International Journal of Computer Assisted Radiology and Surgery manuscript No.
(will be inserted by the editor)

Case–based Fracture Image Retrieval

Xin Zhou · Richard Stern · Henning Müller

Received: date / Accepted: date

Abstract

Purpose Case–based fracture image retrieval can assist surgeons in decisions regarding new cases by supplying visually similar past cases. This tool may guide fracture fixation and management through comparison of long–term outcomes in similar cases.

Methods A fracture image database collected over 10 years at the orthopaedic service of the University Hospitals of Geneva was used. This database contains 2,690 fracture cases associated with 43 classes (based on the AO/OTA classification). A case–based retrieval engine was developed and evaluated using retrieval precision as a performance metric. Only cases in the same class as the query case are considered as relevant. The Scale Invariant Feature Transform (SIFT) is used for image analysis. Performance evaluation was computed in terms of Mean Average Precision (MAP) and early precision (P10, P30). Retrieval results produced with the GNU Image Finding Tool (GIFT) were used as a baseline.

Two sampling strategies were evaluated. One used a dense 40x40 pixel Grid sampling and the second one used the standard SIFT features. Based on dense pixel Grid sampling, three unsupervised feature selection strategies were introduced to further improve retrieval performance. With dense pixel Grid sampling, the

X. Zhou
Geneva University Hospitals and University of Geneva, Rue Gabrielle–Perret–Gentil 4, Geneva, Switzerland.
Tel.: +41-22-3726279, Fax: ++41-22-3728879, E-mail: xin.zhou@unige.ch

R. Stern
Division of Orthopaedics and Trauma Surgery, Geneva University Hospitals and University of Geneva (HUG), Rue Gabrielle–Perret–Gentil 4, Geneva, Switzerland.

H. Müller
MedGIFT group, Business Information Systems, University of Applied Sciences Western Switzerland (HES–SO), TechnoArk 3, Sierre, Switzerland.

image is divided into 1,600 (40x40) square blocks. The goal is to emphasize the salient regions (blocks) and ignore irrelevant regions. Regions are considered as important when a high variance of the visual features is found. The first strategy is to calculate the variance of all descriptors on the global database. The second strategy is to calculate the variance of all descriptors for each case. A third strategy is to perform a thumbnail image clustering in a first step, and then to calculate the variance for each cluster. Finally, a fusion between a SIFT–based system and GIFT is performed.

Results A first comparison on the selection of sampling strategies using SIFT features shows that dense sampling using a pixel Grid (MAP=0.18) outperformed the SIFT detector–based sampling approach (MAP=0.10). In a second step, three unsupervised feature selection strategies were evaluated. A grid parameter search is applied to optimize parameters for feature selection and clustering. Results show that using half of the regions (700 or 800) obtains the best performance for all three strategies. Increasing the number of clusters in clustering can also improve the retrieval performance. The SIFT descriptor variance in each case gave the best indication of saliency for the regions (MAP=0.23), better than the other two strategies (MAP=0.20 and 0.21). Combining GIFT (MAP=0.23) and the best SIFT strategy (MAP=0.23) produced significantly better results (MAP=0.27) than each system alone.

Conclusions A case-based fracture retrieval engine was developed and is available for online demonstration. SIFT is used to extract local features and three feature selection strategies were introduced and evaluated. A baseline using the GIFT system was used to evaluate the salient point based approaches. Without supervised learning, SIFT–based systems with optimized parameters slightly outperformed the GIFT system. A fusion of the two approaches shows that the information contained in the two approaches is complementary. Supervised learning on the feature space is foreseen as the next step of this study.

Keywords content–based image retrieval · feature selection · fracture database · medical imaging · decision support system

1 Introduction

At the orthopaedic service of the University Hospitals of Geneva fracture cases have been collected and stored for more than ten years in a teaching file called Casimage[1] [37]. Images added are mostly radiographs before the initial intervention or immediately after the operation. Whenever available, images are added when the patient comes for a follow–up visit. Cases are classified using the AO/OTA (Arbeitsgemeinschaft für Osteosynthesefragen/Orthopaedic Trauma Association) fracture classification [26]. Subsets of the data have been made available for example in the form of teaching CDs and books [41].

The primary goal of building such a data set is to supply surgeons with examples of fracture cases with a surgical intervention. Fractures are among the most common orthopedic problems. For cases where the bone displacement (fracture

[1] `http://pubimage.hcuge.ch/`

gap) or angulation is large, surgical interventions can be required. Many clinicians choose the method they are most familiar with and perhaps not the one that could lead to best results. It can be beneficial for surgeons to see how similar fractures in past patients were operatively stabilized by other surgeons. Some cases also contain images of follow–up visits, making it possible to evaluate techniques based on long–term outcome.

So far, the search for fracture cases in Casimage is either performed using the AO/OTA classification or sometimes by patient number. The pre–operative and immediate post–operative radiographs or CT images from similar fractures contain essential information to provide decision support to surgeons. Searching for similar cases using images of a new case can be a complementary scenario for a better use of the data stored. In this article, a content–based image retrieval system is developed to enable such a search for fracture cases.

Content–based image retrieval (CBIR) usually extracts visual information of images automatically and then allows searching for images similar to example(s). Many articles have been published on CBIR for general images [39, 10] as well as medical images [33, 21, 30, 29, 32]. Several research groups have worked on fracture image analysis as well, but rarely on fracture retrieval. Leow et al. [22, 25, 17] proposed using contour–based segmentation and a gradient map inside a contour to detect femur bone fractures in X–ray images, Donnelley et al. [13–15] developed a long bone fracture detection system using similar techniques. Systems proposed by Leow or Donnelley were both fracture detection systems that aimed at detecting hard–to–find fractures such as occult bone fractures. They did not provide retrieval functionalities for finding cases with similar fractures. The similar Fracture Image Retrieval technique (IFIR) was proposed in [35] using gray level based co–occurrence matrices to extract features. IFIR provides only single image–based retrieval and no case–based retrieval functionality. These publications did not provide online demos, which makes testing the proposed solutions difficult. Several online demo systems for medical CBIR exist such as MedGIFT[2] (Medical GNU Image Finding Tool [31]), IRMA[3] (Image Retrieval in Medical Applications) [21], FIRE[4] (Flexible Image Retrieval Engine) [12, 11], and SPIRS[5] (Spine Pathology and Image Retrieval System) [18]. To our knowledge, these medical CBIR online demos are currently image–based rather than case–based. The image–based approach uses single images as query, which is usually not the unit of information in clinical use. This approach is suffering from the fact that single images can hardly provide enough description for a complete case [34]. Case–based retrieval taking into account several images and potentially other data of the case has also been proposed by other authors recently [36]. The case–based retrieval system described in this paper focuses on images only. One fracture case contains various possibilities of representation (for example a frontal view and a lateral view, or inter–sectional CT scan). We look for suitable features and parameters to establish a flexible representation to enable comparison of cases. An online demo of this system using a set of the fracture cases extracted in 2009 is available.

