Expert Systems With Applications 89 (2017) 41–51

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

myStone: A system for automatic kidney stone classification

Joan Serrat[a,*], Felipe Lumbreras[a], Francisco Blanco[b], Manuel Valiente[b], Montserrat López-Mesas[b]

[a] Computer Vision Center and Department of Computer Science, Universitat Autònoma de Barcelona, Edifici O, 08193 Bellaterra, Spain
[b] Centre Grup de Tècniques de Separació en Química, Unitat de Química Analítica, Departament de Química, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain

ARTICLE INFO

Article history:
Received 5 April 2017
Revised 14 July 2017
Accepted 15 July 2017
Available online 17 July 2017

Keywords:
Kidney stone
Optical device
Computer vision
Image classification

ABSTRACT

Kidney stone formation is a common disease and the incidence rate is constantly increasing worldwide. It has been shown that the classification of kidney stones can lead to an important reduction of the recurrence rate. The classification of kidney stones by human experts on the basis of certain visual color and texture features is one of the most employed techniques. However, the knowledge of how to analyze kidney stones is not widespread, and the experts learn only after being trained on a large number of samples of the different classes. In this paper we describe a new device specifically designed for capturing images of expelled kidney stones, and a method to learn and apply the experts knowledge with regard to their classification. We show that with off the shelf components, a carefully selected set of features and a state of the art classifier it is possible to automate this difficult task to a good degree. We report results on a collection of 454 kidney stones, achieving an overall accuracy of 63% for a set of eight classes covering almost all of the kidney stones taxonomy. Moreover, for more than 80% of samples the real class is the first or the second most probable class according to the system, being then the patient recommendations for the two top classes similar. This is the first attempt towards the automatic visual classification of kidney stones, and based on the current results we foresee better accuracies with the increase of the dataset size.

1. Introduction

Medicine and healthcare are among the most important fields where expert systems have found application (Wagner, 2017). For instance, computer aided detection and diagnosis systems can enhance the diagnostic capabilities of physicians or reduce the required time. Many of them are based on diverse imaging modalities like brain CT and MRIs, mammographies, chest X-ray and a long etc. Another type of expert systems are decision support systems, whose purpose is not as much to produce a diagnostic but to analyze data (e.g. images) and present some kind of result so that decisions can be made more easily. At the core of such systems one can often find classifiers, which are typically trained on data samples to generate a discrete prediction (a class label) plus a confidence score or a probability for each class, when presented a new sample. This is the case of the work described in this paper, which deals with the problem of kidney stone classification.

Urinary lithiasis –the formation of kidney stones– shows a steady incidence increase in developed countries. Around 10% of population in developed countries suffer a stone episode at least once in his/her life (Romero, Akpinar, & Assimos, 2010; Scales, Smith, Hanley, & Saigal, 2012). Emphasis should be made on the high prevalence affecting this disease. Some European follow-up studies have quantified the stone recurrence rate (repeated stone episodes for the same patient) at 40% in 5 years (Andreassen, Poulsen, Olsen, Aabeck, & Osther, 2007; Hesse, Brndle, Wilbert, Khrmann, & Alken, 2003). These dramatic numbers reflect not only a disturbing and painful disease but also a considerable burden for the national healthcare systems (Strohmaier, 2012).

Once the stone episode has passed, it is widely agreed that an adequate study of the causes of stone formation is required in order to decrease the high recurrence of this disease (Grases, Costa-Bauzá, Ramis, Montesinos, & Conte, 2002; Kok, 2012; Siener & Hesse, 2012). In fact, it has been pointed out that the correct treatment of stone patients can drop further stone formation as much as 46% (Nolde, Hesse, Scharrel, & Vahlensieck, 1993; Strohmaier, 2011). The urinary stone represents a solid description of the metabolic disturbances suffered by the patient, so it should be regarded as the starting point of an individualized treatment.

