



**HAL**  
open science

## Automated endoscopic stone recognition using a multi-view fusion approach and a two-step transfer learning

Francisco Lopez-Tiro, Elias Villalvazo-Avila, Juan Pablo Betancur-Rengifo, Jonathan El-Beze, Jacques Hubert, Gilbert Ochoa-Ruiz, Christian Daul

### ► To cite this version:

Francisco Lopez-Tiro, Elias Villalvazo-Avila, Juan Pablo Betancur-Rengifo, Jonathan El-Beze, Jacques Hubert, et al.. Automated endoscopic stone recognition using a multi-view fusion approach and a two-step transfer learning. XXIXème Colloque Francophone de Traitement du Signal et des Images, GRETSI 2023, Aug 2023, Grenoble, France. hal-04170063

**HAL Id: hal-04170063**

**<https://hal.science/hal-04170063>**

Submitted on 25 Jul 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Automated endoscopic stone recognition using a multi-view fusion approach and a two-step transfer learning.

Francisco LOPEZ-TIRO<sup>1,2</sup>, Elias VILLALVAZO-AVILA<sup>1</sup>, Juan Pablo BETANCUR-RENGIFO<sup>1</sup>,  
Jonathan EL-BEZE<sup>3</sup>, Jacques HUBERT<sup>3</sup>, Gilberto OCHOA-RUIZ<sup>1</sup>, Christian DAUL<sup>2</sup>

<sup>1</sup>Escuela de Ingenieria y Ciencias, Tecnológico de Monterrey  
Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L., Mexico

<sup>2</sup>CRAN (UMR 7039), Université de Lorraine and CNRS,  
2 avenue de la Forêt de Haye, 54518 Vandœuvre-lès-Nancy, France

<sup>3</sup>CHU de Nancy, Service d'urologie de Brabois, 54511 Vandœuvre-lès-Nancy, France  
gilberto.ochoa@tec.mx, christian.daul@univ-lorraine.fr

**Résumé** – Cette contribution présente une méthode d'apprentissage profond pour l'extraction et la fusion d'informations d'images acquises sous différents points de vue, le but étant d'obtenir des caractéristiques plus discriminantes pour l'identification du type des calculs rénaux vus dans des images endoscopiques. Le modèle a été amélioré à l'aide d'une méthode de transfert de connaissances en deux étapes et des modules d'attention pour affiner les cartes de caractéristiques apprises par ce modèle. Ces stratégies de fusion de caractéristiques profondes ont permis d'améliorer les performances des extracteurs à vue unique puisque la précision de la classification des calculs rénaux a augmenté de 6% par rapport aux méthodes de référence.

**Abstract** – This contribution presents a deep-learning method for extracting and fusing image information acquired from different viewpoints, with the aim to produce more discriminant object features for the identification of the type of kidney stones seen in endoscopic images. The model was further improved with a two-step transfer learning approach and by attention blocks to refine the learned feature maps. These deep feature fusion strategies improved the results of single view extraction backbone models by more than 6% in terms of accuracy of the kidney stones classification.

## 1 Introduction

The formation of kidney stones that cannot freely pass through the urinary tract is a major public health issue. In industrialized countries, it has been reported that at least 10% of the population suffers from a kidney stone episode once in their lifetime. In the United States alone, the risk of relapse of the same type of kidney stone has increased by up to 40%. The formation of kidney stones is caused by different factors such as diet, low fluid intake, and a sedentary lifestyle. However, there are other unavoidable factors such as age, genetic inheritance, and chronic diseases that increase the risk of forming kidney stones [1]. Therefore, methods for identifying the different types of kidney stones are crucial for the prescription of appropriate treatments and to reduce the risk of relapses. In order to carry out this identification in clinical practice, different procedures have been developed, such as the Morpho-Constitutional Analysis (MCA), and Endoscopic Stone Recognition (ESR). MCA is commonly accepted as the standard procedure for determining the different types of kidney stones (up to 21 different types and sub-types including pure and mixed compositions are recognized during the MCA). MCA consists of a double laboratory analysis of kidney stone fragments ex-

tracted from the urinary tract during an ureteroscopy [2]. First, a biologist performs a visual inspection of the kidney stone which is observed with a magnifying glass. This inspection aims to describe kidney stones in terms of colors, textures, and morphology. This visual analysis is done both for the surface view (the external part of the kidney stone fragment), and for a cross-section of the kidney stone fragment (the internal stone part may consist of several layers surrounding a nucleus). Then, a biochemical analysis using a Fourier Transform Infrared Spectroscopy (FTIR) provides a detailed description of the chemical composition of the kidney stone. Finally, the MCA analysis returns the type of kidney stone through a detailed report of the biochemical and morphological characteristics of both views of the kidney stone. However, MCA has some major drawbacks : the results are often available only after several weeks, and it is difficult to have a specialized team in each hospital to perform MCA.

