Bag–of–Colors for Biomedical Document Image Classification

Alba G. Seco de Herrera, Dimitrios Markonis, and Henning Müller

University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland
{alba.garcia,dimitrios.markonis,henning.mueller}@hevs.ch
http://medgift.hevs.ch

Abstract. The number of biomedical publications has increased noticeably in the last 30 years. Clinicians and medical researchers regularly have unmet information needs but require more time for searching than is usually available to find publications relevant to a clinical situation. The techniques described in this article are used to classify images from the biomedical open access literature into categories, which can potentially reduce the search time. Only the visual information of the images is used to classify images based on a benchmark database of ImageCLEF 2011 created for the task of image classification and image retrieval. We evaluate particularly the importance of color in addition to the frequently used texture and grey level features.
Results show that bags–of–colors in combination with the Scale Invariant Feature Transform (SIFT) provide an image representation allowing to improve the classification quality. Accuracy improved from 69.75% of the best system in ImageCLEF 2011 using visual information, only, to 72.5% of the system described in this paper. The results highlight the importance of color for the classification of biomedical images.

Keywords: bag–of–colors, SIFT, image categorization, ImageCLEF

1 Introduction

The number of biomedical articles published grew at a double–exponential pace between 1986 and 2006 according to [1]. Images represent an important part of the content in many publications and searching for medical images has become common in applications such as Goldminer[1], particularly for radiologists. Image retrieval has shown to be complementary to text retrieval approaches and images can well help to represent the content of scientific articles, particularly in applications using small interfaces such as mobile phones [2].

Many physicians have regular information needs during clinical work, teaching preparation and research activities [3, 4]. Studies showed that the time for answering an information need with MedLine is around 30 minutes [5], while clinicians state to have approximately five minutes available [6]. Finding relevant information quicker is thus an important task to bring search into clinical

[1] http://goldminer.arrs.org/

routine. To facilitate searching for images in biomedical articles, search engines such as Goldminer include the ability to filter search results by modality, age group or gender [7]. Imaging modalities can include typical classes such as x–ray, computed tomography (CT) or magnetic resonance imaging (MRI). In the biomedical literature, other classes such as photos (e.g. photomicrographs and endoscopic images), graphics (e.g. charts and illustrations) and compound figures also occur frequently [7–9]. For the modality classification, caption information can help if captions are well controlled like in the radiology domain but the more general biomedical literature makes it hard to find the modality information in the caption. Past work has shown that the image modality can be extracted from the image itself using visual features, only [10–12]. Therefore, in this paper, purely visual methods are used for the classification.

Our focus is on implementing, evaluating and developing visual features for representing images for the task of modality classification. To classify images many features based on color [13], texture [14], or shape [15] have been used. Although color information is important many approaches use only grey level information such as the Scale Invariant Feature Transform (SIFT) [16]. Additionally, several color image descriptors have been proposed [17]. Recently, a color extension to the SIFT descriptor was presented by van de Sande et al. [18]. Ai et al. [19] also proposed a color independent components based SIFT descriptor (CIC–SIFT) for image classification. Color and geometrical features combined are expected to improve the results for classification.

This paper extends image classification with SIFT features [**?**] by adding color features using bags–of–colors (BoC) based on [20] to represent the images. SIFT showed to be one of the most robust local feature descriptors with respect to geometrical changes [21]. As it contains only grey level information we fused results with BoC to include both color and texture information. The ImageCLEF 2011 database for medical modality classification was used as results for many research groups using state–of–the–art techniques on this database are available as baselines [22]. Both visual and textual approaches are possible and this paper concentrates on purely visual approaches.

The rest of the paper is organised as follows: Section 2 provides a detailed description of the methods and tools used. Section 3 reports on results while Section 4 concludes this work.

2 Methods

This section describes the dataset and the evaluation methodology used in this article. The main techniques and tools used are also detailed.

2.1 Dataset

The database of the medical ImageCLEF 2011 task [2] [22] is used in this document. The entire database consists of over 230,000 images of the biomedical open

[2] `http://imageclef.org/2011/medical/`

access literature. For modality classification 1,000 training and 1,000 test images were made available with class labels and a standard setup. Labels are one of 18 categories including graphs and several radiology modalities (see Figure 1). The sample images presented in Figure 2 demonstrate the visual diversity of the classes of the data set. Images are unevenly distributed across classes, which can affect the training of the classifiers and the resulting performance. In our study, a subset of 100 training images uniformly distributed across the classes was used for the creation of the visual vocabularies.