* Corresponding author.
E-mail addresses: joans@cvc.uab.es (J. Serrat), felipe@cvc.uab.es (F. Lumbreras), frnblanco@gmail.com (F. Blanco), Manuel.Valiente@uab.cat (M. Valiente), Montserrat.Lopez.Mesas@uab.cat (M. López-Mesas).

http://dx.doi.org/10.1016/j.eswa.2017.07.024
0957-4174/

- uric acid anhydrous and dihydrate (UA)
- mixed uric acid and calcium oxalate (UA + CO)
- cystine (CYS)

When collecting samples to build the dataset, we were able to gather just a few samples of cystine, one of the less frequent ones (around 1%). They were insufficient to properly train the classifier and hence we discarded it. In contrast, we realized that a few subclasses of COM and COD were relatively well populated (see Table 1), thus giving us the chance to provide more specific classification results and consequently better patient recommendations. Hence we decided to expand the first scheme of classes to a second scheme. Along the way, we are going to borrow the class notation of Grases et al. (2002) for which the classes of first scheme are denoted by numbers (2, 3 ...9) whereas the labels for the second are the former plus a suffix if necessary. Accordingly, COM was divided into:

- pure COM (2)
- COM with traces of hydroxiapatite and/or organic matter in the core (2b)
- COM with little amounts of COD in the core (2codt)

and COD was also split into:

- COD with presence of hydroxiapatite (3b)
- COD with COM, resulting from the transformation of COD, which is unstable, to COM (3t)
- COD, transformed to COM plus traces of hydroxiapatite (3bt)

The exhaustive visual description of each class is out of the scope of this paper. We refer the reader to the seminal paper (Daudon et al., 1993) and the more recent work (Cloutier et al., 2015), where they are listed and illustrated with examples. Nonetheless, we can get a glimpse of them in Figs. 1 and 2. They show 8 images of stone fragments for each main class, as acquired by our device. In each case, half the images are from the outer part of a fragment, half from an inner section. We can appreciate that some classes are quite similar in aspect (e.g. STR, HAP and BRU) and also the large intraclass variability common to all of them.

The first problem is a consequence of the lack of a clear frontier between classes because they may share the same components in different proportions. For instance, classes 2b, 3bt, CO - HAP and HAP form a continuum of calcium oxalate with increasing proportion of hydroxiapatite. The same phenomena occurs in the case of UA and UA + CO, and 2, 2codt and 3t. This makes kidney stone classification a tough problem for the human expert and obviously for the classification software.

3. Device

The device, specially developed for the analysis of kidney stones, assembles low-cost off-the-shelf components, being a simple, compact and robust piece of equipment. The final prototype consists of an enclosure that holds the camera, lens, focusing ring and a lighting board, plus a detached base acting as sample holder. Fig. 3 shows an schematic view. All the structural components have been 3d printed in ABS plastic. Dimensions are 70 × 70 × 85 mm^3, making it suitable to be placed on an desktop without disturbance. Fig. 4 shows a picture of the first batch of prototypes where we can appreciate the relative size of the system.

The camera is a conventional RGB CMOS 5 megapixel small camera, model daA2500–14uc from Basler.[1] Its sensor spectral sensitivity allows us to acquire color images both in the visible and near infrared ranges (below 1000 nm) as we will explain. To this end, we got removed the cut-off IR coating and filter from the lens and camera, respectively. Image resolution is 2592 × 1944 pixels/channel. The camera is connected to the host computer with an USB3 cable through which our software obtains the pictures, sets the acquisition parameters, like the exposure time, and also sends two output signals that control the lighting board. In order to obtain similar colors across the several device exemplars we calibrate the gain parameter for each channel with a white background template.

The sample is located just about 60 mm below the sensor plane. At this distance a conventional lens with fixed aperture has a shallow depth of field and consequently a large part of the fragment surface is often out of focus. For this reason, our design includes a modified version of the conventional Evetar M12B1216IR lens, which is 12 mm focal length and F1.6 aperture (wide open). We contacted with the manufacturer[2] who provided us a F16 version of this lens, resulting in a larger depth of field whereby almost all the surface is in focus for almost all fragments. The loss of light by narrowing the aperture is not an issue because we control the time exposure and the light intensity. The focus is manually adjusted by rotating a wheel in contact to the lens.