[2] http://medgift.unige.ch/demo/

[3] http://ganymed.imib.rwth-aachen.de/irma/onlinedemos.php

[4] http://www-i6.informatik.rwth-aachen.de/~deselaers/cgi_bin/fire.cgi?port=12961

[5] http://archive.nlm.nih.gov/proj/spirs.php

The selection of image analysis techniques depends on the nature of the data set. One of the most important challenges of fracture retrieval is that bone fractures generate only very small local changes. Differences in anatomy between individuals are often more important than the difference due to a small fracture. Approaches using local features are thus usually required. A large variety of local visual features to represent the images were proposed during the last decade such as GLOH (Gradient Location and Orientation Histogram) [28], SIFT (Scale Invariant Feature Transform) [24], SURF (Speeded Up Robust Features) [4], and many others [20,27,23]. In our system, image analysis is based on a combination of SIFT and BoF (Bags of Features) [8], which proved to be robust in the ImageCLEF medical image retrieval tasks [29,32]. ImageCLEF[6] [5,6] has started within CLEF[7] (Cross Language Evaluation Forum [38]) in 2003 with the goal to benchmark image retrieval in multilingual document collections. Successful participants in the ImageCLEF medical task [2,43] showed that grid sampling on a single scale often obtains better results than standard SIFT multiscale sampling as medical images are often taken under very standardized conditions and shift invariance has only a limited influence. In order to improve the retrieval performance, sampling strategies and feature selection strategies are investigated in this paper.

2 Methods

This section describes the main techniques and the data set used in this article.

2.1 Dataset used

In this article, a set of fracture cases extracted from the Casimage teaching file in 2009 was used. The data set consists of 23'970 images of 2'690 cases, classified into 43 fracture classes. Among them, 1'467 cases have post–operative images (immediate + long term). The total number of post–operative images is 7'771. Malleolar (ankle) fractures are classified according to the Danis–Weber classification [9,44]. For all other fractures, the AO/OTA classification [26] is used.

Images are grouped into cases with additional free text descriptions in several fields per case. The textual descriptions can include information on the operation, the outcome and also references to the literature, for example describing the techniques used. Each case contains from 1 to 73 images of various modalities. The number of images varies strongly as some cases can contain CT slices based on the selections of the surgeon, and some cases can include many follow up visits. The distribution of the cases based on the number of images is shown in Figure 1, showing that many cases have few images, but almost all cases have at least 3 images.

Table 1 shows the number of cases per fracture class, as well as the total and average number of images for each class. Fracture classes are available as text labels corresponding to particular classes in the AO/OTA and Danis–Weber classifications. The table shows that a few classes are quite dominant with over 10% of the cases such as femur and ankle fractures.

[6] http://www.imageclef.org/

[7] http://www.clef-campaign.org/

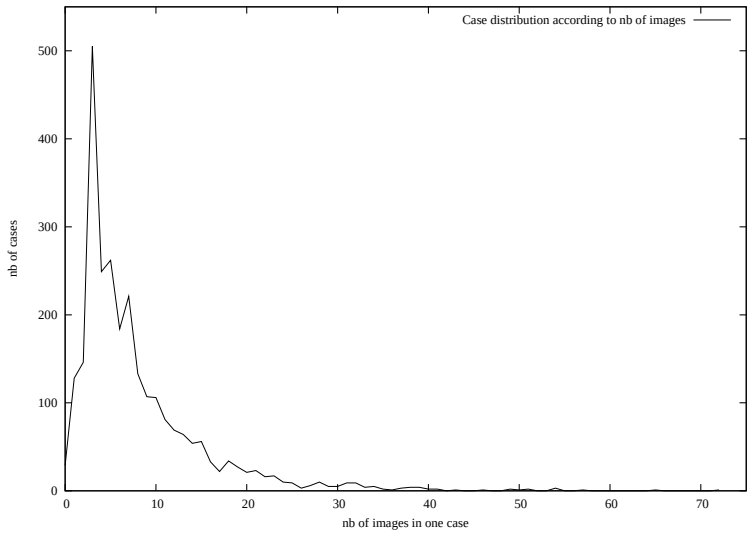

Fig. 1 Number of cases with the respective number of images per case.

Around 90% of the images are x–rays. Other image modalities include CT (Computed Tomography), MRI (Magnetic Resonance Imaging), 3D reconstruction images, Angiography, Scintigraphy, hand–drawn surgery plans and photographs of injuries. X–rays are taken from different views and have varying sizes, sometimes highlighting a region of interest. For most cases, at least two x–rays are contained in a case (one antero–posterior view and one lateral view). Sometimes a third view, oblique or external rotation is given.

2.2 Retrieval techniques

The retrieval system proposed in this paper consists of online and offline processing parts. In Figure 2, the workflow of the system is detailed. The input of the system is a fracture case. The offline part involves mainly image analysis and indexing steps. The online part includes distance measurement and fusion of the results. Image analysis is only required in the online part when a new case is submitted. Query cases can also be selected among the already analyzed cases and then no image analysis is required. As feature selection by variance is considered as one of the main novelties, it is detailed in Section 3.2. Other techniques that are reused in this paper are briefly described in the following paragraphs.

Image analysis and indexing To detect salient regions in the images, two approaches are used:

- the standard SIFT detector;
- a 40x40 *pixel grid* sampling.