Therefore, urologists have proposed, as a possible alternative, the Endoscopic Stone Recognition (ESR) procedure in which the most common kidney stones are visually identified on the video displayed on a screen during the ureteroscopy itself [3]. To follow the process performed in the visual inspection of MCA for ESR, in addition to the surface view, the sec-

tion view is obtained by fragmenting the kidney stone with a laser. However, this visual analysis of the surface and section views requires a great deal of expertise due to the high similarities between classes. Only a limited number of specialists have this expertise. In addition, this technique is more operator dependent and subjective than MCA. Therefore, new approaches based on deep-learning (DL) methods have been proposed to automate and speed-up the kidney stone identification. Such automated recognition can potentially assist urologists for a real-time decision-making during an ureteroscopy.

This paper has two contributions : i) it proposes a novel DL-model for fusing information included in endoscopic images of the two views (surface and section) of a kidney stone fragment with the aim to increase the discrimination performance and, ii) it shows how a multi-branch model can be trained using a two-step transfer learning (TL) approach in order to improve the model generalization capabilities.

This paper is organized as follows. Section 2 reviews the literature on automated ESR and introduces the key concepts used in this work, namely multi-view fusion and two-step TL. Section 3 describes the construction of the dataset, details the two-step TL setup, and presents the pre-training stage of the multi-view model. Section 4 compares the results obtained with the proposed model in several configurations, with that of other models given in previous works. Finally, section 5 discusses future research directions.

## 2 State-of-the-art

Different DL approaches for an automated classification of kidney stones demonstrated encouraging results [4]. However, DL-models require large data amounts to yield accurate results. In ureteroscopy, it is difficult to collect such large datasets. A solution to this issue lies in methods such as TL and fine-tuning from other distributions (ImageNet) as a weight initialization technique. Such techniques also enable to avoid training from scratch. However, for an automated endoscopic stone recognition (aESR), these initialization techniques are not useful, since the distribution of ImageNet and endoscopic (ureteroscopic) images are substantially different. Thus, customized TL methods that initialize useful weights closer to the target domain are required. Furthermore, most models performing aESR were trained on surface or section images taken separately. However, the visual inspection in MCA (by biologists) and ESR (by urologists) is based on both views by jointly exploiting information from fragment surfaces and sections. So far, the DL-models in the literature did not use together surface and section information to improve the classification efficiency. Multi-View (MV) classification is exploited in this contribution to combine the features observed in the two fragment-type views.

The aim of this paper is to show that an MV-model outperforms models without an elaborated fusion strategy. MV is performed by fusing features (of shallow models) or feature maps (for DL-models) determined for various images with the aim to learn more complete representations and to obtain more effective classifiers [5]. Contrary to a MV-approach, previous works

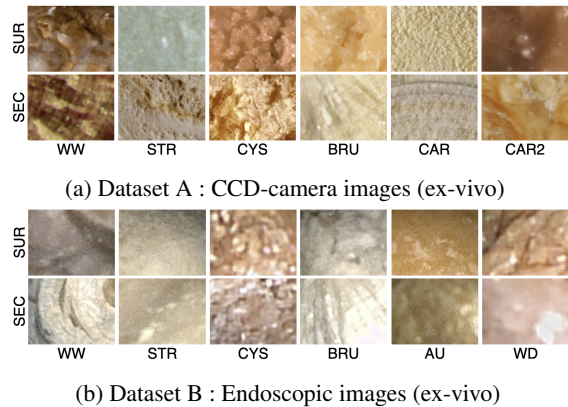


FIGURE 1 – Examples of ex-vivo kidney stone images acquired with (a) a CCD camera and (b) an endoscope. SEC and SUR stand for section and surface views, respectively.

for aESR were based on a DL-model, trained three times (only with section data, only for surface data, and for surface and section data gathered in the same class). This contribution leverages recent advances in DL-based models that combine information from multiple viewpoints and improve the results using domain adaptation techniques.

