Knee X-Ray Image Analysis Method for Automated Detection of Osteoarthritis

Lior Shamir*, Shari M. Ling, William W. Scott, Jr., Angelo Bos, Nikita Orlov, *Member, IEEE,*
Tomasz J. Macura, *Student Member, IEEE*, D. Mark Eckley, Luigi Ferrucci, and Ilya G. Goldberg, *Member, IEEE*

Abstract—We describe a method for automated detection of radiographic osteoarthritis (OA) in knee X-ray images. The detection is based on the Kellgren–Lawrence (KL) classification grades, which correspond to the different stages of OA severity. The classifier was built using manually classified X-rays, representing the first four KL grades (*normal, doubtful, minimal*, and *moderate*). Image analysis is performed by first identifying a set of image content descriptors and image transforms that are informative for the detection of OA in the X-rays and assigning weights to these image features using Fisher scores. Then, a simple weighted nearest neighbor rule is used in order to predict the KL grade to which a given test X-ray sample belongs. The dataset used in the experiment contained 350 X-ray images classified manually by their KL grades. Experimental results show that moderate OA (KL grade 3) and minimal OA (KL grade 2) can be differentiated from normal cases with accuracy of 91.5% and 80.4%, respectively. Doubtful OA (KL grade 1) was detected automatically with a much lower accuracy of 57%. The source code developed and used in this study is available for free download at www.openmicroscopy.org.

Index Terms—Automated detection, image classification, Kellgren–Lawrence (KL) classification, osteoarthritis (OA), X-ray.

I. INTRODUCTION

OSTEOARTHRITIS (OA) is a highly prevalent chronic health condition that causes substantial disability in late life [31]. It is estimated that ∼80% of the population over the age of 65 have radiographic evidence of OA [29], and given the prolonged life expectancy in the United States and the aging of the "baby boomer" cohort, the prevalence of OA is expected to increase further.

Although newer methods, such as MRI, offer an assessment of periarticular as well as articular structures, the availability of plain radiographs makes them the most commonly used tools in the evaluation of OA joints [2], despite known limitations in detecting early disease and subtle changes over time. While several methods have been proposed [3], the Kellgren–Lawrence (KL) system [12], [13] is a validated method of classifying individual joints into one of five grades, with 0 representing normal and 4 being the most severe radiographic disease. This classification is based on features of *osteophytes* (bony growths adjacent to the joint space), *narrowing* of part or all of the tibial–femoral joint space, and *sclerosis* of the subchondral bone. Based on these three indicators, KL classification is considered more informative than any of the three elements individually.

Since the parameters used for OA classification are continuous, human experts may differ in their assessment of OA and therefore reach a different conclusion regarding the presence and severity of OA. This introduces a certain degree of subjectiveness to the diagnosis [10], [32] and requires a considerable amount of knowledge and experience for making a valid OA diagnosis.

Due to the high prevalence of OA, there is an emerging need for clinical and scientific tools that can reliably detect the presence and severity of OA. Boniatis *et al.* [5], [6] proposed a computer-aided method of grading hip OA based on textural and shape descriptors of radiographic hip joint space and showed 95.7% accuracy in detection of hip OA using a dataset of 64 hip X-rays (18 normal and 46 OA). Cherukuri *et al.* [9] described a convex-hull-based method of detecting anterior bone spurs (osteophytes) with accuracy of ∼90% using 714 lumbar spine X-ray images. Browne *et al.* [7] proposed a system that monitors for changes in finger joints based on a set of radiographs taken at different times, which can detect changes in the number and size of osteophytes, and Mengko *et al.* [22] developed an automated method for measuring joint space narrowing in OA knees.

However, despite the prevalence of knee OA, computer-based tools for OA detection based on single knee X-ray images are not yet available for either clinical or research purposes. Here, we describe a method for automated detection of OA by using computer-based image analysis of knee X-ray images. While at this point we do not suggest that the proposed method can completely replace a human reader, it can serve as a decision-supporting tool and can also be applied to the classification of large numbers of X-rays for clinical research trials. In Section II,

*L. Shamir is with the Image Informatics and Computational Biology Unit, Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health (NIH), Baltimore, MD 21224 USA (e-mail: shamirl@mail.nih.gov).

S. M. Ling is with the Clinical Research Branch, National Institute on Aging, Baltimore, MD 21225 USA, and also with the Division of Geriatric Medicine, Gerontology and Rheumatology, Johns Hopkins University School of Medicine, Baltimore, MD 21205 USA. He is also with the University of Maryland, College Park, MD 20742 USA.

W. W. Scott, Jr., is with the Arthritis Center, Johns Hopkins School of Medicine, Baltimore, MD 21205 USA.

A. Bos was with the Clinical Research Branch, National Institute on Aging, Baltimore, MD 21225 USA. He is now with the Institute of Geriatrics and Gerontology, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90690, Brazil.

N. Orlov, D. M. Eckley, and I. G. Goldberg are with the Image Informatics and Computational Biology Unit, Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health (NIH), Baltimore, MD 21224 USA.

T. J. Macura is with the Image Informatics and Computational Biology Unit, Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health (NIH), Baltimore, MD 21224 USA, and also with the Computer Laboratory, Trinity College, University of Cambridge, Cambridge CB2 1TQ, U.K.

L. Ferrucci is with the Clinical Research Branch, National Institute on Aging, Baltimore, MD 21225 USA.

TABLE I
X-RAY IMAGE DISTRIBUTION BY KL GRADE

KL grade	KL description	No. of images images
0	No osteophytes, normal joint space	154
1	Doubtful narrowing, possible osteophytes	102
2	Minimal but definite osteophytes, joint space	39
3	Definite and moderate osteophytes, joint space narrow, some subchondral sclerosis	55

we describe the data used for training and testing the proposed method. In Section III, we present the detection of the joint in the X-ray. In Section IV, we describe the automated classification of the knee X-rays. In Section V, the experimental results are discussed.