The device is equipped with a specific lighting board. A ring of LEDs is arranged around the aperture for the camera lens, as shown in Fig. 3. There are two sets of four white LEDs `w1`, `w2`, plus two other sets of four infrared LEDs `ir1`, `ir2`. The later emit light around 880 nm and 940 nm, respectively. The LEDs in each set are placed in cross shape and are switched on at the same time, to avoid self shadowing. Two on/off triggered signals control through a decoder the four groups.

4. Dataset

The dataset of images is built upon a collection of 454 samples kindly provided by the urology department of the Hospital Universitary de Bellvitge (Barcelona, Spain) in the time span of several years. They cover all the main 9 classes but cystine, for which just 4 samples were available so we discarded this class as mentioned above. As for the rest, we tried to get all second scheme classes balanced and, at the same time, to record as much examples as possible to account for intraclass variability. The resulting percentages are shown in Table 1. Note that HAP, BRU, STR and AU + CO have slightly more samples than their natural frequency.

One sample consists of two stone fragments from one same unique stone producer and episode, in order to assure independence. Stones are expelled either naturally or after extracorporeal shock wave lithotripsy. In the first case one of the stones is cut with the aid of a scalpel to expose its inner part. In the second, the operator has to select two fragments, each one with an inner and an outer surface. This is important because some classes exhibit characteristic traits in the inner part, like formation of nuclei or concentric layers, as mentioned. The classification of a kidney stone based just on its the external surface or just one fragment is faster, because it requires less time for fragment handling and image capture and processing. However, in preliminary experiments it proved to be less accurate than using both fragments, so we discarded this way.

For each side of each fragment we capture a series of 8 images by varying the exposure time —0.5 and 1 s— and light source —`w1`, `w2`, `ir1` and `ir2` LEDs. Hence, the dataset has a total of 14,528 color images with a resolution of 2592 × 1944 pixels. Fig. 5 shows some of them for two fragments.

In spite of this large figures, we are not going to use all the 32 images per sample in the classification. We recorded them in order

[1] www.baslerweb.com.

[2] www.leadingoptics.com.