The SIFT detector uses standard parameters proposed by Lowe et al. [24] (using 3 octaves, a Gaussian kernel $\sigma = 1.6$ without up–sampling the image). Both

Table 1 Number of cases and images of each fracture class.

fracture class	nb of cases	nb images	aver. nb images/case
Acetabulum	31	433	13.97
Ankle	8	60	7.50
Ankle Weber A	44	270	6.14
Ankle Weber B	244	1784	7.31
Ankle Weber C	156	1222	7.83
Calcaneus	33	340	10.30
Carpal–Other	1	5	5.00
Clavicle	40	232	5.80
Elbow	4	21	5.25
Femur–Subtrochanteric	132	1342	10.17
Femur Diaphysis	169	1960	11.60
Femur Distal–Extraarticular	55	614	11.16
Femur Distal–Intraarticular	50	673	13.46
Femur Proximal–Head	2	18	9.00
Femur Proximal–Intertrochanteric	54	406	7.52
Femur Proximal–Neck	72	499	6.93
Femur Proximal–Pertrochanteric	419	2132	5.09
Foot	3	36	12.00
Hip	3	25	8.33
Humerus Diaphysis	119	1146	9.63
Humerus Distal–Extraarticular	30	325	10.83
Humerus Distal–Intraarticular	61	590	9.67
Humerus Proximal	172	1522	8.85
Knee	7	34	4.86
Metacarpal–Phalanx hand	10	30	3.00
Metatarsal–Phalanx foot	62	476	7.68
Patella	34	198	5.82
Pelvic Ring Fracture	43	438	10.19
Radius/Ulna Diaphysis	46	262	5.70
Radius/Ulna Distal	14	68	4.86
Radius/Ulna Proximal	58	357	6.16
Scapula	10	131	13.10
Shoulder	16	73	4.56
Spine Cervical	1	1	1.00
Spine Lumbar	2	8	4.00
Spine Thoracic	1	3	3.00
Talus	32	497	15.53
Tarsal–Other	17	244	14.35
Tibia/Fibula Diaphysis	205	2160	10.54
Tibia/Fibula Distal–Extraarticular	52	737	14.17
Tibia/Fibula Distal–Intraarticular	57	599	10.51
Tibia/Fibula Proximal–Extraarticular	23	310	13.48
Tibia/Fibula Proximal–Intraarticular	98	1689	17.23
total	2690	23970	8.91

described approaches use SIFT descriptors as visual features and the BoF (also called BoW, Bag of visual keyWords [8]) as image representation. In order to reduce the feature space, visual features need to be categorized. This step is commonly

[8] The definition of visual keywords varies in the literature. In certain articles [3] *visual keywords* imply a supervised learning process to attach visual features to semantic labels. Other articles [1,19] refer to *visual keywords* meaning that the feature set is obtained by unsupervised learning. This includes feature clustering, where the supervised learning and well defined semantic meaning are not necessary. In this paper, the second definition is chosen.

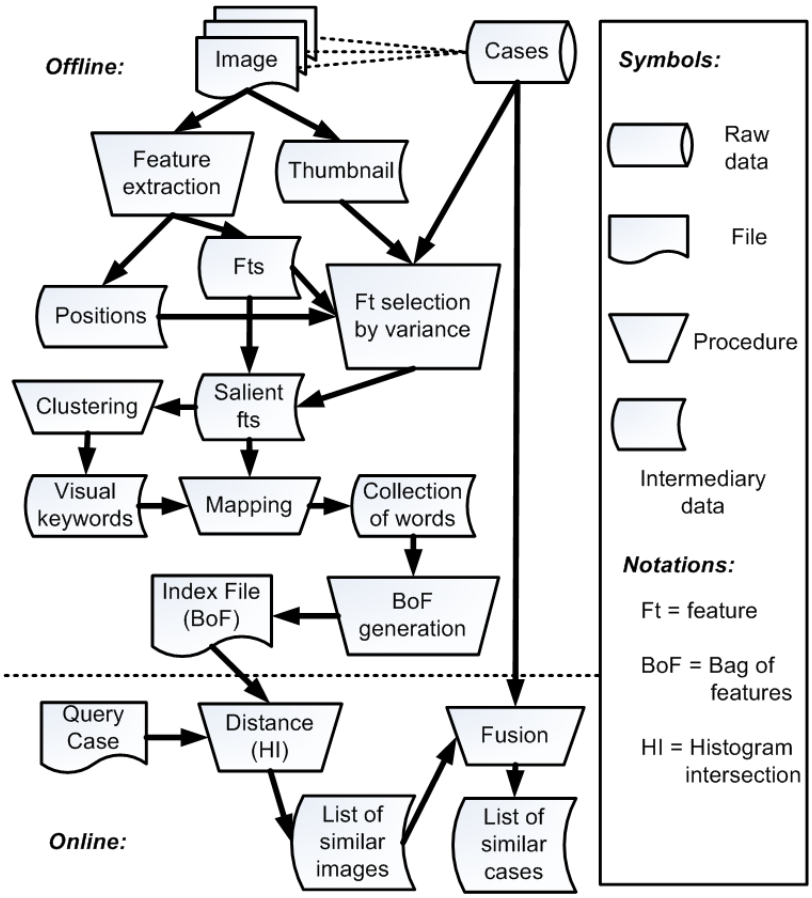

Fig. 2 Workflow of fracture retrieval.

named feature clustering step. A griddified version of the hierarchical KMeans quantizer [49] was developed for this step. The obtained cluster centers are also named *visual keywords*, a *visual dictionary* or a *visual codebook* in the literature.

The optimal number of visual keywords k_d has been studied by various research groups [2,43] in the context of the ImageCLEF medical image classification task. A choice of $k_d = 1000$ is widely used and showed to be robust for many types of medical images. This is taken as default setting in our online system. However in this study other values of k_d were tested and obtained better performance. After mapping all features to the visual keywords a statistical analysis is performed. Each image is described in terms of a histogram of visual words (BoF representation). The distance between two images is calculated with the HI (histogram intersection) [42]. No supervised machine learning technique is used in this paper.

Fusion strategies Technically, case–based retrieval differs from image–based retrieval in both query formulation and result presentation. In this paper we concentrate on purely visual case–based retrieval. In contrast to image–based retrieval, where visual similarity of single images is used for cases–based retrieval the visual

similarities of all images in a case need to be taken into account to calculate visual distances between cases. Most often, also clinical data of the cases is used to calculate case similarities. As for the fractures no structured data is available and a goal of the text was really to concentrate on the visual aspects for similarity calculation. The fusion of image similarities is performed in the following way to obtain case similarities:

Let the query case C_q contains n images $I_{q(i)}$, where $0 < i \leq n$ and q represents the query. The distances between $I_{q(i)}$ and all other images in the database are calculated with the HI. Thus, $I_{q(i)}$ generates a set of similar images $\mathcal{I}_{s(i)}$ ranked by distance (s for *similar*). To obtain a set of similar cases $\mathcal{C}_{s(i)}$, each image of $\mathcal{I}_{s(i)}$ is replaced by the associated case. There can be several images in $\mathcal{I}_{s(i)}$ representing the same case in each list $\mathcal{C}_{s(i)}$. We use the combMAX fusion strategy proposed by Fox et al. [16]. This strategy means that if several images represent the same case in a results list, only the image with the highest score is kept. After the first level fusion, n lists of similar cases $\mathcal{C}_{s(i)}$ $(i = 1..n)$ need to be fused into one list, depending on the number of images in the query case. A second level fusion is applied based on the combMNZ [16] strategy. The reason for using combMAX at the first level is to avoid bias due to the number of images per case that can vary strongly. This could otherwise favor cases with many images.