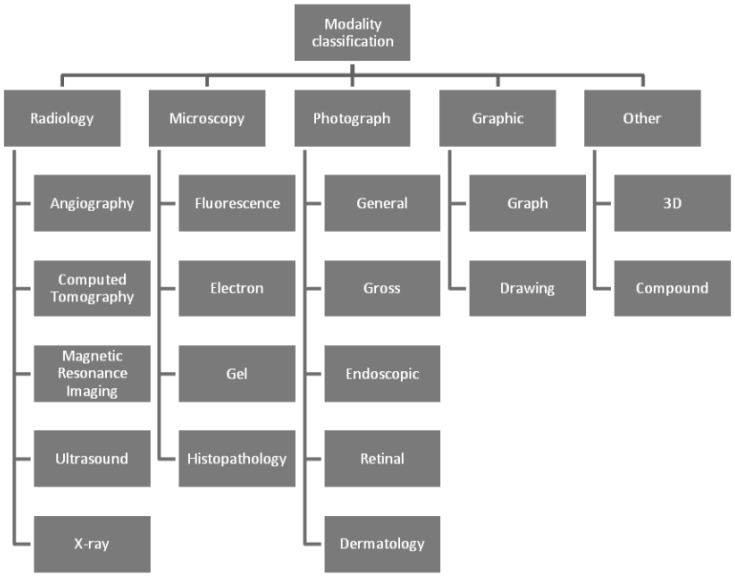
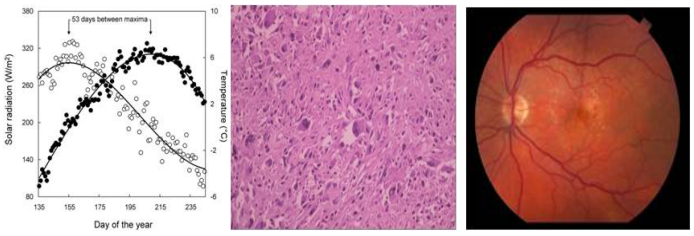


Fig. 1. Modality categories of the ImageCLEF 2011 medical task.

2.2 The CIELab Color Space

We used the CIE (International Commission on Illumination) 1976 L*a*b (CIELab) space for our method because it is a perceptually uniform color space recommended by CIE[3] and used in many applications [23]. CIELab is a 3–D component space defined by L for luminance and a, b for the color–opponent dimensions for chrominance [23, 24].

[3] CIE is the primary organization responsible for standardization of color metrics.



(a) Graph (b) Histopathology (c) Retinography

(d) General Photo (e) Radiology

Fig. 2. Sample images from ImageCLEF 2011 medical data set including their class labels.

2.3 Bags–of–Colors

Bags–of–Colors (BoC) is a method to extract a color signature from images introduced by [20]. The method is based on the Bag–of–Visual–Words (BoVW) image representation [25]. Each image is represented by a BoC from a color vocabulary C previously learned on a subset of the collection.

A color vocabulary $C = \{c_1, \ldots, c_{k_c}\}$, with $c_i = (L_i, a_i, b_i) \in CIELab$ is constructed by first finding the most frequently occurring colors in each image of the subset of the collection. In our case, frequent colors of the 100 selected images are used. A color is considered frequent if it occurs more than once for every 10,000 pixels in an image.The selected colors are clustered using a k–means algorithm [26]. We use for our experiments mainly $k_c = 200$ found by an analysis on the training set (see Table 1). For illustrations (Figures 4 and 5) the visual vocabularies, so the cluster centers of the color clusters for $k_c = 10, 20$ are shown including the histograms for example image types.

The BoC of an image I is defined as a vector $h_{BoC} = \{\bar{c}_1, \ldots, \bar{c}_k\}$ such that, for each pixel $p_k \in I \; \forall k \in \{1, \ldots, n_p\}$, with n_p being the number of pixels of the image I:

$$\bar{c}_i = \sum_{k=1}^{n_p} \sum_{j=1}^{n_p} g_j(p_k) \; \forall i \in \{1, \ldots, k_c\}$$

where

$$g_j(p) = \begin{cases} 1 \; if \; d(p, c_j) \leq d(p, c_l) \; \forall l \in \{1, \ldots, k_c\} \\ 0 \; otherwise \end{cases} \quad (1)$$

and $d(x, y)$ being the Euclidean distance between x and y.