II. DATA

The data used for the experiment are consecutive knee X-ray images taken over a course of two years, as part of Baltimore Longitudinal Study of Aging (BLSA) [30], which is a longitudinal normative aging study. X-ray images were obtained in all participants, irrespective of symptoms or functional limitations, thereby providing an unbiased representation of knee X-rays in an aging sample.

The fixed-flexion knee X-rays were acquired with the beam angle at $10°$, focused on the popliteal fossa using a Siremobile Compact C-arm (Siemens Medical Solutions, Malvern, PA). Original images were 8-bit 1000×945 grayscale Digital Imaging and Communications in Medicine (DICOM) images, converted into tagged image file format (TIFF) format. Left knee images were flipped horizontally in order to avoid an unnecessary variance in the data.

Each knee image was independently assigned a KL grade (0–4) as described in the Atlas of Standard Radiographs [13] by two different readers, with discordant grades adjudicated by a third reader. In 79.8% of the cases, the two readers assigned the same KL grade, and the remaining images were adjudicated by a third reader.

The X-ray readers were radiologists with at least 25 years of reading experience who read from 50 to 100 musculoskeletal X-rays per day. To maximize comparability between readers, all readers received training using a set of "gold standard" X-rays. Each knee image was also assessed for osteophytes, joint space narrowing and sclerosis of the medial and lateral compartments, and tibial spine sharpening. The total number of knee X-ray images used was 350, divided into four KL grades as described in Table I.

In the proposed classifier, each KL grade is considered a class, so that a complete automated KL grade detection is a four-way classifier. KL grades 4 (severe OA) and 5 (knee replaced) remained outside the scope of this study due to the severe symptoms of pain that accompany these grades of OA, making a computer-based detection less effective at KL grade 4, and irrelevant at KL grade 5.

Fig. 1 shows four knee X-rays of KL grades 0 (normal), 1 (doubtful), 2 (minimal), and 3 (moderate). As can be seen in the
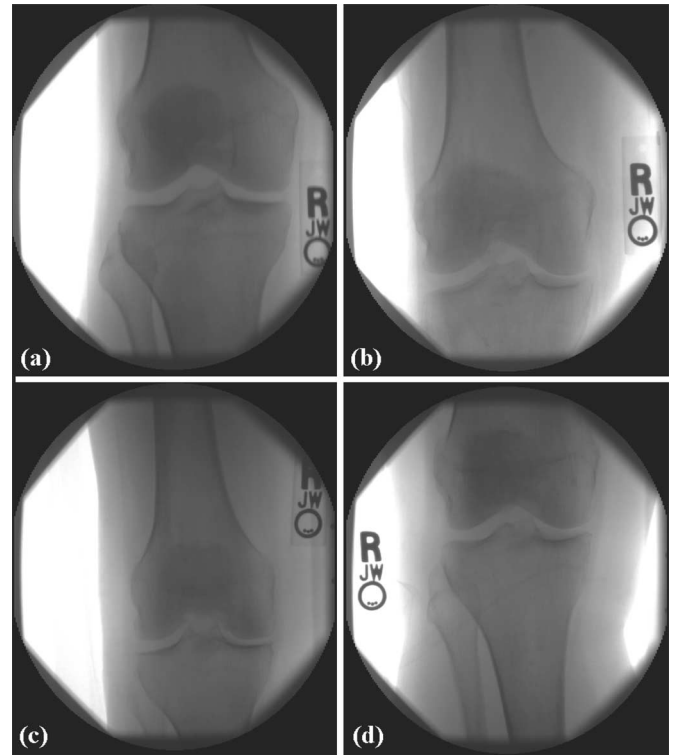

Fig. 1.   X-ray images of four different KL grades. (a) 0 (normal). (b) 1 (doubtful). (c) 2 (minimal). (d) 3 (moderate).

figure, most parts of an X-ray image are background or irrelevant parts of the bones, while only the area around the joint contains useful information for the purpose of OA detection. The X-rays also contain some metadata in the form of the letters *R* and *L*. These letters are the effect of small letter-shaped metal plates that are placed near the knee when the X-ray is taken, preventing any chance of confusion between the left and right knees.

The samples of Fig. 1 demonstrate how the different KL grades may look fairly similar to the untrained eye. Even experienced and well-trained experts can find it difficult to assign the KL grade and assess the OA severity, and a second (and sometimes also a third) analysis is often required to obtain a reliable and accurate diagnosis.

Despite providing a valuable approximation of the status of the knee, the manual KL classification cannot be considered a gold standard for the actual OA severity. The reason is that KL classification is based on a set of radiography visual elements that have been observed to be correlated with OA, but no claim has been made that this set of elements is complete. Therefore, one can reasonably assume that the OA can be detected by more elements that have not yet been isolated and well characterized, or that their signal is too weak to be sensed by an unaided eye. The incompleteness of the set of KL elements can be evident by the partial correlation between pain symptoms and the KL classification.

Another important feature of the data is that the progress of OA is continuous, while KL grades are discrete. That is, a knee X-ray classified as KL grade 1 can actually be somewhere between grade 0 and 1, but since KL grades are discrete
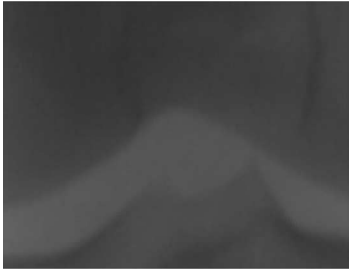
Fig. 2. Window of the joint center used for finding the center of the joint in the X-rays.

Fig. 3. Joint area of a knee X-ray.

classes, this in-between case will be classified as pure KL grade 1. This transition from a continuous variable to a discrete variable can be affected by the many in-between cases, resulting in weaker ground truth for both building and testing the image classifier.