The choice for the second level fusion is based on previous experience [16,7, 45] showing that using combMNZ obtained stable results for a variety of problems and often has the best performance in benchmarks. The following equations detail the fusion approaches:

$$D_{\texttt{combMAX}} = \arg \max_{l=1:m} D(l), \tag{1}$$

$$D_{\texttt{combMNZ}} = (\sum_{l=1}^{n} D(l)) * F(l), \tag{2}$$

where l is a returned image, $D(l)$ the distance to l, $F(l)$ the frequency of l, m the number of images belonging to a returned case, and n the number of images in the query.

Evaluation All cases are used as query to evaluate the system. Among the returned cases, only those belonging to the same fracture class are considered as relevant, although some very similar classes exist that could be regarded as relevant as well. All other cases are considered non relevant. The query case itself is not taken into account for the performance calculation as the distance to itself is not calculated. Performance measures such as MAP (Mean Average Precision, defined in Equation 6) and early precision (P10, P30, defined in Equation 4) are used.

$$P(n_s, C_q) = n_r/n_s \tag{3}$$

$P(n_s, C_q)$ is the precision after n_s returned cases for query C_q. n_s is the number of returned cases taken into account and n_r is the number of relevant cases among the n_s returned cases.

$$P(n_s) = \sum_{q=1}^{n_c} P(n_s, C_q)/n_c \tag{4}$$

n_c is the number of cases, which is constant at 2'690. Early precision at n_s is then averaged over the precision values of all queries.

$$AverageP(C_q) = \sum_{n_s=1}^{n_c} (P(n_s, C_q) * rel(n_s))/n_{tr} \quad (5)$$

$AverageP(C_q)$ is the average precision for one query C_q. n_{tr} is the total number of relevant cases. $rel(n_s)$ is a binary function, returning 1 if the n_sth case is relevant, 0 if not.

$$MAP = \sum_{i=1}^{n_c} (AverageP(C_i))/n_c \quad (6)$$

MAP is calculated for the entire database, meaning that the average precision for each case $AverageP(C_i)$ is averaged.

Measures described in Equation 4 and Equation 6 average the performance over all cases, favoring good performance in large classes, particularly for early precision. A class–based average is necessary to measure the stability across classes. $MAP_{cl}, P10_{cl}, P30_{cl}$ are thus calculated for each class. The values averaged across classes are noted $\overline{MAP_{cl}}, \overline{P10_{cl}}, \overline{P30_{cl}}$. There are thus six performance measures (MAP, P10, P30, $\overline{MAP_{cl}}, \overline{P10_{cl}}, \overline{P30_{cl}}$) for the evaluation.

As small classes contain very few cases, the best possible precision scores are well below 1 even for a perfect system. Among 43 classes, 12 consist of less than 10 cases, 16 of less than 30 cases. Thus, the best possible $\overline{P10_{cl}}$ is 0.8095 and the best possible $\overline{P30_{cl}}$ is 0.7087).

The GNU Image Finding Tool (*GIFT*[9]) is an open source image retrieval engine [40], which was also used as a baseline in ImageCLEF for the past eight years [48, 47, 46]. Grey level and Gabor texture features are used to describe images both locally and globally in the form of local blocks and a global histogram. In this work, *GIFT* is used as a baseline. Each case is used as query with GIFT using all images inside the case separately as query. To transform the list of similar images into a list of similar cases, the combMAX fusion strategy introduced in Equation 1 is used.

3 Results

In this section, results are represented in three sub–sections. In the first sub–section, a comparison is used to evaluate the SIFT+BoF approach with two sampling strategies on the entire image collection. In the second sub–section, the data set is divided into training set and test set. Three feature selection strategies are used and a varying number of clusters are tested to optimize the performance. The image collection is divided by 50%–50% per class into training set (1'337 cases) and testing set (1'353 cases). 50%–50% per class means if one class contains $2 * a$ cases, a cases are randomly selected into training set and the rest of cases are taken by testing set. As not all the classes have even number of cases, for classes containing odd number $(2 * a + 1)$ of cases, only a cases are randomly selected into training set, and the testset will have one more case($a + 1$ cases) than training

[9] `http://www.gnu.org/software/gift/`

Table 2 Comparison of sampling strategies.

	MAP	P10	P30	$\overline{MAP_{cl}}$	$\overline{P10_{cl}}$	$\overline{P30_{cl}}$
SIFT detector	0.10	0.12	0.11	0.08	0.06	0.06
40x40 pixel grid sampling	0.18	0.25	0.21	0.10	0.14	0.10
GIFT baseline	0.23	0.38	0.31	0.15	0.20	0.15

set. The training data are only used to select suitable parameters for the number of clusters k_d and regions N. No kernel–based feature space transformation is performed in this stage (i.e no machine learning such as SVM or neural network is used, only the choice of two parameters is learned based on the training data). Cross–validation is performed 10 times to obtain average values for the 6 evaluation measures. In the third sub–section, fusion is applied to combine the SIFT–based and GIFT–based approaches, which further improves the result.

3.1 Sampling strategies

Table 2 shows the performance obtained using the SIFT+BoF approaches with various sampling strategies.

In our experiment, a 40x40 pixel grid sampling provides better results than using the SIFT detectors. Both SIFT+BoF approaches are below the GIFT baseline. GIFT uses a *tf/idf* feature weighting on a large number of low level features (over 80'000 possible features), whereas our SIFT+BoF approach does not apply any feature weighting or selection strategy.

3.2 Feature selection strategies

To improve the retrieval performance, three unsupervised feature selection strategies based on variance are proposed. Based on a 40x40 pixel grid, the image is divided into 1600 square blocks, noted $blk(x, y)$ $(0 < x <= 40, 0 < y <= 40)$. One feature is extracted from each block. Blocks are considered salient when high variance of the features is found in the region. The goal is to sort the blocks by the variance of features. Then, a feature selection is performed to keep only the N most salient blocks and not the others before the creation of the visual keywords. The first strategy is to calculate the variance $Var(x, y)_{overall}$ of all features inside a block $blk(x, y)$ for the entire database. The second strategy is to calculate the variance $Var(x, y)_{case}$ of features inside a block $blk(x, y)$ for each case separately C_j $(0 < j <= 2690)$. The third strategy is to perform a thumbnail clustering with a parameter k_t in the first step, and then to calculate the variance $Var(x, y)_{thClst}$ inside a block $blk(x, y)$ for each cluster $Clst_g$ $(0 < g <= k_t)$.