Generally speaking, given a color vocabulary $C = \{c_1, \ldots, c_{k_c}\}$ defined by automatically clustered color occurrences in the CIELab color space, a BoC of an image is obtained by finding for each pixel of the image the closest color in the color vocabulary. The number of times each color appears in the image is then entered into a color histogram. The procedure is the following:

1. Convert images into the CIELab color space.
2. Create a color vocabulary:
   2.1. Find frequently occurring colors in each image from the 100 images selected.
   2.2. Cluster colors using the k–means algorithm.
3. Create a BoC for each image:
   3.1. Select for each pixel of each image, the closest color in the vocabulary using the Euclidean distance.
   3.2. Increment the corresponding bin of the output k_c–dimensional histogram.
4. Normalize to make the vectors comparable.

This procedure is described graphically in Figure 3.

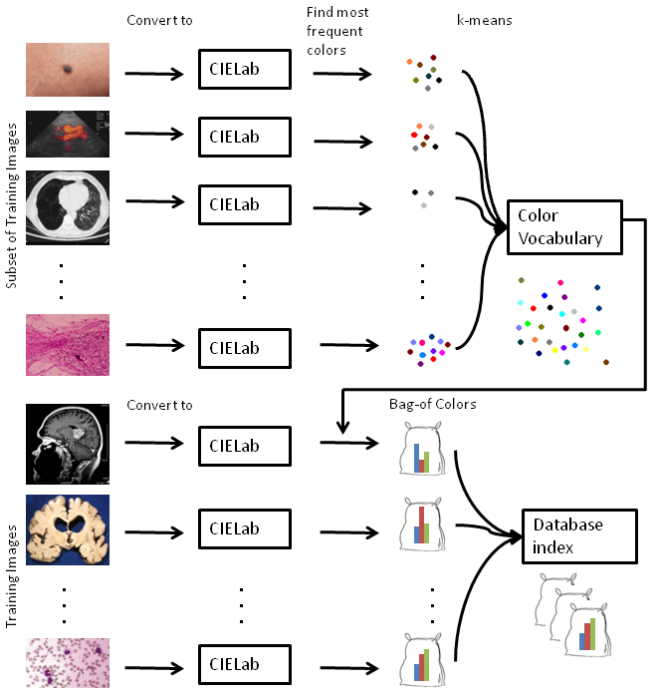


Fig. 3. The procedure for constructing the BoC.

2.4 SIFT

SIFT [16] has been used for general image retrieval and also for medical image classification by several authors in the past [**?**]. The Fiji image processing package[4] was used for the extraction of the SIFT features. In this work, we use SIFT features represented as BoVW. For the creation of the vocabulary, our implementation of the DENCLUE (Density Clustering) [27] algorithm was used to increase the speed of the clustering.

2.5 Representation and Fusion

The images are represented as histograms and the similarity between images is calculated by comparing their histograms. The distance measure used in this article is the histogram intersection [28].

Late fusion [29] is used to combine the results of SIFT and BoC. First we obtain similarity scores separately using SIFT BoVW and BoC descriptors. Then, these scores are fused by voting. The image is classified into one class by a k–NN voting [30]. There are many ways for obtaining the optimal k_{nn} value. We show results with varying k_c and k_{nn} in Tables 1 and 2.

As accuracy we take the percentage of correctly classified images of the entire test set of 1,000 images. This procedure allows for a fair comparison of our different schemes with and without the use of BoC.

