Pattern Recognition 61 (2017) 629–640

Contents lists available at ScienceDirect

Pattern Recognition

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

Bacterial colony counting with Convolutional Neural Networks in Digital Microbiology Imaging

Alessandro Ferrari [a,b], Stefano Lombardi [a], Alberto Signoroni [a,*]

[a] Information Engineering Department, University of Brescia, Brescia, Italy
[b] Futura Science Park, Copan Italia S.p.A., Brescia, Italy

ARTICLE INFO

Article history:
Received 30 January 2016
Received in revised form
1 June 2016
Accepted 7 July 2016
Available online 9 July 2016

Keywords:
Convolutional Neural Networks
Deep learning
Image classification
Handcrafted feature extraction
Image analysis
Bacterial colony counting
Digital Microbiology Imaging
Full Laboratory Automation

ABSTRACT

Counting bacterial colonies on microbiological culture plates is a time-consuming, error-prone, nevertheless essential quantitative task in Clinical Microbiology Laboratories. With this work we explore the possibility to find effective solutions to the above issue by designing and testing two different machine learning approaches. The first one is based on the extraction of a complete set of handcrafted morphometric and radiometric features used within a Support Vector Machines solution. The second one is based on the design and configuration of a Convolutional Neural Networks deep learning architecture. To validate, in a real and challenging clinical scenario, the proposed bacterial load estimation techniques, we built and publicly released a fully labeled large and representative database of both single and aggregated bacterial colonies extracted from routine clinical laboratory culture plates. Dataset enhancement approaches have also been experimentally tested for performance optimization. The adopted deep learning approach outperformed the handcrafted feature based one, and also a conventional reference technique, by a large margin, becoming a preferable solution for the addressed Digital Microbiology Imaging quantification task, especially in the emerging context of Full Laboratory Automation systems.

1. Introduction

What it has been considered, for a long time and till a very recent past, the dream of fully automatizing the complex workflow of Clinical Microbiology Laboratories (CML) [1] is now becoming a tangible reality thanks to the advent and rapid diffusion of the so-called Full (or Total) Laboratory Automation (FLA) systems [2–4]. A main characteristic of FLA is the possibility to acquire (and then process, visualize, store and communicate) high-quality digital images of the bacterial cultures on solid agar plates, during their incubation. Diagnostic Microbiology Imaging (DMI) in the context of FLA systems is triggering a new digital revolution in CML.

Thanks to the recent advancements of image analysis and interpretation technologies, there is a broad range of crucial tasks where DMI technologies can have a key role in favoring faster, more robust and reliable diagnostic procedures: from the automation of mundane tasks (e.g. the screening of negative plates) to the development of advanced computer-aided image interpretation techniques, capable to face the very high degree of variability that characterizes clinical microbiology applications and to support accurate and early diagnosis for the reduction of turn-around times for diagnosis and patient treatment.

1.1. Objectives and contribution

The number of colonies on culture plates, also called *colony forming units* (CFU), is used in clinical microbiology as a rough estimate of the number of viable bacteria in a sample. Therefore, an important step in many CM protocols toward the estimation of the nature and degree of infections is bacterial colony counting (BCC). However, BCC is inherently a complex task, in particular if thought as a component of a system that cannot fail (in fact, underestimation errors in this phase may become missed diagnosis of infectious diseases, thus missed treatment of the patient) and that must operate in routinary clinical conditions (characterized by a high variability in bacterial growing patterns and by the possible presence of other disturbing elements on the plate). As a consequence, what can be thought as a mundane and relatively simple task for microbiologists, it is actually a very complex task for a machine. While for higher bacterial loads, when massive confluent areas occur, the count is not required to be exact, and some coarse estimation can be done (this is a different problem with respect to BCC and it is not addressed in this work), an accurate and precise count of single colonies and countable confluent agglomerates must be guaranteed, especially for relatively low bacterial loads

* Correspondence to: Alberto Signoroni, PhD Università degli Studi di Brescia Information Engineering Dept. Via Branze, 38 - 25123 Brescia, Italy
E-mail address: alberto.signoroni@unibs.it (A. Signoroni).

http://dx.doi.org/10.1016/j.patcog.2016.07.016
0031-3203/

(e.g. less than 80–100 colonies), so that the software can safely sort the plates in terms of bacterial load intervals or screen out negative samples (i.e. presenting a number of colonies lower than a given threshold). The problem complexity derives from the high morphological variability of the colony agglomerates in terms of number, dimension and configuration of the CFUs in each single aggregate.