In Figure 3 the variance for all descriptors in each position is shown as an example. Brightness represents the regions where high variance occurs. We can see that border regions most often have a lower variance and are thus less important.

The 1600 regions are sorted by variance as presented in Figure 3. As the variance is calculated using features from the entire database it favors large image classes. We highlight half (800) of the regions with globally high variance on two example images (Figure 4) to show one important drawback of this strategy: the

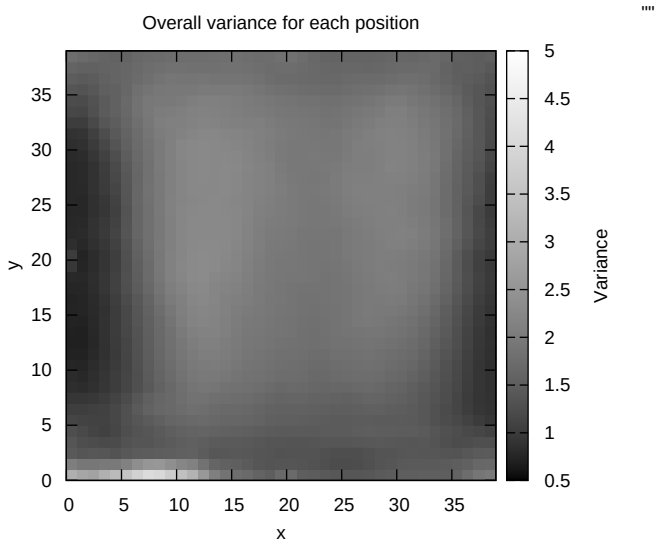

Fig. 3 Overall variance $Var(x, y)_{overall}$ for each position in the images.

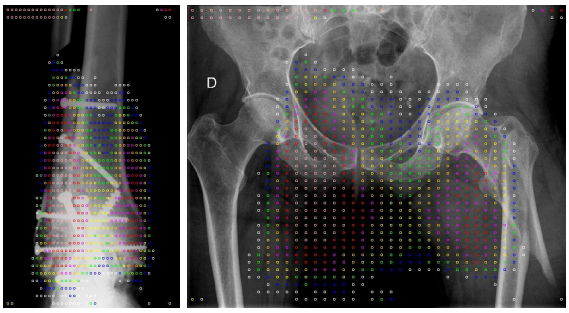

Fig. 4 Example of the 800 most salient regions in two fracture classes (left: ankle, right: femur), Different colors represent the variance (pink: 1-100, red: 101-200, magenta: 201-300], orange: 301-400, yellow: 401-500, green: 501-600, blue: 601-700, white: 701-800).

variation of images between different classes is not taken into account. For example, in Figure 4 the ankle is perfectly covered by the 800 regions, whereas the femur bone structure is in a large part outside of the 800 regions with the highest global variance. Regions other than the selected 800 are not used for the extraction of features, which can create an information loss for some classes.

To solve the problem of global variance, the variance of all images in a case $Var(x, y)_{case}$ and per cluster $Var(x, y)_{thClst}$ (based on thumbnail images) are calculated. Each case/cluster thus has specific salient regions selected. Figure 5 illustrates the frequency of each region in the 800 best positions when using the variance on a per case basis.

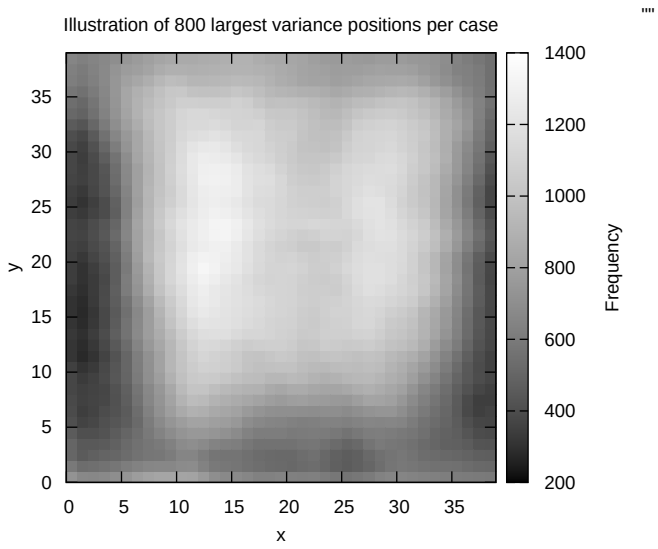

Fig. 5 Frequency of the 800 regions with highest variance per case.

Feature selection reduces the number of features to be passed to the K–means clustering, which influences the selection of the clustering parameter k_d. For the first two strategies, we study both N and k_d in the range of $[100 : 1600]$ in steps of 100, which forms a parameter grid. For the third strategy, a dimension k_t in needed. Only $k_t = 50, 70, 100$ are investigated in our case to limit computation time. Grid search is applied for 10 randomly generated training sets to optimize the parameter settings. In total $16 * 16 * 10 = 2560$ calculations are required for the first two strategies. For the third strategy, $3 * 2560$ calculations are required to evaluate the performance. Only the best parameters are afterwards validated on the test set.

In Figure 6, Figure 7 and Figure 8 the performance for these calculations is presented. For each curve, N is fixed and k_d varies between 100 and 1600. For each feature selection strategy, 16 curves can be printed (from $N = 100$ to $N = 1600$). In order to avoid overloading the image, only the curves of the five best results are shown. Curves are compared based on their best overall MAP.

In Figure 6, Figure 7 and Figure 8, results with small k_d obtain often low performance. This can be due to K–means clustering depending on the starting points, which are randomly selected. When $k_d > 700$, all curves become more stable. Results show that in our case, the performance is always increasing when features are clustered into a larger number of clusters. Best results are always obtained by $k_d = 1600$, which implies that higher k_d may obtain even better results.