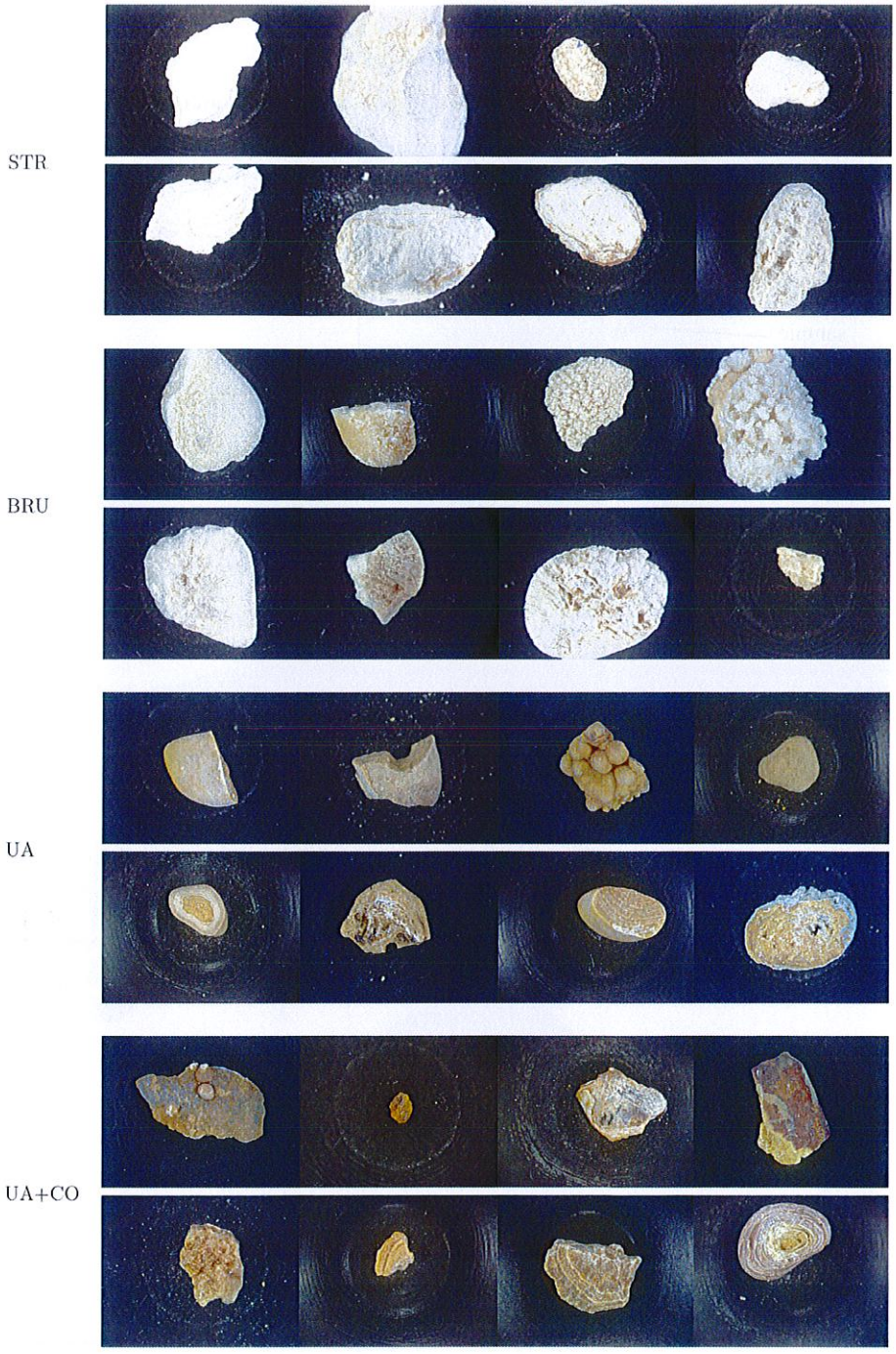

Fig. 2. Continuation of Fig. 1.

lighting (see for instance the two rightmost columns of Fig. 5). Actually, this is their main use, since features computed on them did not add much discriminative power to those computed on visible lighting images. Specifically, a simple thresholding on the image captured with `ir1` LEDs and exposure time equal to 0.5 s, followed by selection of the largest region, already yields a good segmentation in all cases. Given the clear bimodality of the histogram, the threshold value is determined by the classical Otsu method (Otsu, 1979; Sezgin & Sankur, 2004), which has proven quite reliable in our case.

We tried many combinations of feature types and their parameters, on all four light sources and exposure times, which we do not report here. In the end, the best results were achieved with the following features computed on images under white LEDs and 1 s exposure time:

- Rotational invariant local binary patterns (Ojala, Pietikäinen, & Mäenpää, 2002), computed with radius 2, 6 and 8 pixels, and 8, 8 and 10 sample points, respectively. They seem to represent well the different multiscale textures. Even though they are color textures, it was sufficient to compute them on a single channel (red).
- Color histogram, quantizing the color space into 16 bins per channel, in total thus a vector of dimension 4096.