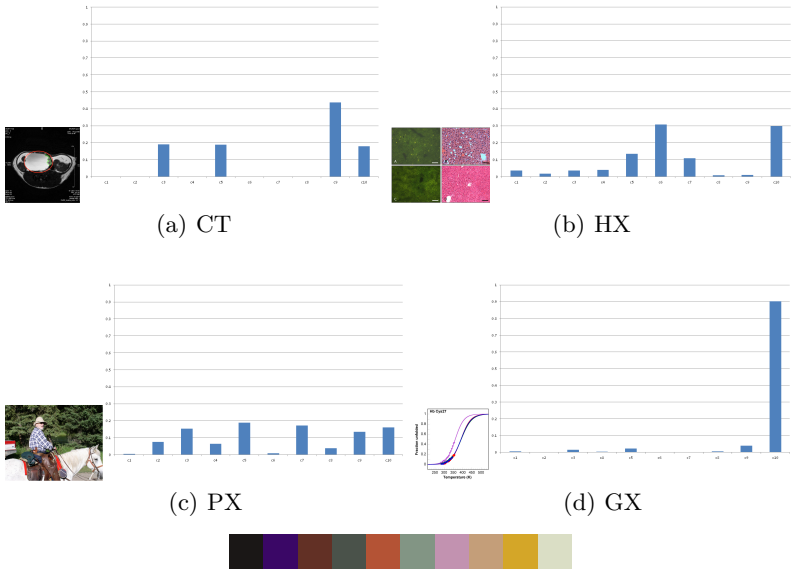
3 Results

To analyze results using the BoC, we present two examples with 10 and 20 color terms (see Figures 4 and 5).

Given $k_c = 10$, a vocabulary $C_{10} = \{c_1, \ldots, c_{10}\}$ is created. The vocabulary contains ten colors corresponding to the ten cluster centers (Figure 4(e)). In Figures 4(a), 4(b), 4(c) and 4(d)) the averages of the BoC corresponding to C_{10} of the classes computed tomography (CT), histopathology (HX), general photos (PX) and graphs (GX) are presented. We can observe that CTs (Figure 4(a)) are not only represented by black and white but also a few other colors. These colors are not represented stronger because several of the CT images in the database used are not fully grayscale images but RGB representations that have some color components. There are also color annotations on the images as they were used in journal texts. The HX BoC (Figure 4(b)) contains mainly red, green, pink and white colors, which is consistent with expectations. The PX BoC (Figure 4(c)) consists of a large variety of colors since it is a class with a very varied content. In the last example, we observe that the GX BoC 4(c) includes mostly white and some black, which is also consistent with expectations.

Given $k_c = 20$, a vocabulary $C_{20} = \{c_1, \ldots, c_{20}\}$ is shown in Figure 5(e) with the same examples. Since there are more colors each modality is represented by a BoC with a larger variety, follow similar patterns as with C_{10}.

[4] `http://fiji.sc/`



(a) CT (b) HX

(c) PX (d) GX

(e) $C_{10} = \{c_1, \ldots, c_{10}\}$.

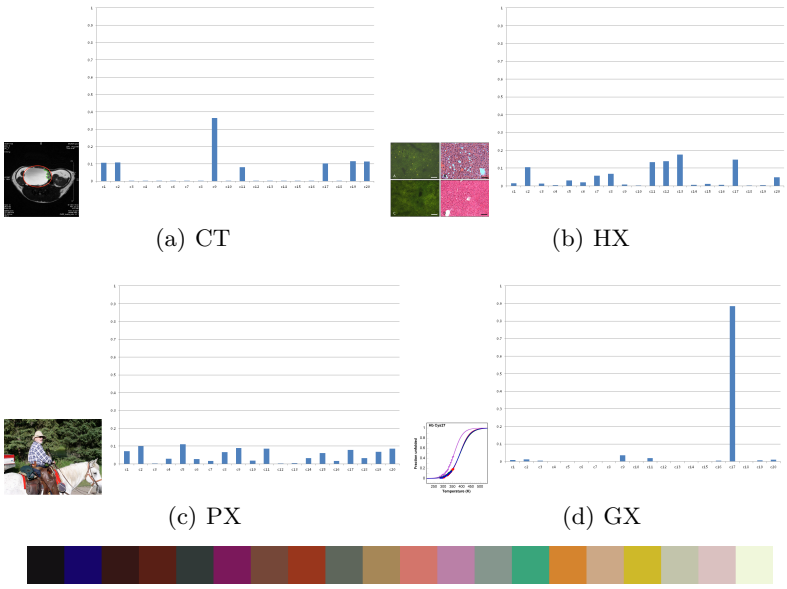
Fig. 4. Average BoC for four modalities corresponding to the color vocabulary $C_{10} = \{c_1, \ldots, c_{10}\}$ (4(e)) learned with $k_c = 10$.