III. JOINT DETECTION

Despite the attempts of making the X-rays as consistent as possible, the nature of working with human patients, especially at elderly ages, makes it practically impossible to repeat the procedure in the exact same fashion due to the differences in the flexibility of the patients and their ability to place their knee in the right position, as well as their ability to sustain these poses until the X-rays are taken. As a result, the position of the joint within the X-ray image can vary significantly.

Since in each X-ray image the joint can appear at different image coordinates, a joint detection algorithm is required to find the joint and separate it from the rest of the image. This is done by using a fixed set of 20 preselected images, such that each image is a 150×150 window of a center of a joint. For example, Fig. 2 is the center of the joint of the X-ray of panel (a) in Fig. 1. These images are then downscaled by a factor of 10 into 15×15 images.

Finding the joint in a given knee X-ray image is performed by first downscaling the image by a factor 10, and then scanning the image with a 15×15 shifted window. For each position, the Euclidean distances between the 15×15 pixels of the shifted window and each of the 20 15×15 predefined joint images are computed using

$$d_{i,w} = \sqrt{\sum_{y=1}^{15} \sum_{x=1}^{15} (I_{x,y} - W_{x,y})^2} \qquad (1)$$

where $W_{x,y}$ is the intensity of pixel x, y in the shifted window W, $I_{x,y}$ is the intensity of pixel x, y in the joint image I, and $d_{i,w}$ is the Euclidean distance between the joint image I and the 15×15 shifted window W.

Since the proposed implementation uses 20 joint images, 20 different distances are computed for each possible position of the shifted window, but only the shortest of the 20 distances is recorded.

After scanning the entire $(\text{width}/10 - 15) \times (\text{height}/10 - 15)$ possible positions, the window that recorded the smallest Euclidean distance is determined as the center of the joint,

and the 250×200 pixels around this center form an image that is used for the automated analysis. Fig. 3 shows an example of the 250×200 joint area. Since each image contains exactly one joint, and since the rotational variance of the knees is fairly minimal, this simple and fast method was able to successfully find the joint center in all images in the dataset.

Using these smaller images for the automated classification keeps them clean from the various background features and makes the images invariant to the position of the joint within the original X-ray. No attempt is made to fix for the rotational variance of the images, as this small variance is not expected to affect the image analysis described in Section IV and does not justify automated correction, which can introduce some inaccuracies resulting from the nontrivial task of estimating the exact angle of the joint.

IV. IMAGE CLASSIFICATION

Modern radiography instruments are many times more sensitive than the human eye that observes the resulting images. Therefore, it can be reasonably assumed that OA can be detected by more elements than those proposed by Kellgren and Lawrence, but are not used for the classification due to inability of the human eye to sense them. Since computers are substantially more sensitive to these small intensity variations, an effective use of the strength of computer-based detection would be based on a data-driven classification, rather than an attempt to follow each of the manually classified elements used by Kellgren and Lawrence.

The method works by first extracting a large set of image features, from which the most informative features are then selected. Since image transforms can often provide additional information that is difficult to deduce when analyzing the raw pixels [11], image features are computed not only on the raw pixels, but also on several transforms of the image, and also on transforms of transforms. These image content descriptors

extracted from transforms and compound transforms have been found highly effective in classification and similarity measurement of biological and biometric image datasets [11], [25], [26], [28].

For image feature extraction, we use the following algorithms, described more thoroughly in [25].

1) *Zernike features* [35] are the absolute values of the coefficients of the Zernike polynomial approximation of the image as described in [24], providing 72 image content descriptors.
2) *Multiscale histograms* computed using various numbers of bins (3, 5, 7, and 9), as proposed by [17], providing $3 + 5 + 7 + 9 = 24$ image content descriptors.
3) *First four moments* of mean, standard deviation, skewness, and kurtosis computed on image "stripes" in four different directions ($0°$, $45°$, $90°$, $135°$). Each set of stripes is then sampled into a three-bin histogram, providing $4 \times 4 \times 3 = 48$ image descriptors.
4) *Tamura texture features* [34] of *contrast*, *directionality*, and *coarseness*, such that the coarseness descriptors are its sum and three-bin histogram, providing $1 + 1 + 1 + 3 = 6$ image features.
5) *Haralick features* [18] computed on the image's cooccurrence matrix as described in [24] and contribute 28 image descriptor values.
6) *Chebyshev statistics* [16]—A 32-bin histogram of a 1×400 vector produced by Chebyshev transform of the image with order of $N = 20$.

The image transforms used by the proposed method are the commonly used wavelet (Symlet 5, level 1) transform, Fourier transform, and Chebyshev transform. Additionally, three compound transforms are also used, which are Chebyshev transform followed by Fourier transform, wavelet transform followed by Fourier transform, and Fourier transform followed by Chebyshev transform. Since human intuition and analysis of image features extracted from transforms and compound transforms are limited, the compound transforms were selected empirically by testing all two-level permutations of the three basic transforms (wavelet, Fourier, and Chebyshev) and assessing the contribution of the resulting content descriptors to the accuracy of the classification.

Extracting a set of 210 image content descriptors from seven different image transforms (including the raw pixels) results in a feature vector of dimensionality of 1470. However, not all image features are equally informative, and some of these features are expected to represent noise. In order to select the most informative image features while rejecting the noisy features, each image content descriptor is assigned with a Fisher score [4], described as

$$W_f = \frac{1}{N} \sum_{c=1}^{N} \frac{(\overline{T_f} - \overline{T_{f,c}})^2}{\sigma_{f,c}^2} \quad (2)$$

where W_f is the Fisher score of feature f, N is the total number of classes, $\overline{T_f}$ is the mean of the values of feature f among the images allocated for training, and $\overline{T_{f,c}}$ and $\sigma_{f,c}^2$ are the mean and variance of the values of feature f among all training images

of class c. The Fisher score can be conceptualized as the ratio of variance of class means from the pooled mean to the mean of within-class variances. All variances used in the equation are computed after the values of feature f are normalized to the interval $[0, 1]$.