This work builds on our preliminary study [5] where we tested the suitability of a Convolutional Neural Network (CNN) approach to address the considered task. The objectives of the present work are (1) to confirm the promising results of [5] on a significantly extended and improved database of colony segments (the adopted CNN topology has been confirmed to be the most suitable one by an accurate cross-validation process), (2) to test various dataset augmentations and normalizations and their suitable combination and (3) to demonstrate the superiority of the proposed deep learning approach with respect to what can be considered a strong competitor, i.e. a properly designed and implemented handcrafted feature based SVM classification system that works on the same target counting classes and that exploits a complete and discriminative set of colorimetric and morphometric features extracted from the bacterial colonies. Comparisons have also been extended to the results obtained by watershed based counting [6,7] which is currently the most diffused solving approach. An effort has been made to significantly extend the database from 17k to 28.5k images, all taken from clinical samples, with reduced class skewness and with a completely revised labeling by a second expert. This database is made public in order to serve as a reference for research works on the same or related tasks in the emerging Digital Microbiology Imaging application field.

The rest of the paper is organized as follows: in Section 2 we give an overview on BCC approaches we found in the literature as well as on the recent applications of CNN to a variety of challenging biomedical tasks. A concise overview of the acquisition system is given in Section 3, while Section 4 is devoted to the presentation of our new fully labeled image segments database and related pre-processing and enhancement solutions. The proposed CNN-based approach to the estimation of the cardinality of CFU aggregates and outlier rejection is described in Section 5 as well as the description of the handcrafted feature-based classification pipeline proposed as a competing technique. Experimental results and performance comparisons are presented and discussed in Section 6, while concluding remarks are provided in Section 7.

2. Related work

The literature on bacterial colony counting accounts for many interesting approaches claiming, in some cases, performances comparable to human technicians. However, the experimental settings of all of these studies were quite specific. Many literature methods were tested just on one or few strains, with few different bacterial load tested generally on a limited amount of plates and/or controlled lab conditions. None of the existing works have been tested against a large variety of samples coming from clinical routine, representative of the real complexity of the task. This is probably a main reason of why the existing methods are not widely adopted in CML. Another reason may be that many of the existing solutions are *software-centric*, claiming good flexibility regardless of the recording system. However, the adopted segmentation methods are parametric and, in practice, even if slightly adjusted, they are incapable to effectively handle differences among recording systems at a large scale. Thus, many of the existing colony counter proposed a “best effort approach”: they attempt to count, then they propose to the user a visual overlay of the segmentation and segments enumeration results, so that the user can understand if something went wrong and, whether necessary, to roll back and count herself/himself colonies on the plate. In clinical applications, reliability and cost-effectiveness are of primary importance, thus such best effort approach is not acceptable, because the human would need to be strongly present in the loop, deteriorating the throughput instead of increasing it. A high degree of reliability and a boosted expectation for a widespread adoption of such technology can be reached only with a controlled, repeatable and replicable recording systems. This degree of image acquisition standardization is exactly what the increasingly diffusing FLA systems are capable to provide [4]. Bacterial colony counting has been first attempted by pioneers in [8–10]. These works proposed custom recording and processing systems for performing bacterial colony counting on agar plates. In [11], it is proposed to apply distance transform on a segmented binarized image and to consider as colonies the local maxima with values over a certain threshold. A method that exploits a particular lighting producing countable spot reflections on certain colonies is proposed in [12]. A method that uses the watershed transform for splitting clumps once the colonies are segmented is reported both in [6] and in [7]. Another grayscale morphological analysis solution for counting is introduced in [13]. The above methods usually rely on an involved parameters hand-tuning that can be effectively adjusted only for some limited settings. However, they may have difficulties when facing the large variety encountered in clinical routine. For example, it is hard to set correct threshold for a watershed splitting when dealing at the same time with micro-colonies with about 10 pixels diameter on high resolution images and macro-colonies with hundreds or even thousands pixels diameter, especially in clinical settings where not all the colonies have a regularly rounded morphology. The shortcomings of traditional image processing methods have suggested the design of handcrafted feature-based classification approaches for determining the segments cardinality. In [14] a method based on shape classification of the segments is presented. A Sanger neural network is adopted for performing dimensionality reduction of the binary segments that are then classified in categories from 1 to 7 colonies, or outliers, obtaining results similar to those attainable by watershed based techniques. The OpenCFU solution [15] proposes a segmentation method based on multiple thresholding, then a score map is computed from the multiple segmentation by means of a particle filter on simple segments moments; confluent segments are separated by means of a watershed transform on the distance map. In [16], a multistage classification identifies isolated colonies, which are detected by means of a classifier taking as inputs Zernike moments representations of the binary segments and, where detected, isolated colonies are subsequently classified to recognize different bacterial species.