For the $Var(x, y)_{overall}$ strategy, the best results from the first to fifth are respectively $N = 700, N = 800, N = 600, N = 400$ and $N = 900$. For $Var(x, y)_{case}$,

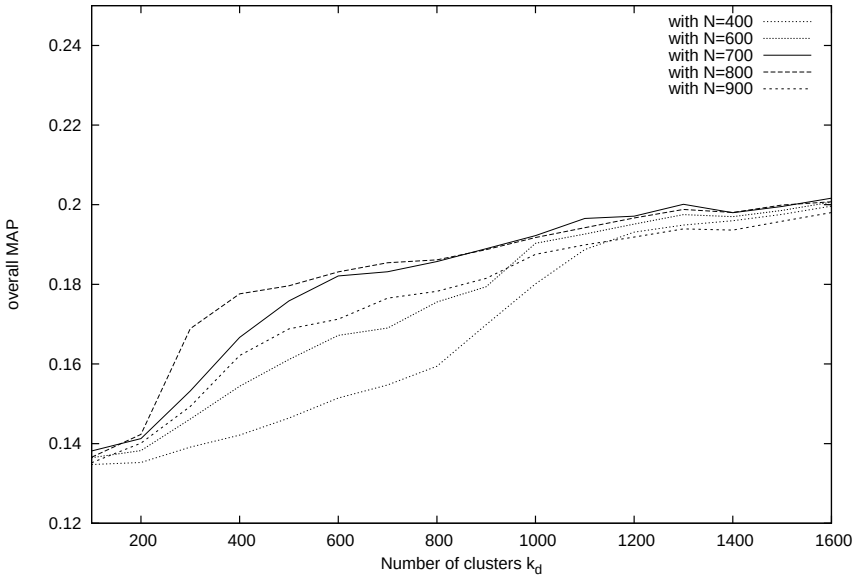

Fig. 6 The curves of the five best results for feature selection by overall variance.

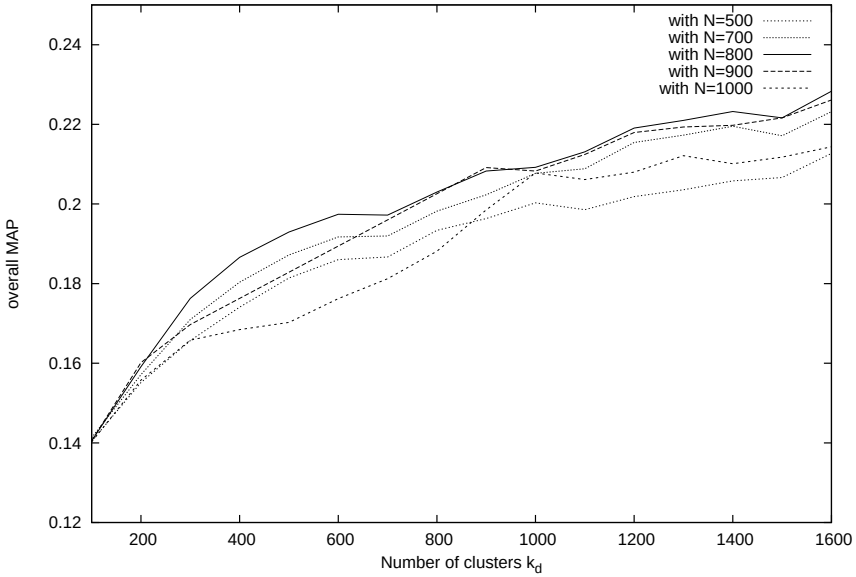

Fig. 7 The curves of the five best results for feature selection by variance per case.

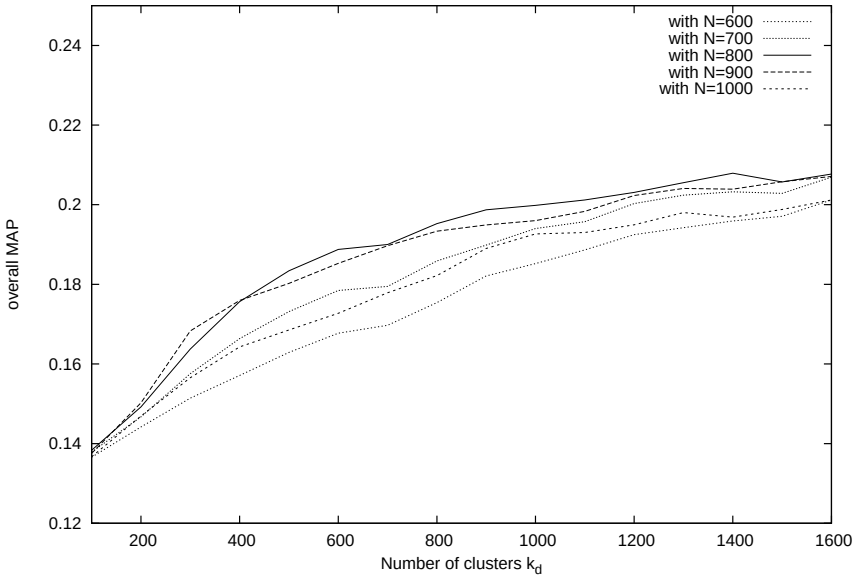

Fig. 8 The curves of the five best results (with $k_t = 100$) for feature selection by variance per cluster.

best results from the first to fifth are respectively $N = 800, N = 900, N = 700, N = 1000$ and $N = 500$. For $Var(x, y)_{cluster}$, best results are obtained with $N = 800, N = 900, N = 700, N = 1000$ and $N = 600$. The differences of performance are often around 0.001–0.002, so rather small. Compared to the default settings ($k_d = 1000$), using $k_d = 1600$ improves the MAP by 0.01–0.02.

Best runs for the three feature selection strategies are shown in Table 3. Parameters and scores are also presented. Case–based feature selection ($Var(x, y)_{case}$) obtained the best performance and stability. It slightly outperforms the GIFT baseline in terms of MAP and $\overline{MAP_{cl}}$ both on the training and the test data set, although GIFT is slightly better in early precision.

Compared with results listed in Table 2, feature selection improves the performance significantly. Even without supervised machine learning, 800 regions of high variance per image constantly obtain good results, better than using all regions.

3.3 Fusion of GIFT and SIFT

Combining the SIFT–based system and GIFT using combMNZ as defined in Equation 2 improves the results. The SIFT–based approach is best with the $Var(x, y)_{case}$ feature selection and use the best parameters learned from training data. Results show that fusion improves the performance for both MAP and early precision. This increase in the fused result shows that both approaches model different information and are thus partly complementary.

Table 3 Comparison of the three feature selection strategies.