Fisher score values rank the image features by their informativeness and are assigned to all 1470 features computed by extracting the 210 image content descriptors from the seven different image transforms (including the original image). Once each of the 1470 image features is assigned with a Fisher score, the weakest 90% of the features (with the lowest Fisher scores) are rejected, resulting in a feature space of 147 image content descriptors. As discussed in Section V, this setting provided the best performance in terms of classification accuracy. The distribution of the different types of image features and image transforms is described in Table II.

As the table shows, the most informative image content descriptors are the Zernike polynomials and both Haralick and Tamura texture features. The radial nature of Zernike polynomials allows these features to reflect variations in the unit disk, so that these features are expected to be sensitive to the joint space in the X-ray images. As observed and thoroughly discussed by Boniatis *et al.* [5], the pixel intensity variation patterns mathematically described by the texture features correlate with biochemical, biomechanical, and structural alterations of the articular cartilage and the subchondral bone tissues [1], [23]. These processes have been associated with cartilage degeneration in OA [8], [27] and are therefore expected to affect the joint tissues in a fashion that can be sensed by the radiographic texture.

Since not all radiographic elements of OA have been isolated and well characterized, not all discriminative image features are expected to correspond to a known OA element. Therefore, a small portion of the discriminative values such as the statistical distribution of the pixel values of a Fourier transform followed by a Chebyshev transform is difficult to associate with a known radiographic OA element.

Additional algorithms for extracting image features have been tested but were found to be less informative for the classification of the KL grades. These image content descriptors include Radon transform features [20], which were expected to correlate with joint space, but were outperformed by Zernike polynomials. Gabor filters [14] capture textural information that could also be useful for the analysis but were found to be less informative than the Haralick and Tamura textures. Since the areas of interest have relatively low intensity variations, high contrast features such as edge and object statistics as described in [26] do not provide useful information.

After computing all feature values of a given test image, the resulting feature vector is classified using a simple weighted nearest neighbor rule, which is one of the most effective routines for nonparametric classification [4], [36]. The weights, in this case, are the Fisher scores computed by (2). The result of this classification can be a KL grade class determined by the training sample with the shortest distance to the given test sample but can also be an interpolated value based on the two nearest training

TABLE II
DISTRIBUTION OF TYPES OF FEATURES BY ALGORITHMS AND IMAGE TRANSFORMS

Transform	Zernike polynomials	Haralick textures	Tamura textures	First four moments	Multi-scale histogram	Chebyshev statistics	Total
Raw pixels	10	8	1	0	0	6	25
Wavelet	22	1	4	3	0	0	30
Chebyshev	26	0	2	1	2	4	35
Fourier	0	2	2	2	0	0	6
Fourier + Chebyshev	0	10	0	5	7	0	22
Fourier + Wavelet	9	0	1	0	0	4	14
Chebyshev + Fourier	0	10	0	4	1	0	15
Total	67	31	10	15	10	14	147

TABLE III
CONFUSION MATRIX OF MODERATE OA AND NORMAL

	Normal (KL grade 0)	Moderate OA (KL grade 3)
Normal (KL grade 0)	613	87
Moderate OA (KL grade 3)	32	668

TABLE IV
CONFUSION MATRIX OF MINIMAL OA AND NORMAL

	Normal (KL grade 0)	Minimal OA (KL grade 2)
Normal (KL grade 0)	222	58
Minimal OA (KL grade 2)	52	228

samples that do not belong to the same class, described as

$$\text{KL} = \frac{(K_1/d_1 + K_2/d_2)}{(1/d_1 + 1/d_2)} \qquad (3)$$

where KL is the resulting interpolated KL grade, and d_1 and d_2 are the distances from the nearest two samples that belong to different KL grades K_1 and K_2, respectively.

The main advantage of this interpolation is that it can potentially provide a higher resolution estimation by determining the OA severity within the grade (e.g., KL grade 1.6 for a case of OA severity between KL grade 0 and 1). This, however, is more difficult to evaluate for accuracy since the data used as ground truth only specify the OA severity in resolution of KL grades.

V. EXPERIMENTAL RESULTS

The proposed method was tested using the dataset described in Section II, where the ground truth was the manual classification of the X-rays. In the first experiment, the proposed method was tested by automatically classifying moderate OA (KL grade 3) and normal knees (KL grade 0). The experiment was performed by using 55 X-ray images from each grade, such that 20 images from each grade were used for training and 35 for testing.

This test was repeated 20 times, where each run used a different random split of the training and test images. Since the dataset has more grade 0 images than grade 3, each run used a different set of 55 grade 0 images, randomly selected from the dataset. The classification accuracy was 91.5% ($P < 0.000001$), as can be learned from the confusion matrix of Table III.

As the confusion matrix shows, the specificity of *moderate* OA detection is ∼86.5% and the sensitivity is 95%. This can be important for a potential practical use of the described classifier, since it may be overprotective in some cases, but is less likely to dismiss a positive X-ray as *normal*. It should be noted that while the human readers had a consensus rate of ∼80%, the disagreements were usually between neighboring KL grades,

and none of the knees was classified by one reader as *normal* and by the other reader as *moderate*. Therefore, experienced and knowledgeable readers should be able to tell *moderate* OA from normal knees with accuracy of practically 100%.