CNNs are hierarchical models that are attaining state-of-the-art performances for many object classification and detection applications. They have been first introduced for overcoming known problems of fully connected deep neural networks when handling high dimensionality structured inputs, such as images or speech [17]. Convolutional layers topologically encode spatial correlation by means of local receptive fields, moreover they enforce shift and distortion invariance by feature pooling and by forcing shared weights on features map, so that each feature map replicates the same features on the local receptive fields all over the input. Encoding those constraint in the network topology reduces the number of parameters to learn, thus the training set dimensionality and the computational resources needed, while theoretically just slightly impacting the expressive power of the model. Recently, favored by the advent of fast GPU and smart devices added to the original design, CNNs have become state-of-the-art solutions for large scale object classification [18,19] and object detection tasks [20,19]. CNNs have been already applied to a variety of

biomedical imaging problems. In [21] cells and nuclei of developing embryos of *Caenorhabditis elegans* roundworm were segmented and located. A system that performs automatic segmentation of neuronal structures on electronic microscope images was presented in [22], while in [23] a CNN system is designed for mitosis detection on cell nucleus in breast cancer histology images, significantly outperforming previous solutions. Another approach for detecting mitosis in live-imaging microscopy images has been proposed in [24]. In [25] an architecture for segmenting neuronal structures in electron microscopic images is experimented. In [26] a fast scanning deep convolutional neural network has been applied for histopathological image segmentation. Eventually, without wanting to be exhaustive, an increasing number of original investigations related to various task in the vast field of radiology diagnostic imaging are appearing. For example, chest radiography retrieval [27], mass lesion classification in mammography [28], spine metastases detection in Computed Tomography [29] and knee segmentation in Magnetic Resonance Imaging [30].

3. Image acquisition

This study focuses on clinical urine samples (collected from the routine activity of CML sites) that numerically represent a large portion of the whole set of specimens examined in CMLs [31]. Similar to what happens in many laboratories worldwide, specimens have been inoculated on Trypticase Soy Agar with 5% sheep blood (an example is shown in Fig. 1). The urine samples are representative of the majority of human pathogens, not only those of the urinary tract, so a computer vision system trained on this type of samples has inherently good generalization properties. Images are produced by the acquisition equipment within the WaspLab™ (Copan Italia S.p.a., Italy) FLA system, a combination of high-resolution linear camera, multiple lighting system and sample conveyor. The camera has a linear sensor that acquires at each shot 3 lines (RGB channels) of more than 4000 pixels. The plate slides under the camera by an automated sledge. The result of the scan is a high resolution image of the camera that counts more than 4000×4000 isotropic pixels (0.0265 mm/pixel ratio). Since the bacteria colony size may be very small, a low noise acquisition system is important in order to have good results even at high zoom. Particular attention has been spent also on the setting of the illumination conditions. A led lighting bar has been used in order to spotlight the acquired sample stripes. Reflections help to visually appreciate the degree of convexity and the morphology of bacterial colonies. The acquisition system can acquire a series of image of each blood agar plate at different times. Here we use images taken at time 0 (just after plating) and time N (N=16–48 h, depending on the culture protocol).