The results for the training data with varying k_c over BoC and k_{nn} are shown in Table 1. And the results for the test data with varying k_{nn} are shown in Table 2. We use these results to choose the parameters. Since the selection of number of clusters affects the result of accuracy during image retrieval, five numbers of clusters were chosen: 10, 20 ,200, 500 and 1,000. Results indicate that classification performance has been improved by using 200 clusters.

We applied the optimal vocabulary size $k_c = 200$ on the test data As seen in the confusion matrices in Figure 6, there are more misclassified color images using SIFT than BoC such as in histopathology (HX), general photos (PX) or fluorescence (FL). On the other hand, using SIFT, there are fewer mistakes in radiology images (grey level) such as magnetic resonance imaging (MR), angiography (AN) or x–ray (XR). Figure 6(c) shows that the fusion of SIFT and BoC reduces the number of errors in both, color and typically grey level image types.

Using only SIFT, the best accuracy is 62.5% and results are stable for varying k_{nn}. For BoC the best accuracy is 63.96%, also quite stable across varying k_{nn}. For each k_{nn}, the fusion of BoC and SIFT produces an improved accuracy. The best overall fused result is 72.46%.

The k_{nn} voting is a very simple but often powerful tool [31]. We looked for the optimal k_{nn} value using the accuracy on the training data (Table 1). We also showed several k values on the training and test data to show the relative



(a) CT (b) HX

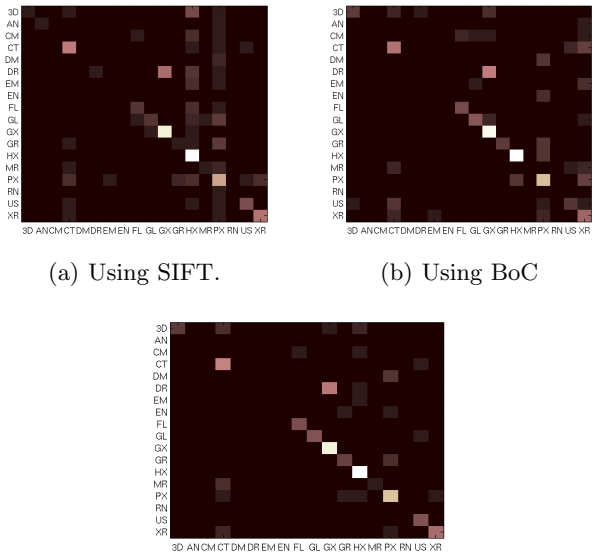
(c) PX (d) GX

(e) $C_{20} = \{c_1, \ldots, c_{20}\}$.

Fig. 5. Average BoC for four modalities corresponding to the color vocabulary $C_{20} = \{c_1, \ldots, c_{20}\}$ (4(e)) learned with $k_c = 20$.

Table 1. Classification accuracy using BoC/SIFT/both with varying k_c and k_{nn} over the training data.

k_{nn}	SIFT	BoC & k_{10}	BoC & k_{20}	BoC & k_{200}	BoC & k_{500}	BoC & k_{1000}
2	27.27	14.83	24.40	24.83	24.84	24.40
3	28.06	17.80	27.58	26.81	26.81	27.03
4	28.56	16.15	28.13	28.46	28.46	28.24
5	28.85	17.80	27.80	28.90	29.23	28.46
6	29.15	17.14	28.57	**30.11**	29.45	30
7	28.95	19.01	28.57	29.78	29.34	28.57
8	28.85	19.01	29.78	**30.11**	29.01	28.68
9	**29.45**	18.57	29.67	28.68	28.79	28.79
10	29.35	18.35	29.45	29.12	29.23	28.35
11	29.05	18.57	28.79	29.12	29.12	29.23
12	29.15	18.79	28.90	29.01	29.89	29.01
13	28.75	18.79	29.34	29.12	28.68	29.01
14	28.95	18.57	29.23	29.45	28.79	28.68
15	29.35	18.79	29.45	28.79	28.46	28.35
16	29.25	18.02	28.90	28.35	28.68	28.35
17	**29.45**	18.35	28.79	27.91	28.57	28.79
18	29.25	18.57	29.45	27.80	28.46	27.91
19	29.15	18.35	29.56	27.91	28.46	27.80



(a) Using SIFT. (b) Using BoC

(c) Using SIFT and BoC

Fig. 6. Confusion Matrices obtained for the classification results using three feature types.

stability of the results. Furthermore, we tested Support Vector Machines (SVM) for our experiments. We used the Gaussian Radial Basis Function (RBF) kernel provided by WEKA[5] optimizing the parameters over the test set. The results were not as good as k_{nn} (Table 3), probably due to the characteristics of the database or the distribution of the features used.