Training set	N	k	MAP	P10	P30	$\overline{MAP_{cl}}$	$\overline{P10_{cl}}$	$\overline{P30_{cl}}$
$Var(x,y)_{overall}$	700	1600	0.2016	0.3270	0.2626	**0.1731**	0.1965	0.1380
$Var(x,y)_{case}$	800	1600	**0.2283**	**0.3558**	**0.2872**	0.1708	**0.1978**	**0.1391**
$Var(x,y)_{thClst}$	800	1400	0.2079	0.3310	0.2697	0.1680	0.1889	0.1300
GIFT baseline			0.2271	0.3691	0.2914	0.1641	0.2001	0.1413
Testing set	N	k	MAP	P10	P30	$\overline{MAP_{cl}}$	$\overline{P10_{cl}}$	$\overline{P30_{cl}}$
$Var(x,y)_{overall}$	700	1600	0.1997	0.3094	0.2410	0.1667	0.1842	0.1308
$Var(x,y)_{case}$	800	1600	**0.2277**	**0.3479**	**0.2860**	**0.1978**	**0.1881**	0.1303
$Var(x,y)_{thClst}$	800	1400	0.2064	0.3261	0.2644	0.1645	0.1645	0.1495
GIFT baseline			0.2266	0.3557	0.2889	0.1611	0.1966	0.1320

Table 4 Fusion of GIFT and the SIFT–based approach.

	MAP	P10	P30	$\overline{MAP_{cl}}$	$\overline{P10_{cl}}$	$\overline{P30_{cl}}$
GIFT+SIFT (combMNZ)	0.2680	0.4076	0.3192	0.2018	0.2543	0.1711

4 Conclusions

In this article two SIFT sampling strategies together with three variance–based feature selection strategies are proposed for medical case retrieval on a database containing fracture images. The goal was to improve the performance of the visual case retrieval system. The GIFT retrieval system was used as the baseline for the evaluation as it has shown to have good performance in absence of training data in the past.

A dense sampling strategy such as a 40x40 pixel grid performed better than the SIFT detector–based sampling, which is in agreement with the conclusion obtained by [2,43]. This is due to the very standardized image acquisition protocols and thus little need for shift, rotation and scale invariance, which are the strong points of many visual word approaches. Salient point–based region detection such as SIFT provides only a sparse sampling. The advantage of using SIFT is that it provides a smaller number of high quality features, generating a relatively low–dimensional feature space. Supervised machine learning requires the dimensionality of the input feature space to be low, as it may extend this feature space. In cases where supervised learning is not applied, the retrieval performance can be limited, as part of the information is not taken into account. Dense sampling keeps a majority of the global information without a learning process. It can thus outperform the salient point–based approaches.

Feature selection can be considered as an unsupervised learning strategy. Using a variance–based feature selection improved the performance by up to 0.05. Computing variance per case showed to be the best strategy. Overall results with the visual word approach are not very different from the GIFT baseline but the system uses a much smaller number of features and still has several possibilities for optimization. GIFT proved to be very robust but is a rather closed system. Learning strategies on the GIFT feature space have shown to have a very limited potential in the past. Combining the layout–based features of GIFT with the local features of our approach leads to much better results.

Different from the conclusion stated in [2,43], in our tests the retrieval performance continues to increase when the number of clusters k_d increases. This can be due to the fact that only a simple histogram intersection is used as distance measure and no supervised learning strategy such as SVMs. In [2,43] SVMs are always applied for feature–level machine learning. In the future we plan to take into account machine learning on the feature space rather than on specific parameters to increase the performance. Performance should particularly increase for the large classes which have sufficient training data. Stability also needs to be taken into account in order to not reduce performance of small classes too much as the number of cases per class in the described database is extremely heterogeneous.

The current version of the retrieval system is available online at[10]. The system is an important step towards adding a visual retrieval functionality to the Casimage database created by the surgeons or to other clinical case databases.

Acknowledgements This work was partially supported by SWITCH/AAA in the context of the medLTPC project and the 7th Framework program of the European Union in the context of the Khresmoi and Promise projects (grant agreements 257528 and 258191).

References

1. Abdullah, A., Veltkamp, R.C., Wiering, M.A.: Ensembles of novel visual keywords descriptors for image categorization. In: 11th International Conference on Control Automation Robotics Vision (ICARCV), pp. 1206–1211 (2010)
2. Avni, U., Greenspan, H., Konen, E., Sharon, M., Goldberger, J.: X–ray categorization and retrieval on the organ and pathology level, using patch–based visual words. IEEE Trans Med Imaging **73**(11), to appear (2010)
3. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. J. Mach. Learn. Res. **3**, 1107–1135 (2003)
4. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded–up robust features (surf). Computer Vision and Image Understanding **110**(3), 346–359 (2008)
5. Clough, P., Müller, H., Sanderson, M.: The CLEF cross–language image retrieval track (ImageCLEF) 2004. In: C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck, B. Magnini (eds.) Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign, *Lecture Notes in Computer Science (LNCS)*, vol. 3491, pp. 597–613. Springer, Bath, UK (2005)
6. Clough, P., Sanderson, M., Müller, H.: The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In: The Challenge of Image and Video Retrieval (CIVR 2004), Springer Lecture Notes in Computer Science, pp. 243–251 (2004)
7. Croft, W.B.: Combining approaches to information retrieval. In: Advances in Information Retrieval, pp. 1–36. Springer US (2000)
8. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: In Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
9. Danis, R.: Théorie et Pratique de l'Ostéosynthèse. Masson and Cie, Paris, France (1949)
10. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys **40**(2), 1–60 (2008)
11. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. Information Retrieval **11**, 77–107 (2008)
12. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in ImageCLEF 2005: Combining content–based image retrieval with textual information retrieval. In: Working Notes of the CLEF Workshop. Vienna, Austria (2005)
13. Donnelley, M.: Computer aided long–bone segmentation and fracture detection. Ph.D. thesis, Flinders University of South Australia, Adelaide, South Australia (2008)