## 3 Materials and Methods

### 3.1 Datasets

Two kidney stone datasets were used in our experiments [6, 7]. According to the dataset, the images were acquired either with standard CCD cameras or with an ureteroscope (i.e., an endoscope). These datasets are described below.

**Dataset A**, [6]. This ex-vivo dataset of 366 CCD camera images (see, Fig. 1a) is split in 209 surface and 157 section images, and contains six different stone types sorted by sub-types denoted by WW (Whewellite, sub-type Ia), CAR (Carbapatite, IVa), CAR2 (Carbapatite, IVa2), STR (Struvite, IVc), BRU (Brushite, IVd), and CYS (Cystine, Va). The stone fragment images were acquired with a digital camera under controlled lighting conditions and with a uniform background.

**Dataset B**, [7]. The endoscopic dataset consists of 409 images (see Fig. 1b). This dataset includes 246 surface and 163 section images. Dataset B involves the same classes as dataset A, except that the Carbapatite fragments (sub-types IVa1, and IVa2) are replaced by the Weddelite (sub-type IIa) and Uric Acid (IIIa) classes. The images of dataset B were captured with an endoscope by placing the kidney stone fragments in an environment simulating in a quite realistic way in-vivo conditions (for more details, see [7]).

Obtaining the section view of a kidney stone is a difficult task, since it is operator-dependent, therefore, the surface and section sets are not fully paired. Automatic kidney stone classification is usually not performed on full images due to the limited size of the datasets. Therefore, as in previous works [8], patches of  $256 \times 256$  pixels were extracted from the original images to increase the size of the training dataset (for

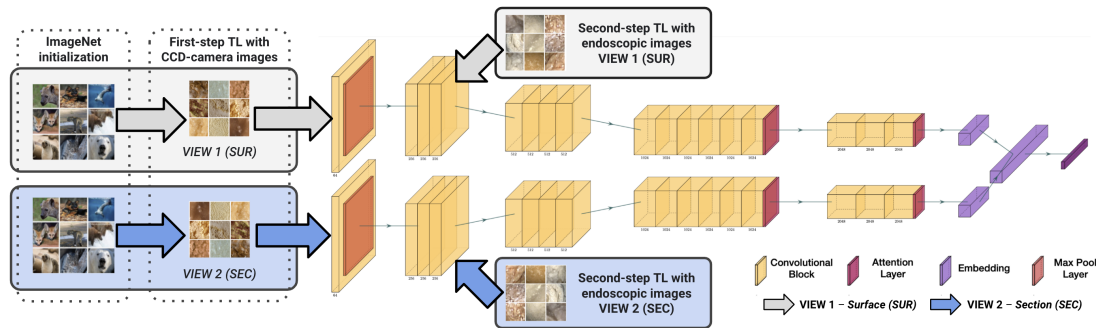


FIGURE 2 – Proposed multiview-fusion model assisted by two-step transfer learning for aESR.

more details, see [4]). A total of 12,000 patches were generated for each dataset which is organized as follows : For dataset A (WW, STR, CYS, BRU, CAR, CAR2) and dataset B (WW, WD, AU, STR, BRU, CYS). For dataset A and B, a thousand patches are available for each class and view (SUR, SEC). For each data set, 80% of the patches (9600 patches) are used for the training and validation steps, while the remaining 20% of the patches (2400 patches) act as test data. Patches of the same image contribute either only to the training/validation data or solely to the test data. The patches were also “whitened” using the mean  $m_i$  and standard deviation  $\sigma_i$  of the color values  $I_i$  in each channel [4].