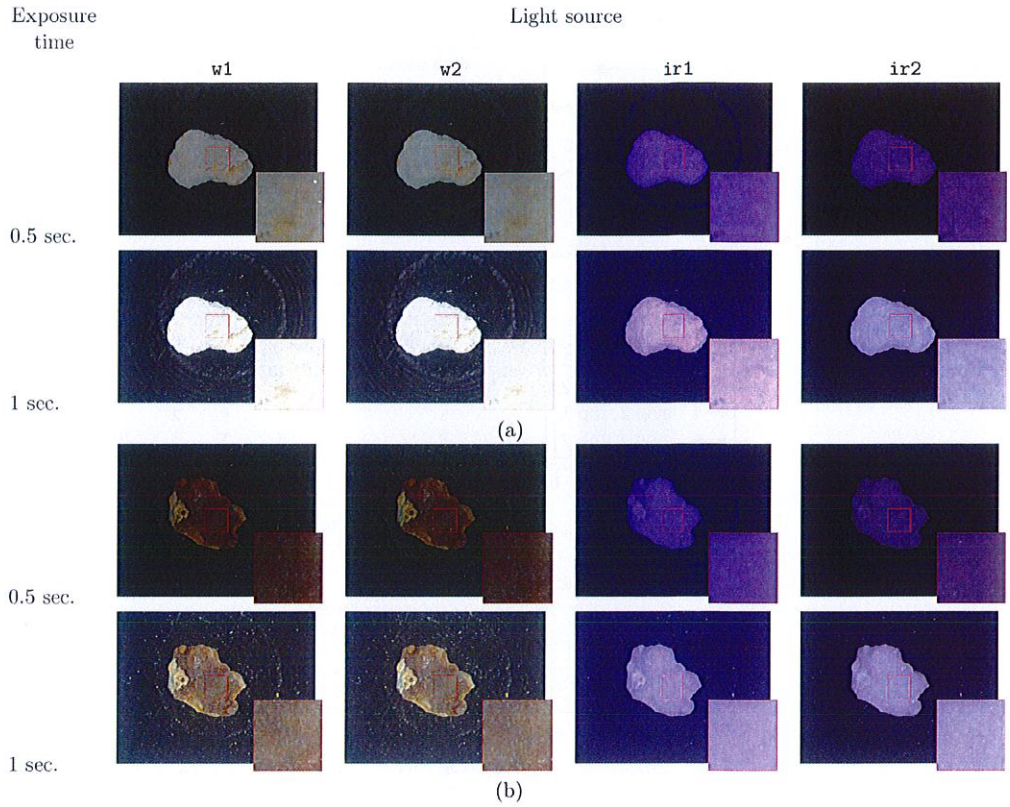

Fig. 5. Set of 8 images taken for the external surface of an (a) struvite, (b) COM fragment. Typically struvite stones are the brightest and COM the darkest.

classifier or the feature computation, we had to train N classifiers using the remaining $N-1$ samples, being $N = 454$ the size of the dataset.

This decision had an important implication with regard to the optimization of the many parameters of the features and classifier.[3] Unless training one leave-one-out classifier could be performed fast, searching for the optimum parameters would have been an extraordinary long task. Fortunately, our development environment was a CentOS Supermicro cluster with 28 nodes, each with 16 x86 64 GB cores. Thanks to this facility we were able to launch all the N training jobs in parallel and get the result in a mere matter of minutes. We implemented the feature extraction and classification in Python on the Scikit-learn and Scikit-image (van derWalt et al., 2014; Pedregosa et al., 2011) libraries.

6. pH level as a feature

Major chemical components of a kidney stone start to precipitate below or above a certain pH level of the urine. Table 2 contains the ranges for each component according to several studies (Bichler et al., 2002; Grases et al., 2002; Spettel, Shah, Sekhar, Herr, & White, 2013). We realize from this table that the pH level range, while not different for every class, clearly separates them into two groups. It thus may help to differentiate samples of visually similar classes like HAP and STR from BRU, and UA, UA + CO from calcium oxalate stones COM, COD, CO + HAP. In addition, it is a measure very easy and fast to obtain. Hence, we have integrated it into the classification pipeline, giving rise to a variant of the former classifier.

Table 2
Ranges of urinary pH level per class.

Class	pH level
COM	< 6.3
COD	< 6.3
CO + HAP	> 6.0
HAP	> 6.5
STR	> 6.8
BRU	< 6.6
UA	< 5.5
UA + CO	< 5.5

There are however two caveats. First, the values in Table 2 are approximate and the pH level may vary with time for a subject, so we can not take them as sharp class borders. Second, we do not have the real pH for the samples in the dataset as we would like. Hence, we have figured out a reasonable pH level from the groundtruth class that we do know. In the intent of achieving a realistic simulation of the actual unknown pH value, given the groundtruth class of a sample, we draw a random value p from a uniform distribution between the class limit in Table 2 and a reasonable minimum/maximum urinary pH level that we have set to 4.5 and 7.0, respectively.