The best result in the modality classification task of ImageCLEF 2011 using visual methods [22] was obtained by Xerox research with 83.59% accuracy. This result is not comparable with our technique as the improvement was mainly due to an increased training set using data other than the original training data. Without the additional training set the obtained performance was at only 62.2% [32]. The best accuracy using visual methods without increasing the training data was 69.72%, obtained by the University of Campinas [33]. Using our fusion strategy of BoC and SIFT a better accuracy was obtained.

4 Conclusions and Future Work

In this paper, we present a BoC model for biomedical image categorisation. This domain is important for applications that aim at the integration of image retrieval tools into real applications We showed that fusing BoC and SIFT leads

[5] `http://www.cs.waikato.ac.nz/ml/weka/`

Table 2. Classification accuracy using BoC/SIFT/both with varying k_{nn} on the test data.

k_{nn}	SIFT	BoC	SIFT+BoC	k_{nn}	SIFT	BoC	SIFT+BoC
2	59.77	54.20	63.96	11	61.23	63.28	**72.46**
3	60.94	59.47	69.14	12	60.94	63.09	71.58
4	62.01	59.96	70.61	13	61.23	62.40	72.17
5	62.21	62.60	71.39	14	61.52	63.18	71.48
6	**62.50**	62.99	70.61	15	61.04	63.18	70.61
7	62.40	62.60	71.19	16	60.55	**63.96**	70.70
8	**62.50**	63.48	71.58	17	61.04	63.67	70.51
9	61.82	62.89	71.29	18	60.64	63.38	70.41
10	61.62	63.48	70.61	19	60.16	63.67	70.12

Table 3. Classification accuracy using BoC/SIFT/both using a simple SVM over the training data.

Features	SIFT	BoC	SIFT+BoC
Accuracy	15.92	63.09	18.95

to good results for classifying biomedical document images. Results obtained by this approach demonstrate the notable improvement using BoC and SIFT together and also compared to 15 other research groups participating in ImageCLEF who work on the same data and in the exact same evaluation setting. There are other classification scheme like the one proposed by in [34]. We chose ImageCLEF because it has established standar database where we could compare aproches. However,the classes and ground truth provided by ImageCLEF are quite limited, ambiguous and, hence, reduces the quality of results obtaines.

Several directions are foreseen for future work. We plan to increase the training set for improved results as shown by Xerox in the competition in 2011. With a database available that is much larger than the training and test data sets this should be easily feasible. Using a different color space or adding additional features can be another option but can be expected to lead to only small improvements. Particularly the text captions can be used for classification improvement as some classes can easily be distinguished using the captions. Such mixtures of visual and textual methods equally have a potential for important performance improvements.

Acknowledgments. This work was partially funded by the EU in the context of the KHRESMOI (257528) and PROMISE (258191) FP7 projects.

References

1. Hunter, L., Cohen, K.B.: Biomedical language processing: What's beyond pubmed? Molecular Cell **21**(5) (Mar 2006) 589–594