[10] `http://arcgift.unige.ch/~xmzh/FractureDemo/RIA.html`

14. Donnelley, M., Knowles, G.: Computer aided long bone fracture detection. In: Proceedings of the Eighth International Symposium on Signal Processing and Its Applications (ISSPA 2005), vol. 1, pp. 175–178. Sydney, AUSTRALIA (2005)
15. Donnelley, M., Knowles, G., Hearn, T.: A cad system for long–bone segmentation and fracture detection. In: Proceedings of the 3rd International Conference on Image and Signal Processing (ICISP 2008), *Lecture Notes in Computer Science*, vol. 5099, pp. 153–162. Springer (2008)
16. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: Text REtrieval Conference, pp. 243–252 (1993)
17. He, J.C., Leow, W.K., Howe, T.S.: Hierarchical classifiers for detection of fractures in x–ray images. In: Proceedings of the 12th International Conference on Computer Analysis of Images and Patterns (CAIP 2007), *Lecture Notes in Computer Science*, vol. 4673, pp. 962–969. Springer, Vienna, Austria (2007)
18. Hsu, W., Antani, S., Long, L.R., Neve, L., Thoma, G.R.: Spirs: A web-based image retrieval system for large biomedical databases. International Journal of Medical Informatics **78**(Supplement 1), S13–S24 (2009). MedInfo 2007
19. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag–of–features for object categorization and semantic video retrieval. In: CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval, pp. 494–501. ACM, New York, NY, USA (2007)
20. Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), vol. 2, pp. 506–513. Washington, DC, USA (2004)
21. Lehmann, T.M., Güld, M.O., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., Wein, B.B.: Content–based image retrieval in medical applications. Methods of Information in Medicine **43**, 354–361 (2004)
22. Lim, S.E., Xing, Y., Chen, Y., Leow, W.K., Howe, T.S., Png, M.A.: Detection of femur and radius fractures in x–ray images. In: In: Proc. 2nd Int. Conf. on Advances in Medical Signal and Information Processing, pp. 249–256. Sliema, Malta (2004)
23. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: SIFT Flow: Dense Correspondence across Different Scenes. In: ECCV '08: Proceedings of the 10th European Conference on Computer Vision, pp. 28–42. Springer-Verlag, Berlin, Heidelberg (2008)
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
25. Lum, V.L.F., Leow, W.K., Chen, Y., Howe, T.S., Png, M.A.: Combining classifiers for bone fracture detection in x-ray images. In: IEEE International Conference on Image Processing (ICIP'2005), vol. 1, pp. 1149–1152. Genoa, Italy (2005)
26. Marsh, J.L., Slongo, T.F., Agel, J., Broderick, J.S., Creevey, W., DeCoster, T.A., Prokuski, L., Sirkin, M.S., Ziran, B., Henley, B., Audigé, L.: Fracture and dislocation classification compendium - 2007: Orthopaedic trauma association classification, database and outcomes committee. Journal of Orthopaedic Trauma **21(10 Suppl)**, S1–133 (2007)
27. Matas, J.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing **22**(10), 761–767 (2004)
28. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis & Machine Intelligence **27**(10), 1615–1630 (2005)
29. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: CLEF 2007 Proceedings, *Lecture Notes in Computer Science (LNCS)*, vol. 5152, pp. 473–491. Springer, Budapest, Hungary (2008)
30. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In: CLEF 2006 Proceedings, *Lecture Notes in Computer Science (LNCS)*, vol. 4730, pp. 595–608. Springer, Alicante, Spain (2007)
31. Müller, H., Fabry, P., Lovis, C., Geissbuhler, A.: medGIFT — retrieving medical image by their visual content. In: World Summit of the Information Society, Forum Science and Society. Geneva, Switzerland (2003)
32. Müller, H., Kalpathy-Cramer, J., Kahn Jr., C.E., Hatt, W., Bedrick, S., Hersh, W.: Overview of the ImageCLEFmed 2008 medical image retrieval task. In: C. Peters, D. Giampiccolo, N. Ferro, V. Petras, J. Gonzalo, A. Peñas, T. Deselaers, T. Mandl, G. Jones, M. Kurimo (eds.) Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum, *Lecture Notes in Computer Science*, vol. 5706, pp. 500–510. Aarhus, Denmark (2009)

33. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content–based image retrieval systems in medicine–clinical benefits and future directions. Internation Journal of Medical Informatics **73**(1), 1–23 (2004)
34. Muller, H., Zhou, X., Depeursinge, A., Pitkanen, M., Iavindrasana, J., Geissbuhler, A.: Medical Visual Information Retrieval: State of the Art and Challenges Ahead. In: Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, ICME'07, pp. 683–686. IEEE (2007)
35. Nehemiah, H.K., Khanna, A., Kumar, D.S.: Intelligent fractured image retrieval from medical image databases. Asian Journal of Information Technology **5**, 448–453 (2006)
36. Quellec, G., Lamard, M., Cazuguel, G., Roux, C., Cochener, B.: Case retrieval in medical databases by fusing heterogeneous information. IEEE Transactions on Medical Imaging **30**(1), 108–118 (2011)
37. Rosset, A., Müller, H., Martins, M., Dfouni, N., Vallée, J.P., Ratib, O.: Casimage project — a digital teaching files authoring environment. Journal of Thoracic Imaging **19**(2), 1–6 (2004)
38. Savoy, J.: Report on CLEF–2001 experiments. In: Report on the CLEF Conference 2001 (Cross Language Evaluation Forum), pp. 27–43. Springer LNCS 2406, Darmstadt, Germany (2002)
39. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content–based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12), 1349–1380 (2000)
40. Squire, D.M., Müller, W., Müller, H., Raki, J.: content–based query of image databases, inspirations from text retrieval: inverted files, frequency–based weights and relevance feedback. Tech. Rep. 98.04, Computer Vision Group, Computing Centre, University of Geneva, rue Général Dufour, 24, CH-1211 Genève, Switzerland (1998)
41. Stern, R., Hoffmeyer, P., Rosset, A., Garcia, J.: Fractures. University of Geneva, Geneva, Switzerland (2003)
42. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1), 11–32 (1991)
43. Tommasi, T., Orabona, F., Caputo, B.: CLEF2008 image annotation task: an SVM confidence–based approach. In: Working Notes of the 2008 CLEF Workshop. Aarhus, Denmark (2008)
44. Weber, B.: Die verletzungen des oberen sprungge-lenkes. Aktuelle Probleme in der Chirurgie., Stuttgart: Huber (1966)
45. Zhou, X., Depeursinge, A., Müller, H.: Information fusion for combining visual and textual image retrieval. In: International Conference on Pattern Recognition, ICPR'10. IEEE Computer Society, Los Alamitos, CA, USA (2010)
46. Zhou, X., Eggel, I., Müller, H.: The MedGIFT group at ImageCLEF 2009. In: Working Notes of CLEF 2009 (Cross Language Evaluation Forum). Corfu, Greece (2009)
47. Zhou, X., Gobeill, J., Müller, H.: The MedGIFT group at ImageCLEF 2008. In: CLEF 2008 Proceedings, *Lecture Notes in Computer Science (LNCS)*, vol. 5706, pp. 712–718. Springer, Aarhus, Denmark (2009)
48. Zhou, X., Gobeill, J., Ruch, P., Müller, H.: University and hospitals of geneva participating at imageclef 2007. In: CLEF 2007 Proceedings, *Lecture Notes in Computer Science (LNCS)*, vol. 5152, pp. 649–656. Springer, Budapest, Hungary (2008)
49. Zhou, X., Krabbenhöft, H., Niinimäki, M., Depeursinge, A., Möller, S., Müller, H.: An easy setup for parallel medical image processing: Using Taverna and ARC. In: Proceedings of HealthGrid 2009. Berlin, Germany (2009)