Now, we have to combine this new feature with those derived from images. The random forest classifier we employ produces not only the class label for the most probable class of a sample but also the probability of belonging to each one of the classes. Let n be the number of classes and $P(C_i|\text{visual features})$, $i = 1 \ldots n$ these probabilities. A second vector of probabilities is obtained from the

[3] For example, just for the LBP features we have to find the best number and size of the radius, the number of points on the circle in each case, from which channel to compute them, and whether we want standard, rotational invariance or non-uniform LBPs.

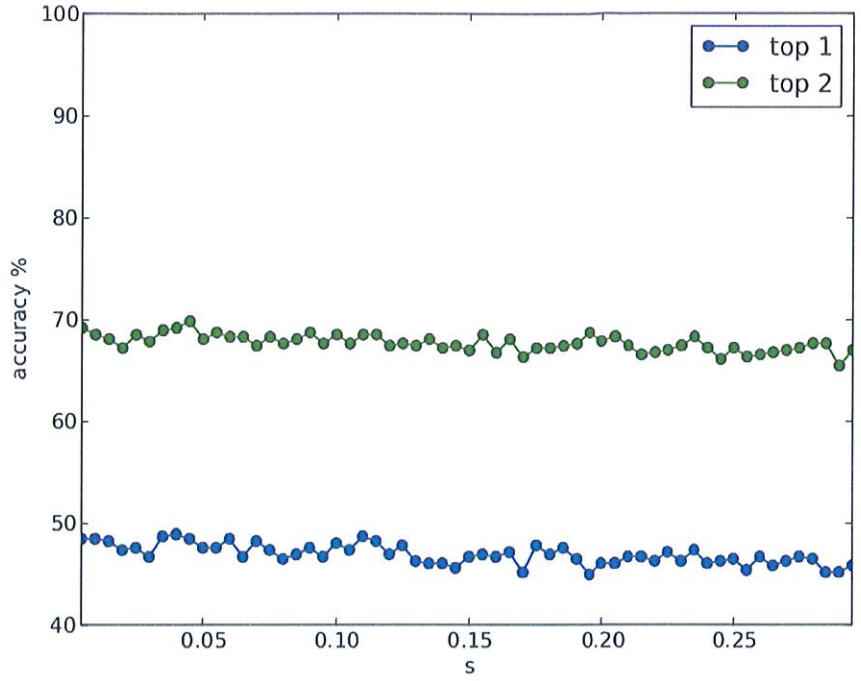

Fig. 7. Accuracy of classification of detailed classes depending on the value of s.

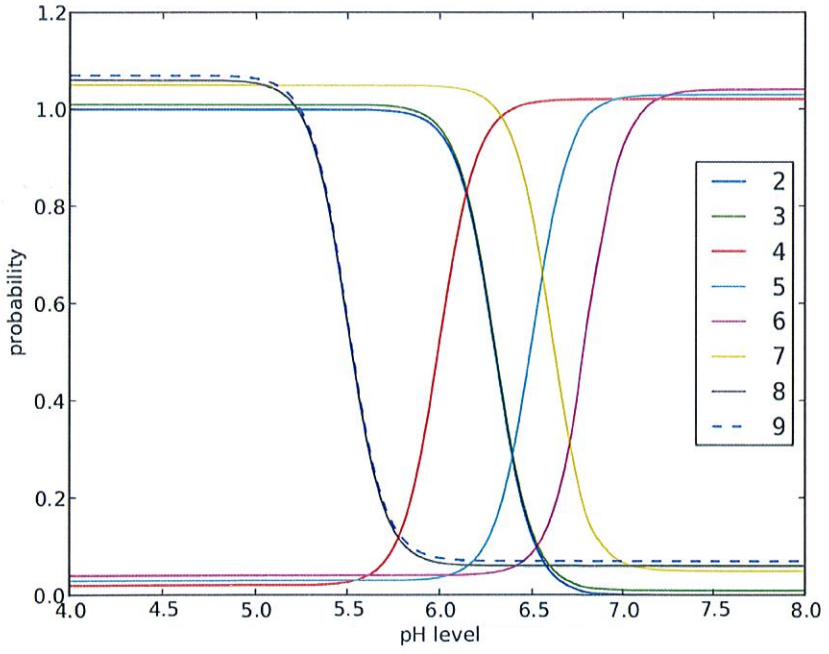

Fig. 8. Class probabilities given pH level for $s = 0.1$. Curves have been vertically shifted for better visualization.