2. Depeursinge, A., Duc, S., Eggel, I., Müller, H.: Mobile medical visual information retrieval. IEEE Transactions on Information Technology in BioMedicine **16**(1) (January 2012) 53–61
3. Hersh, W., Jensen, J., Müller, H., Gorman, P., Ruch, P.: A qualitative task analysis for developing an image retrieval test collection. In: ImageCLEF/MUSCLE workshop on image retrieval evaluation, Vienna, Austria (2005) 11–16
4. Müller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., Geissbuhler, A.: Health care professionals' image use and search behaviour. In: Proceedings of the Medical Informatics Europe Conference (MIE 2006). IOS Press, Studies in Health Technology and Informatics, Maastricht, The Netherlands (August 2006) 24–32
5. Hersh, W.R., Hickam, D.H.: How well do physicians use electronic information retrieval systems? Journal of the American Medical Association **280**(15) (1998) 1347–1352
6. Hoogendam, A., de Vries F. Robbé, A.F.S.P., Overbeke, A.J.: Answers to questions posed during daily patient care are more likely to be answered by uptodate than pubmed. Journal of Medical Internet Research **10**(4) (2008)
7. Kahn, C.E., Thao, C.: Goldminer: A radiology image search engine. American Journal of Roentgenology **188**(6) (2007) 1475–1478
8. Barry Rafkind, Minsuk Lee, S.f.C.H.Y.: Exploring text and image features to classify images in bioscience literature. In: Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis, New York, NY, USA (2006) 73–80
9. Demner-Fushman, D., Antani, S., Siadat, M.R., Soltanian-Zadeh, H., Fotouhi, F., Elisevich, K.: Automatically finding images for clinical decision support. In: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops. ICDMW '07, Washington, DC, USA, IEEE Computer Society (2007) 139–144
10. Pentland, A.P., Picard, R.W., Scarloff, S.: Photobook: Tools for content–based manipulation of image databases. International Journal of Computer Vision **18**(3) (June 1996) 233–254
11. Lakdashti, A., Moin, M.S.: A new content-based image retrieval approach based on pattern orientation histogram. In Gagalowicz, A., Philips, W., eds.: MIRAGE. Volume 4418 of Lecture Notes in Computer Science., Springer (2007) 587–595
12. Jain, A.K., Vailaya, A.: Image retrieval using color and shape. Pattern Recognition **29**(8) (1996) 1233–1244
13. van de Sande, K.E., Gevers, T., Snoek, C.G.: A comparison of color features for visual concept classification. In: Proceedings of the 2008 international conference on Content-based image and video retrieval. CIVR '08, New York, NY, USA, ACM (2008) 141–150
14. Tou, J.Y., Tay, Y.H., Lau, P.Y.: Recent trends in texture classification: A review. In: Symposium on Progress in Information & Communication Technology, Kuala Lumpur, Malaysia (2009) 63–68
15. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognition **37**(1) (2004) 1–19
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
17. J.Burghouts, G., Geusebroek, J.M.: Performance evaluation of local colour invariants. Compututer Vision and Image Understanding **113**(1) (2009) 48–62
18. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 32. (sep 2010) 1582 – 1596

19. Ai, D., Han, X.H., Ruan, X., Chen, Y.W.: Adaptive color independent components based sift descriptors for image classification. In: ICPR, IEEE (2010) 2436–2439
20. Wengert, C., Douze, M., Jégou, H.: Bag–of–colors for improved image search. In: Proceedings of the 19th ACM international conference on Multimedia. MM '11, New York, NY, USA, ACM (2011) 1437–1440
21. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis & Machine Intelligence **27**(10) (2005) 1615–1630
22. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., García Seco de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum). (September 2011)
23. Sharma, G., Trussell, H.J.: Digital color imaging. IEEE Transactions on Image Processing **6**(7) (1997) 901–932
24. Banu, M., Nallaperumal, K.: Analysis of color feature extraction techniques for pathology image retrieval system, IEEE (2010)
25. Grauman, K., L.B.: Visual Object Recognition. (2011)
26. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume 1., University of California Press (1967) 281–297
27. Hinneburg, A., Keim, D.A.: An efficient approach to clustering in large multimedia databases with noise. In: Conference on Knowledge Discovery and Data Mining (KDD). Volume 5865., AAAI Press (1998) 58–65
28. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1) (1991) 11–32
29. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, New York, NY, USA, ACM (November 2005) 399–402
30. Hand, D.J., Mannila, H., Smyth, P.: Principles of Data Mining (Adaptive Computation and Machine Learning). The MIT Press (2001)
31. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Computer Vision and Pattern Recognition. (2008) 1–8
32. Csurka, G., Clinchant, S., Jacquet, G.: XRCE's participation at medical image modality classification and ad–hoc retrieval task of ImageCLEFmed 2011. In: Working Notes of CLEF 2011. (2011)
33. Faria, F.A., Calumby, R.T., Torres, R.d.S.: RECOD at ImageCLEF 2011: Medical modality classification using genetic programming. In: Working Notes of CLEF 2011. (2011)
34. Deserno, T.M., Antani, S., Long, L.R.: Content–based image retrieval for scientific literature access. Methods of Information In Medicine **48**(4) (July 2009) 371–380