A similar experiment tested the classification accuracy of KL grade 2 (*minimal* OA). Since this classification problem is more difficult, five more training images were used for each class so that the training set consisted of 25 images of grade 0 and 25 images of grade 2. The use of a larger training set improved the classification accuracy, while increasing the training set in the grade 3 detection did not contribute to the performance, as will be discussed later in this section. The classification accuracy was ∼80.4% ($P < 0.0001$), and the confusion matrix is given in Table IV.

As can be learned from the table, the specificity of the detection is ∼79.3%, with sensitivity of ∼81.4%. These numbers show that the detection of KL grade 2 is less effective than the automated detection of KL grade 3. This can be explained by the fact that the progress of OA is continuous, so that KL grade 2 is expected to be visually more similar to 0 than KL grade 3.

We also tried to classify KL grade 1 (*doubtful* OA) from 0, which provided classification accuracy of 54% when using 25 images per class for training and 14 images for testing. Increasing the size of the training set to 70 samples per class (and 32 samples for testing) marginally increased the performance to 57%. This can be explained by the fact that these two grades are visually very similar, and even experienced human readers often have to struggle to differentiate between the two. Also, since the KL grades are discrete while the actual OA progress is continuous, the very many in-between cases can significantly contribute to the confusion.

Table V shows the classification accuracy when classifying between any pair of KL grades. In all cases, 25 images were used for training, 14 images for testing, and the classification accuracy was computed by averaging the accuracy of 20 random splits to sets of training and test images. Maximum differences

TABLE V
TWO-WAY CLASSIFICATION ACCURACY (IN PERCENT) OF ALL
PAIRS OF KL GRADES

	KL grade 1	KL grade 2	KL grade 3
KL grade 0	54 ± 19	80 ± 8	91 ± 4
KL grade 1	-	60 ± 13	82 ± 7
KL grade 2	-	-	65 ± 10

TABLE VI
CONFUSION MATRIX OF A FOUR-WAY CLASSIFIER OF ALL KL GRADES

	Grade 0	Grade 1	Grade 2	Grade 3
Grade 0	144	96	21	19
Grade 1	68	60	45	31
Grade 2	10	45	118	107
Grade 3	12	36	105	127

Fig. 4. Classification accuracy (%) of KL grades 2 and 3 as a function of the size of the training set.

from the mean (among the 20 runs) are also specified in the table.

As the table shows, classification between two neighboring KL grades is less accurate than classification of nonneighboring grades. We can also see that the two neighboring KL grades 2 and 3 can be differentiated with accuracy of 65%, while the classification of grades 1, 2 and grades 0, 1 provide accuracy of 60% and 54%, respectively. This may indicate that neighboring KL grades are visually more similar to each other in the early stages of OA. It is also noticeable that the classification accuracy improves as the difference between the KL grades gets larger. For example, the classification accuracy of KL grades 3 and 0 is better than the classification accuracy of KL grades 3 and 1. These results are in agreement with the continuous nature of OA.

Testing the classification accuracy when classifying all four KL grades was performed by randomly selecting 25 images from each class for training and then selecting 14 of the remaining images of each class for testing. This experiment was repeated 20 times, such that in each run, the training and test sets were determined in a random manner. The overall classification accuracy of this classifier was 47%, as described by the confusion matrix of Table VI.

Classification accuracy of 47% may not be considered strong, and it is significantly lower than the 79.8% of agreement between the two human experts. However, the confusion matrix shows that from all cases of KL grade 3, only ~4% were classified as non-OA (KL grade 0) and ~13% as doubtful (KL grade 1). Examining the false positives, the confusion matrix shows that ~7% of the X-rays classified manually as KL grade 0 were falsely classified by the proposed method as KL grade 3 (moderate), and also ~11% of KL grade 1 were mistakenly classified as KL grade 3.

Since the actual OA progress is a continuous variable, there are many in-between cases that can be classified to either of the two closest classes of the given case. For example, OA progress between KL grade 1 and KL grade 2 can be classified to either 1 and 2. As described in the end of Section IV, the resulting value of the image classification can also be an interpolation of the two nearest KL grades, rather than a single KL grade class. The

efficacy of this interpolation can be demonstrated by computing the Pearson correlation coefficient between the predicted and actual KL grades (when using all four KL grades as described in Table VI). When using the interpolated value as the predicted grade, Pearson correlation is 0.73, compared to 0.49 when the predicted KL grade is simply the class of the nearest training sample.

A knee is classified as OA positive if it has been classified as KL grade 2 or higher. By merging the images of KL grades 2 and 3, we introduced a new class called *OA positive* that contained 78 images (39 from each class). An experiment was performed by building a two-way classifier using this newly defined class and KL grade 0, such that 60 images from each class were used for training and 18 for testing. The purpose of this experiment was to classify OA positive X-rays from OA negatives. The classification accuracy of this classifier was ~86.1%, with sensitivity and specificity of ~88.7% and ~83.5%, respectively.

It is also important to mention that the X-rays were taken from randomly selected human subjects who participate in the BLSA study, and not necessarily from patients who reported on pain symptoms or were diagnosed as OA positive in one of their other joints. This policy provides a uniform representation of the elderly population, and therefore it can be assumed that the results presented here can be generalized to the entire elderly population.

The classification accuracy may be affected by the size of the training set such that it is expected to increase as the number of training samples gets higher. Fig. 4 shows the classification accuracy of the classification of KL grades 2 and 3 from KL grade 0 as a function of the size of the training set. As can be learned from the graph, for KL grade 3 the classification accuracy improves as the size of the training set gets larger, but stabilizes at ~14 training images per class. Classification accuracy of KL grade 2 also increases with the number of training images, but our dataset is not large enough to determine whether it reached the peak of accuracy.