much more complex and varied. Sometimes even the human expert hesitates and resorts to complementary analysis (like infrared spectroscopy) to support his/her decision. The subdivision of COM and COD to form the second scheme, with the introduction of new four classes, only aggravates the problem. This is appreciated in the drop of recall for classes 2, 2b and 2codt for example, formerly 80 and 77% and now between 30 and 40%. The cause is to be found in the top-down nature of the random forest classifier, whereby at each node a certain feature and threshold value are selected to perform a good partition of training samples. The relative good news is that most of the confusions for the new classes are among themselves, that is, most of the errors in the group 2-2b-2codt are confined to the same group, and similarly for 3b-3t-3bt. This is just another manifestation of the former cause. Moreover, the rest of mistakes for each of these two groups are with the other group, keeping the pattern of the confusion matrix of scheme 1.

A last observation is that for class scheme 1, classes 5 and 7 have a very low recall, though the use of pH slightly increases it. The reason is that they have the lower number of samples, 4.2 and 3.1%, respectively.

In addition to the global accuracy we have delved into the contribution of each feature. They have been selected on the basis

design any color or texture feature like us but used the signature of an hyperspectral camera at each pixel, a vector of more than 100 features. Despite their dataset was only composed of 6–8 fragments per class, the fact that they got this result with a classifier as simple as a quadratic discriminant analysis is rather surprising. We interpret it as the extreme importance of selecting a good set of features.

Given these results, we wonder which alternative features and/or classifier could perform better on regular color camera images. The present trend, deep learning, is to learn the features themselves and the classifier altogether. Hence we have already done a number of experiments with convolutional neural networks for feature learning, followed by a few fully connected layers to output class probabilities. After trying several combinations of network architectures and layers plus the typical ingredients of regularization (dropout, weight regularization), we have just reached an accuracy short of 3% with regard the approach described here. We can not afford doing leave-one-out partitions and parallelization of the training phase as done here, because of the long time required to train a neural network and the huge number of folds. Instead, a conventional k-fold partition for some small k needs be adopted.

As future work we have identified in this context the following four lines. First, we hypothesize that our dataset needs to grow considerably before we can improve the present results with deep learning techniques, which are quite demanding in supervised data. Thus, we intend to add more samples to the dataset, trying at the same time to balance the number of samples per class. Second, the network architecture for classifying our samples can not be the same as the standard deep convolutional networks proposed for single images: each sample consists of four images which are not pixelwise compatible and thus can not be merged into a multichannel single image. Some new type of multiview architecture must be designed. Third, we have to deal with the problem of unbalanced classes, for instance through weights or a non-uniform sampling scheme to build the minibatches. And finally, deep networks are known for being overconfident on their predictions. This means that some calibration procedure needs to be applied to the scores they provide so as to approach them to actual probabilities. This is relevant to ours because the output of the system is not only the most probable class but also the confidence or belief on each of them.

Acknowledgments

This research has been supported by the Accio (Generalitat de Catalunya), project VALTEC13-1-0148 and the CaixaImpulse program (CI15-00013). We also thank NVIDIA Corporation for the donation of a GPU board.

References

Andreassen, K., Poulsen, A., Olsen, P., Aabeck, J., & Osther, P. (2007). Classification of urolithiasis in Denmark: Anational survey. *European Urology, 2*(1), 126.

Bichler, K.-H., Eipper, E., Naber, K., Braun, V., Zimmermann, R., & Lahme, S. (2002). Urinary infection stones. *International Journal of Antimicrobial Agents, 19*(6), 488–498.

