In the following, the ensemble approach is evaluated using the features described in section 2.3. An extensive study of all combinations has been performed. Some of the results are shown in Table 3. The best classification rate of $62.5 \pm 1.1\%$ is achieved with a combination of the Bag-of-Features, GIST and the face detection based features. Since the dataset is highly unbalanced, the categories *landscape* and *portrait* are highly overrepresented while some of the other categories do not have more than a few samples. Therefore, the overall recognition rate as well as the average class-wise recognition rate over all nine classes is shown. While the classification rate is $62.5 \pm 1.1\%$ the respective class average is only around $22 \pm 0.7\%$ due to the limited amount of samples. Note that the ensemble improves the overall classification rate but on the other hand tends to slightly reduce the class average. The results also show that it is not beneficial to use all possible features. This is more clearly shown by the ensemble combination using the color histogram which reduces the classification rate. Although it might help to identify the visual domain, e.g., distinguish photographs from color sketches, this information does not appear to pose valuable information for the categorization. The face features on the other hand add some very specific information about the postcards that helps recognizing them correctly. It especially helps to recognize the categories *portrait* and *love & poem* very well.

The face detection method that is adapted to the historical images yields a precision of 94.7% and an estimated recall of approximately 31%. Due to fading and clutter in some of the images the exact number of faces is not known. The results in Table 4 also show that the proposed method outperforms a cascade detection without any adaption.

Dataset	Method	Features	Classification rate	Class avg.
256 Postcards	Bag-of-Features	SIFT	$59.2 \pm 1.5\%$	$23.0 \pm 1.3\%$
	Ensemble	SIFT GIST Faces	$62.5 \pm 1.1\%$	$22.0 \pm 0.7\%$
	Ensemble	SIFT GIST Color	$50.1 \pm 2.7\%$	$17.8 \pm 0.9\%$
	Ensemble	All	$60.4 \pm 1.6\%$	$21.8 \pm 0.8\%$
Unexplored set	Ensemble	SIFT GIST Faces	47.4%	25.2%

Table 3. Classification results on both sets of historical postcards. In all cases a random forest has been used for classification. For the unexplored set the recognition rates could only be estimated as no ground truth annotations are available.

Method	Precision	Recall	F_1 score
Cascade	89.2%	22.4%	35.8%
Cascade & adaption	94.7%	31.0%	46.7%

Table 4. Face detection on the 256 postcard dataset (GT oriented). The detector that is adapted to the historical postcards is compared to a baseline detection cascade using a fixed threshold $\epsilon = 6$ and a scaling factor of 1.3, which showed the best results.

3.4 Unexplored set

The best performing method has then been evaluated on the so far unexplored set. While this set poses a realistic test scenario, no annotations are available for this set. Therefore, the results are estimated by manually verifying the results of 50 samples in each category. The results are shown in Table 3. Here, a recognition rate of 47.4% is achieved. This estimation is rather a lower bound as a majority of the images belong to the categories *landscape* and *portrait* which are recognized with 80% and 70% respectively. The estimated class average of 25.2% is comparable to the 256 postcard dataset and slightly improved due to the larger number of training samples for the rare categories. On the negative side no *weapon, patriotism* or *frontline* themed images were recognized as these categories are very rare and can easily be mistaken.

4 Conclusion

In this paper a new task for scene categorization has been introduced: a challenging set of historical postcard images from the period of World War I. First approaches to address the challenges that arise on this image collection have been proposed.

The orientation of postcards as an image in the image has been evaluated and an ensemble approach has been applied for categorization. Also features that are based on detection results, in this case based on face detections, have been incorporated in the ensemble. It could be shown that all three steps show improvements for recognizing thematic categories in the historical postcards. Recognition rates of more than 62.5% have been shown on a validation set and 47.4% have been estimated on a large unexplored test set.

published in: German Conference on Pattern Recognition (GCPR) 2014

References

1. Bley, B.: Feldpostkarten im 1. Weltkrieg (Feldpost Postcards of World War I). Private Collection
2. Brocks, C.: Die bunte Welt des Krieges: Bildpostkarten aus dem Ersten Weltkrieg 1914-1918 (The Colorful World of the War: Picture Postcards from the First World War 1914-1918). Klartext-Verlag, Essen (2008), (in German)
3. Divvala, S., Hoiem, D., Hays, J., Efros, A., Hebert, M.: An empirical study of context in object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1271–1278 (2009)
4. Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., Schmid, C.: Evaluation of gist descriptors for web-scale image search. In: Proceedings of the ACM International Conference on Image and Video Retrieval. ACM (2009)
5. Grzeszick, R., Rothacker, L., Fink, G.A.: Bag-of-features representations using spatial visual vocabularies for object classification. In: Proc. IEEE Intl. Conf. on Image Processing (2013)
6. Hoffman, J., Rodner, E., Donahue, J., Darrell, T., Saenko, K.: Efficient learning of domain-invariant image representations. In: Proc. Intl. Conf. on Learning Representations (ICLS) (2013)
7. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
8. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin-color modeling and detection methods. Pattern recognition 40(3), 1106–1122 (2007)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2169–2178 (2006)
10. Ojala, T., Pietikäinen, M., Mäenpää, T.: Gray scale and rotation invariant texture classification with local binary patterns. In: Computer Vision-ECCV 2000, pp. 404–420. Springer (2000)
11. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in brain research 155 (2006)
12. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22(10), 1090–1104 (2000)
13. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(4), 754–766 (2011)
14. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. In: ACM Transactions on Graphics (TOG). vol. 30, p. 154. ACM (2011)
15. Talele, K., Kadam, S.: Face detection and geometric face normalization. In: TENCON 2009-2009 IEEE Region 10 Conference. pp. 1–6. IEEE (2009)
16. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(11), 1958–1970 (2008)
17. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: IEEE 12th International Conference on Computer Vision. pp. 606–613. IEEE (2009)

published in: German Conference on Pattern Recognition (GCPR) 2014

18. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 511–518. IEEE (2001)
19. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on. pp. 3485–3492. IEEE (2010)
20. Zhu, S.s., Yung, N.H.: Improve scene categorization via sub-scene recognition. Machine Vision and Applications pp. 1–12 (2014)

published in: German Conference on Pattern Recognition (GCPR) 2014