### 3.2 Proposed approach

Several approaches [4] have demonstrated the ability of DL-based models to recognize in single views (SUR or SEC) different types of kidney stones with high performance. However, in most cases, they have been trained by fine-tuning with a totally different distribution than kidney stones, or worse, they have been trained from scratch with the endoscopic images for individual views. On the other hand, so far no elaborated technique was exploited to combine the surface and section information. Usually, to exploit the information from SUR and SEC images, the patches of the two views of a fragment are simply seen as instances of the same class. Although such methods fuse both views information and more data available for the training, the way in which image features are extracted and combined is far from being optimal, as it does not emulate how the visual inspection of MCA/ESR is performed. To make matters worse, mixing the features in this way does not always improve the classification results. As can be observed in the MIX column of Table 1 (values marked by the \* symbol), in some cases fusing features from SUR and SEC patches does not produce better feature maps, as this information combination is not optimal and hinders the model performance [8].

In order to exploit the best features of both views, the proposed DL-model (see Fig. 2) combines the information in a systematic way using a fusion strategy based on a multi-view scheme, introducing attention mechanisms to further filter out unnecessary features maps of our CNN-model. Moreover, instead of training from the scratch the individual branches, we assist the model training with a two-step TL approach as a method of initializing weights from a similar distribution (CCD-

camera images, ex-vivo) to the endoscopic images (in-vivo).

### 3.3 Two-step Transfer Learning

The DL-model acquires knowledge in several ways. During the HeTL (HeTL stands for heterogeneous TL), the pre-training is performed with a general domain. The model weights are updated during a HoTL (homogeneous TL) using a domain whose data distribution is the closest to that of the target (domain adaptation process, see[8]). In the kidney stone application, the pre-training on ImageNet improves the generalization capabilities of the DL-model and the CCD camera images of ex-vivo fragments are used as a first fine-tuning. This fine-tuning is finalized using a part of the target dataset B (fragment images acquired with endoscopes), the remaining patches of dataset B being used for the validation and testing steps. More specifically, during the HeTL-step, the large ImageNet dataset is used to transfer knowledge into a ResNet50 network which is fine-tuned by the smaller kidney stone image set acquired under controlled acquisition conditions (dataset A) as shown on the left part of Fig. 2. Then, fine-tuning is achieved for each branch (i.e. individual model for each view) during the HoTL-step. This final tuning exploits dataset B which is composed of endoscopic images close to dataset A, but with higher variability in terms of image contrast, noise, and resolution, emulating thus the illumination and scene conditions actually encountered in ureteroscopy when patient data are acquired with an endoscope. The second-step TL is performed for each of the views (SUR/SEC) by obtaining two independent models trained with dataset B of endoscopic images for their respective views (for more details, see [8]). As described below, a MV-model, assisted by the second TL-step, is used to combine the SUR and SEC views into a mixed model (MIX).

### 3.4 Multi-view model

Once the two SUR and SEC models are trained through the previous two-step TL, the feature extraction layers of this single-view network are frozen to ensure that each branch of the multi-view model extracts the same features and that any variation in performance depends on the non-frozen layers (merge and full connection layers). These frozen layers are connected to a fusion layer, which is responsible for mixing the information of the two views. In this work, the two late-fusion methods pro-

TABLE 1 – Comparison of the performance of various aESR DL-methods. The classification accuracy (in percentage) overall classes was determined with test dataset B for all methods.

Method	SUR	SEC	MIX
Martinez, et al. [10]	56.2±23.3	46.6±12.8	*52.7±18.9
Estrade, et al. [3]	73.7±17.9	78.8±10.6	*70.1±22.3
Black, et al. [9]	73.5±19.0	76.2±18.5	80.1±13.8
Lopez-Tiro, et al. [4]	81.0±03.0	88.0±02.3	*85.0±03.0
<b>This contribution</b>	<b>83.2±01.2</b>	<b>90.4±04.8</b>	<b>91.2±0.50</b>

posed in [5] were exploited. On the one hand, the first method concatenates the feature vectors obtained from each view and merges the resulting representation through a fully connected layer. On the other hand, in the second method, feature vectors are stacked and max-pooling is applied to them. Two configurations were used to implement max-pooling. The first corresponds to a model without attention mechanisms. The second consists of two layers of attention (arranged as shown in Fig. 2). The results presented in this work correspond to the second configuration "max-pooling" (for more details, see [5]). Lastly, the output of the late-fusion layer is connected to the remaining part of the MV-model, which merely consists of the classifier. The full proposed model is shown in Fig. 2.