Blanco, F., Lpez-Mesas, M., Serranti, S., Bonifazi, G., Havel, J., & Valiente, M. (2012). Hyperspectral imaging based method for fast characterization of kidney stone types. *Journal of Biomedical Optics, 17*(7). 076027-1-12.

Blanco, F., Lumbreras, F., Serrat, J., Siener, R., Serranti, S., Bonifazi, G., et al. (2015). Taking advantage of hyperspectral imaging classification of urinary stones against conventional infrared spectroscopy. *Journal of Biomedical Optics, 19*(12), 126004–1–9.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cloutier, J., Villa, L., Traxer, O., & Daudon, M. (2015). Kidney stone analysis: "Give me your stone and I will tell who you are". *World Journal of Urology, 33*, 157–169.

Daudon, M., Bader, C., & Jungers, P. (1993). Urinary calculi: Review of classification methods and correlations with etiology. *Scanning Microscopy, 7*(3), 1081–1104.

van derWalt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., et al. (2014). Scikit-image: Image processing in Python. *PeerJ, 2*, e453.

Friedlander, J. I., Antonelli, J. A., & Pearle, M. S. (2015). Diet: From food to stone. *World Journal of Urology, 33*(2), 179–185.

Grases, F., Costa-Bauzá, A., & Prieto, R. M. (2006). Renal lithiasis and nutrition. *Nutrition Journal, 23*.

Grases, F., Costa-Bauzá, A., Ramis, M., Montesinos, V., & Conte, A. (2002). Simple classification of renal calculi closely related to their micromorphology and etiology. *Clinica Chimica Acta, 322*(12), 29–36.

Hesse, A., Brndle, E., Wilbert, D., Khrmann, K.-U., & Alken, P. (2003). Study on the prevalence and incidence of urolithiasis in germany comparing the years 1979 vs. 2000. *European Urology, 44*(6), 709–713.

Kok, D. J. (2012). Metaphylaxis, diet and lifestyle in stone disease. *Arab Journal of Urology, 10*(3), 240–249. Stones / Endurology.

Lumbreras, F., Serrat, J., & Rotger, G. (2017). *myStone web site*. http://www.cvc.uab.es/mystone [Accessed March 2017].

Nolde, A., Hesse, A., Scharrel, O., & Vahlensieck, W. (1993). Modellprogramm zur Nachsorge bei rezidievierenden Harnsteinpatienten (Model program for follow-up of recurrent urinary stone formers). *Urologe B*, (33), 148–154.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 971–987.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics, 9*(1), 62–66.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Piqueras, S., Duponchel, L., Tauler, R., & DeJuan, A. (2011). Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares. *Analytica Chimica Acta, 705*(1), 182–192.

Romero, V., Akpinar, H., & Assimos, D. G. (2010). Kidney stones: A global picture of prevalence, incidence, and associated risk factors. *Reviews in Urology, 12*(2–3), 89–96.

Scales, C., Smith, A., Hanley, J., & Saigal, C. (2012). Prevalence of kidney stones in the United States. *European Urology, 62*(1), 160–165.

Schubert, G. (2006). Stone analysis. *Urological Research, 34*(2), 146–150.

Sezgin, M., & Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging, 13*(1), 146–168.

Siener, R., & Hesse, A. (2012). *Comparative costs of various treatment strategies and preventive measures* (pp. 897–901). London: Springer London.

Spettel, S., Shah, P., Sekhar, K., Herr, A., & White, M. D. (2013). Using hounsfield unit measurement and urine parameters to predict uric acid stones. *Urology, 82*(1), 22–26.

Strohmaier, W. L. (2011). *Economic implications of medical and surgical management* (pp. 245–250). London: Springer London.

Strohmaier, W. L. (2012). Economics of stone disease/treatment. *Arab Journal of Urology, 10*(3), 273–278. Stones / Endourology.

Wagner, W. P. (2017). Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies. *Expert Systems with Applications, 76*, 85–96.