As described in Section IV, 10% of the image features with the highest Fisher scores were used for the classification, while
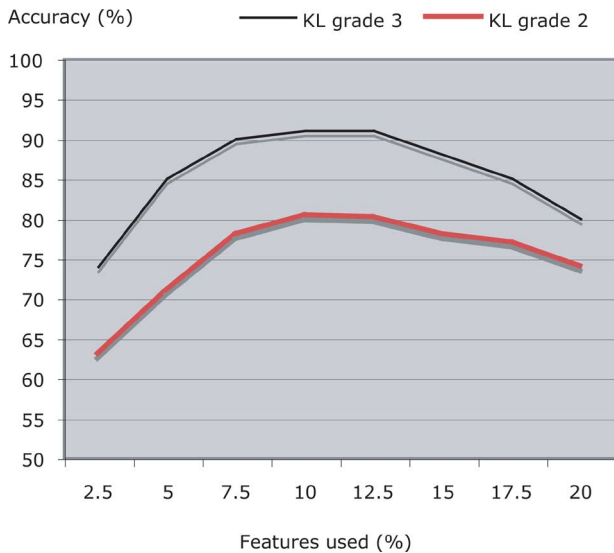
Fig. 5. Classification accuracy (%) of KL grades 2 and 3 as a function of the number of image features used.

the rest of the image content descriptors were ignored. Fig. 5 shows how the number of features used for the classification affects the classification accuracy of KL grade 3. According to the graph, the best performance is achieved when using between 7.5% and 12.5% of the total number of image features (1470).

In terms of computational complexity, classifying one X-ray image takes 105 s, using a system with a 2 GHZ Intel processor and 1 MB of RAM. While the time required for the joint detection described in Section III is negligible, nearly all of the CPU time is sacrificed for transforming the image and extracting image features. Major contributors to this complexity are the 2-D Zernike features, which are known to be computationally expensive [19], [37]. This downside of the classifier makes it unsuitable for real-time applications or other tasks in which speed is a primary concern but may be fast enough for classification of single X-ray images, especially considering the fact that the entire procedure of taking X-rays can take a much longer time.

An improvement in the response time of the classifier can be achieved by parallelization of the feature extraction algorithms. Since most of the image features are not dependent on each other, the algorithm becomes trivial to parallelize so that many image features can be computed concurrently by different processors. For instance, while one processor can compute Haralick features on the Chebyshev transform, another processor can extract Haralick features from the Fourier transform, or Zernike features from the Chebyshev transform. In order to take full advantage of the parallelization, the algorithm has been also implemented using the Open Microscopy Environment (OME) software suite [15], [33], which is a platform for storing and processing microscopy images, designed to optimize the execution and data flow of multiple execution modules. A detailed description of the parallelization of the image transform and image feature extractions in OME can be found in [21].

VI. CONCLUSION

In this paper, we described an automated approach for the detection of OA using knee X-rays. In the absence of an accurate method of OA diagnosis, the manually classified KL grade is used here as a "gold standard," although it is known that this method is less than perfect. The classification is not performed in a way that attempts to imitate the human classification, but is based on a data-driven approach using manually classified X-rays of different KL grades, representing different stages of OA severity.

Experimental results suggest that more than 95% of moderate OA cases were differentiated accurately from normal cases, with a false positive rate of ∼12.5%. Classification accuracy of differentiating *minimal* OA from normal cases was ∼80%, and detection of *doubtful* OA cases was far less convincing. Future attempts of improving the detection accuracy of *doubtful* OA will include integration of relevant clinical information such as history of knee injury, body weight, and knee alignment angle, and will also use more X-ray samples as they become available. However, due to the subjective nature of the "gold standard," it is possible that a 100% correlation between computer-based and manual classifications may not be achieved. KL grades 4 (severe OA) and 5 (knee replaced) remained outside the scope of this study due to the severe symptoms that accompany these stages, and the relatively easy detection of OA in these stages.

While the classification accuracy of KL grades 1 and 2 cannot be considered strong, it is important to note that radiograph readers are often challenged in attempting to distinguish between these grades, and therefore the confusion of the automated detection between these two grades cannot be considered surprising.

We acknowledge that the equipment used to obtain the knee images does not allow for maximal resolution of joint structures. We speculate conventional radiographic images would predictably allow for better delineation of OA grades, particularly in cases with less severe disease. Certainly, conventional radiography is capable of delineating joint structures more readily, but achieves this at the cost of greater radiation exposure. The application of this imaging software to the interpretation of X-ray images occurs without the bias inherent in a clinical interpretation. Finally, this study was conducted within the context of a longitudinal aging study that will enable the comparison of imaging data not only to clinical OA features as pain, but also to physiologic measures relevant to aging body systems that might contribute to OA severity.

Future plans include testing of this automated technique in the evaluation of longitudinal knee images obtained over time, to check whether OA can be detected before radiographic evidence are noticeable by a human reader. We also plan to develop similar techniques to classify hand X-rays with attention to the signal joints predisposed to the development of OA.

The full source code used for the experiment is available for free download under standard GNU general public license (GPL) via concurrent versions system (CVS) at www.openmicroscopy.org. Scientists and engineers are encouraged to download, compile, and use this code for their needs.

REFERENCES

[1] T. Aigner and L. McKenna, "Molecular pathology and pathobiology of osteoarthritic cartilage," *Cell. Mol. Life Sci.*, vol. 59, pp. 5–18, 2002.

[2] R. D. Altman and G. E. Gold, "Atlas of individual radiographic features in osteoarthritis, revised," *Osteoarthritis Cartilage*, vol. 15, pp. A1–A56, 2007.

[3] S. Ahlback, "Osteoarthrosis of the knee: A radiographic investigation," *Acta Radiol. Suppl.*, vol. 277, pp. 7–72, 1968.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[5] I. Boniatis, L. Costaridou, D. Cavouras, I. Kalatzis, E. Panagiotopoulos, and G. Panayiotakis, "Osteoarthritis severity of the hip by computer-aided grading of radiographic images," *Med. Biol. Eng. Comput.*, vol. 44, pp. 793–803, 2006.