## 4 Results and Discussion

Three experiments were carried out to assess the performance of the two-step TL approach applied to the patch data described in Section 3.1. In the first and second experiments, the two-step TL approach described in Section 3.3 was used to predict kidney stone types in endoscopic images for SUR and SEC views, respectively. Then, in the third experiment, the models trained on SUR and SEC data were combined using the MV-model described in Section 3.4. The results of these experiments are gathered in Table 1 and discussed below.

The results obtained for the first and second experiments follow the trend observed in [4]. For the SUR view, a mean accuracy of  $83.2 \pm 1.2$  (in %) was obtained. On the other hand, the results observed for the SEC view (accuracy of  $90.4\% \pm 4.8$ ) are better than those obtained for the SUR view, probably due to the extraction of more discriminant features. The importance of section data was highlighted in previous works and is confirmed by this contribution. In comparison to the state-of-the-art, the highest performances were reached by the presented TL-method, both for the SUR and SEC views taken separately.

For the third experiment, fusion through MV, an accuracy of  $91.42\% \pm 0.5$  was obtained. This result is given for the max-pooling configuration with attention. However, the concatenation configuration ( $89.8\% \pm 3.03$ ) presents results very close to the max-pooling configuration. Regardless of the configuration selected for the MV-model, the fusion shows promising results. First, the accuracy obtained in the MIX column in Table 1 suggests a clear improvement over the-state-of-the-art. Secondly, it shows that combining in an efficient way both views

(SUR/SEC) in a "mixed" model can maintain the performance, contrary to the models marked by the \* symbol. The latter shows that combining the SUR and SEC information of stones in a single class leads to a performance decrease.

## 5 Conclusion and future work

This contribution shows that, by mixing information from two views, it is possible to train more accurate models to identify kidney stones acquired with endoscopes. Thus, AI technology can be an interesting solution for assisting urologists. However, these contributions used a very limited dataset in terms of class number and patch samples. The learning approaches on few samples must be improved to cope with the small amount of training data, and especially to increase the class separability when more kidney stone types have to be identified.

## Références

- [1] J.I. Friedlander, J. Antonelli and M.S. Pearle. *Diet : from food to stone*. World journal of urology, 2015.
- [2] M. Daudon, P. Jungers, B. Dominique and J.C. Williams. *Recurrence rates of urinary calculi according to stone composition and morphology*. Urolithiasis, 2018.
- [3] V. Estrade, M. Daudon, E. Richard, J.C. Bernhard, F. Bladou and G. Robert. *Towards automatic recognition of pure & mixed stones using intraoperative endoscopic digital images*. BJU International, 2021.
- [4] F. Lopez-Tiro. *Assessing deep learning methods for the identification of kidney stones in endoscopic images*. Int. Conf. of the IEEE EMBC Society, 2021.
- [5] E. Villalvazo-Avila, F. Lopez-Tiro, J. El-Beze, J. Hubert, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, and C. Daul *Improved Kidney Stone Recognition Through Attention and Multi-View Feature Fusion Strategies*, IEEE International Symposium on Biomedical Imaging, April 2023.
- [6] M. Corrales, S. Doizi, Y. Barghouthy, O. Traxer, M. Daudon. *Classification of stones according to Michel Daudon : a narrative review*, European Urology Focus, 2021.
- [7] J. El-Beze, C. Mazeaud, C. Daul, G. Ochoa-Ruiz, M. Daudon, P. Eschwège, and J. Hubert *Evaluation and understanding of automated urinary stone recognition methods*. BJU international, 2022.
- [8] F. Lopez-Tiro, J.P. Betancur-Rengifo, A. Ruiz-Sanchez, I. Reyes-Amezcuca, J. El-Beze, J. Hubert, M. Daudon, G. Ochoa-Ruiz, C. Daul *Boosting Kidney Stone Identification in Endoscopic Images Using Two-Step Transfer Learning*.
- [9] K.M. Black, H. Law, A. Aldoukhi, J. Deng and K.R Ghani. *Deep learning computer vision algorithm for detecting kidney stone composition*. Wiley Periodicals Inc., 2020.
- [10] A. Martínez, D-H. Trinh, J. El Beze, J. Hubert, P. Eschege, V. Estrade, L. Aguilar, C. Daul and G. Ochoa. *Towards an automated classification method for ureteroscopic kidney stone images using ensemble learning*. Int. Conf. of the IEEE EMBC Society, 2020.