[6] I. Boniatis, D. Cavouras, L. Costaridou, I. Kalatzis, E. Panagiotopoulos, and G. Panayiotakis, "Computer-aided grading and quantification of hip osteoarthritis severity employing shape descriptors of radiographic hip joint space," *Comput. Biol. Med.*, vol. 37, pp. 1786–1795, 2007.

[7] M. A. Browne, P. A. Gaydeckit, R. F. Goughll, D. M. Grennanz, S. I. Khalilt, and H. Mamtoras, "Radiographic, image analysis in the study of bone morphology," *Clin. Phys. Physiol. Meas.*, vol. 8, pp. 105–121, 1987.

[8] A. Buckwalter and H. J. Mankin, "Instructional course lectures, the American Academy of Orthopaedic Surgeons Ñ Articular Cartilage. Part II: Degeneration and osteoarthrosis, repair, regeneration, and transplantation," *J. Bone Joint Surg.*, vol. 79, pp. 612–632, 1997.

[9] M. Cherukuri, R. J. Stanley, R. Long, S. Antani, and G. Thoma, "Anterior osteophyte discrimination in lumbar vertebrae using size-invariant features," *Comput. Med. Imag. Graph.*, vol. 28, pp. 99–108, 2004.

[10] P. Croft, "An introduction to the atlas of standard radiographs of arthritis," *Rheumatol. Suppl.*, vol. 4, p. 42, 2005.

[11] I. B. Gurevich and I. V. Koryabkina, "Comparative analysis and classification of features for image models," *Pattern Recognit. Image Anal.*, vol. 16, pp. 265–297, 2006.

[12] J. H. Kellgren and J. S. Lawrence, "Radiologic assessment of osteoarthritis," *Ann. Rheum. Dis.*, vol. 16, pp. 494–501, 1957.

[13] J. H. Kellgren, M. Jeffrey, and J. Ball, *Atlas of Standard Radiographs*. Oxford, U.K.: Blackwell, 1963.

[14] D. Gabor, "Theory of communication," *J. Inst. Electr. Eng.*, vol. 93, no. 26, pp. 429–457, Nov. 1946.

[15] I. G. Goldberg, C. Allan, J. M. Burel, D. Creager, A. Falconi, H. Hochheiser, J. Johnston, J. Mellen, P. K. Sorger, and J. R. Swedlow, "The Open Microscopy Environment (OME) data model and XML file: Open tools for informatics and quantitative analysis in biological imaging," *Genome Biol.*, vol. 6, p. R47, 2005.

[16] I. Gradshtein and I. Ryzhik, *Table of Integrals, Series and Products*, 5th ed. New York: Academic, 1994.

[17] E. Hadjidementriou, M. Grossberg, and S. Nayar, "Spatial information in multiresolution histograms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 702–709.

[18] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 269–285, Nov. 1973.

[19] L. Kotoulas and I. Andreadis, "Real-time computation of Zernike moments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 6, pp. 801–809, Jun. 2005.

[20] J. S. Lim, *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1990, pp. 42–45.

[21] T. J. Macura, L. Shamir, J. Johnston, D. Creager, H. Hocheiser, N. Orlov, P. K. Sorger, and I. G. Goldberg, "Open microscopy environment analysis system: End-to-end software for high content, high throughput imaging," *Genome Biol.*, to be published.

[22] T. L. Mengko, R. G. Wachjudi, A. B. Suksmono, and D. Danudirdjo, "Automated detection of unimpaired joint space for knee osteoarthritis assessment," in *Proc. 7th Int. Workshop Enterprise Netw. Comput. Healthcare Ind.*, 2005, pp. 400–403.

[23] J. Martel-Pelletier and J. P. Pelletier, "Osteoarthritis: Recent developments," *Curr. Opin. Rheumatol.*, vol. 15, pp. 613–615, 2003.

[24] R. F. Murphy, M. Velliste, J. Yao, and G. Porreca, "Searching online journals for fluorescence microscopy images depicting protein subcellular location patterns," in *Proc. 2nd IEEE Int. Symp. Bioinf. Biomed. Eng.*, 2001, pp. 119–128.

[25] N. Orlov, J. Johnston, T. Macura, L. Shamir, and I. Goldberg, "Computer vision for microscopy applications," in *Vision Systems*, G. Obinata and A. Dutta, Eds. Vienna, Austria: Adv. Robot. Syst., 2007, pp. 221–242.

[26] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckely, and I. Goldberg, "WND-CHARM: Multi-purpose image classification using compound image transforms," *Pattern Recognit. Lett.*, vol. 29, pp. 1684–1693, 2008.

[27] E. L. Radin and R. M. Rose, "Role of subchondral bone in the initiation and progression of cartilage damage," *Clin. Orthop.*, vol. 213, pp. 34–40, 1986.

[28] K. Rodenacker and E. Bengtsson, "A feature set for cytometry on digitized microscopic images," *Anal. Cell. Pathol.*, vol. 25, pp. 1–36, 2006.

[29] S. F. St. Clair, C. Higuera, V. Krebs, N. A. Tadross, J. Dumpe, and W. K. Barsoum, "Hip and knee arthroplasty in the geriatric population," *Clin. Geriatr. Med.*, vol. 3, pp. 515–533, 2006.

[30] N. W. Shock, R. C. Greulich, R. Andres, D. Arenberg, P. T. Costa, E. G. Lakatta, and J. D. Tobin, *Normal Human Aging: The Baltimore Longitudinal Study of Aging, NIH Publication No. 84-2450*. Washington, DC: GPO, 1984.

[31] R. A. Stockwell, "Cartilage failure in osteoarthritis: Relevance of normal structure and function—A review," *Clin. Anat.*, vol. 4, pp. 161–191, 1990.

[32] Y. Sun, K. P. Gunther, and H. Brenner, "Reliability of radiographic grading of osteoarthritis of the hip and knee," *Scand. J. Rheumatol.*, vol. 26, pp. 155–165, 1997.

[33] J. R. Swedlow, I. Goldberg, E. Brauner, and P. K. Sorger, "Informatics and quantitative analysis in biological imaging," *Science*, vol. 300, pp. 100–102, 2003.

[34] H. Tamura, S. Mori, and T. Yamavaki, "Textural features corresponding to visual perception," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 6, pp. 460–472, Jun. 1978.

[35] M. R. Teague, "Image analysis via the general theory of moments," *J. Opt. Soc. Amer.*, vol. 70, pp. 920–920, 1979.

[36] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2003.

[37] C. Y. Wee and P. Paramesran, "Efficient computation of radial moment functions using symmetrical property," *Pattern Recognit.*, vol. 39, pp. 2036–2046, 2006.

Lior Shamir received the Bachelor's degree in computer science from Tel Aviv University, Tel Aviv, Israel and the Ph.D. degree from Michigan Technology.

He is a Postdoctoral Fellow in the Image Informatics and Computational Biology Unit, Laboratory of Genetics, National Institute on Aging, National Institutes of Health, Baltimore, MD.

Shari M. Ling received the Bachelor's degree from the University of Santa Clara, CA, the Master's degree in direct service from the Ethel Percy Andrus School of Gerontology, University of Southern California, and the M.D. degree from Georgetown University, D.C.

She is currently a Staff Clinician in the Clinical Research Branch, Intramural Research Program, National Institute on Aging, Baltimore, MD. She is also with the Division of Geriatric Medicine, Gerontology and Rheumatology, Johns Hopkins University School of Medicine, Baltimore. She is on the clinical faculty at the University of Maryland, College Park. Her current research interests include aging, osteoarthritis, and the identification and development of novel diagnostic and prognostic indicators of arthritis development in the elderly.

William W. Scott, Jr., received the B.A. degree in physical sciences (*magna cum laude*) from Harvard University, Cambridge, MA, in 1967, and the M.D. degree from Johns Hopkins University School of Medicine, Baltimore, MD, in 1971.

He did radiology residency training at Johns Hopkins University School of Medicine, where he was engaged in radiology, board certified in diagnostic radiology in 1975, formerly in charge of orthopaedic radiology, and is currently an Associate Professor of radiology. He was at the United States Army Medical Corps (USAMC) for two years. He has been one of the Radiographic Interpreters for Baltimore Longitudinal Study of Aging (BLSA) for many years.

D. Mark Eckley received the B.A. degree majoring in biochemistry from the University of Colorado, Boulder, in 1984, and the M.A. degree in molecular biology from the University of California, Santa Barbara, in 1989.

He was a Technician at the University of Texas Medical Branch, Galveston, where he was engaged in recombinant DNA techniques. He was engaged in molecular biology at the University of California, Santa Barbara, during 1989. He was also involved in training in cell biology through the Biochemistry, Cellular and Molecular Biology (BCMB) Program, Johns Hopkins School of Medicine. He is currently a Staff Scientist in the Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health (NIH), Baltimore, MD, where he focuses on microscopy as a genetic screening tool.

Angelo Bos received the M.D. degree from the Fundacao Faculdade Federal de Ciencias Medicas of Porto Alegre, Porto Alegre, Brazil, in 1983, and the Ph.D degree in medicine from Tokai University, Kanagawa-ken, Japan.

Until 2008, he was a Senior Statistician for the Longitudinal Studies Section, National Institute on Aging. He is currently an Associate Professor at the Institute of Geriatrics and Gerontology, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre.

Luigi Ferrucci received the M.D. and Ph.D. degrees in biology and pathophysiology of aging from the University of Florence, Florence, Italy.

He is the Director of the Baltimore Longitudinal Study of Aging, National Institute on Aging, National Institutes of Health, Baltimore, MD, where he is also a Senior Investigator at the Clinical Research Branch. His current research interests include frailty and mobility disability in the elderly.

Nikita Orlov (M'06) received the Ph.D. degree in applied mathematics from Moscow State University, Moscow, Russia, in 1989.

He was a Researcher and a Senior Researcher in the Computational Electromagnetics Laboratory, Moscow University, and then with ADE Optical Systems, Charlotte, NC, and ADE Corporation, Norwood, MA. In 2003, he joined the Group of Ilya Goldberg, Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health, Baltimore, MD, where he is currently a Senior Research Fellow. His current research interests include pattern recognition, machine vision, and biological applications of numerical modeling.

Ilya G. Goldberg (M'06) received the B.S. degree in biochemistry from the University of Wisconsin, Madison, in 1990, and the Ph.D. degree in biochemistry and cell biology from Johns Hopkins University School of Medicine, Baltimore, MD, in 1997.

He was a Postdoctoral Fellow of crystallography at Harvard. He started the Open Microscopy Environment (OME) Project as a Postdoctoral Researcher at Massachusetts Institute of Technology (MIT). In 2002, joined the Laboratory of Genetics, National Institute on Aging, National Institutes of Health, Baltimore, MD, where he started a group dedicated to quantitative morphology and systematic functional genomics.

Tomasz J. Macura (S'06) received the B.S. degrees (*magna cum laude*) in mathematics and computer science from the University of Maryland, Baltimore. He is currently working toward the Doctoral degree at the Computer Laboratory, Trinity College, University of Cambridge, Cambridge, U.K.

He is a Predoctoral Research Trainee at Dr. Ilya Goldberg's Image Informatics and Computational Biology Unit (IICBU), Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health (NIH), Baltimore, MD. He is supported by a National Institutes of Health–University of Cambridge Health Science Scholarship.