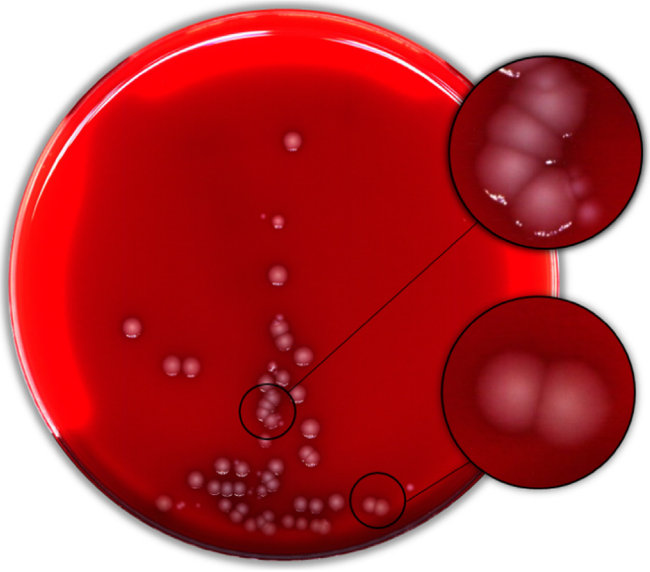

Fig. 1. An example of an acquired image of a bacterial growth on a blood agar plate with details showing colony aggregates.

4. Colony segments dataset

4.1. Segmentation

The captured images of blood agar plates are segmented by a software made available on the WASPlab system. Colonies are considered as foreground objects that grows over the agar which is a fairly uniform background. Time 0 image is useful to make a differential operation that can discriminate and isolate the bacterial growth, so as to achieve isolated growth segments which can contain single colony or isolated colonies agglomerates. The segmentation pipeline is composed of the following two phases: (1) a coregistration of the image of the bacterial culture (*time N image*) to the image of the plate right after inoculation (*time 0 image*) by means of a normalized roto-translational correlation, and (2) a mixed global thresholding [32] and adaptive thresholding [33] through which a segment and a binary mask can be extracted.

CFU aggregates can have very different sizes from really tiny micro-colonies of 0.1 mm^2 to huge confluent configurations of 3 cm^2. Accordingly, the resulting rectangular image segments containing the isolated or aggregated segmented CFUs have variable dimension. Also image contrast may vary a lot from highly contrasted colonies with possible halos around them (due to blood haemolysis typical of certain bacteria species) to almost transparent small colonies that are barely visible at certain visualization scales.

4.2. Dataset creation

Each of the obtained segments can be assigned to a class, depending on the number of colonies it contains, from 1 to 6, or labeled as an outlier (the seventh class) if, instead of colonies, contains bubbles, dust or dirt on the agar. Fig. 2 shows illustrative image segments belonging to the different classes. Segments that contain seven or more colonies (very rare) can be considered as confluences and therefore they are excluded from this classification problem. Segments have been labeled by experienced operators by means of a dedicated GUI (see Fig. 3) and the corresponding data have been stored using a custom metadata format. A significant effort was made to reach a fully labeled database size of about 28,500 images. The dataset presents skewed classes, since most of the segments contain only one colony: 50.2% of segments contain an isolated colony, 19.0% contain 2 colonies, 12.8% 3 colonies, 6.5% 4 colonies, 3.4% 5 colonies, 3.6% 6 colonies and 4.5% of segments contain outliers. This is not surprising as it reflects normal clinical conditions, where isolated colonies are more common compared to clustered ones. Skewness was then only partially corrected (with respect to [5]) to guarantee a significantly large and representative number of examples for each class. In order to stimulate further investigations the entire labeled database is released for research use.[1] As stated, the dataset consists of variable-size images, however traditional CNN requires a constant input dimensionality. Therefore, segments have been resized to

[1] The database is released under Creative Commons license at the following web address: http://www.microbia.org/index.php/resources.

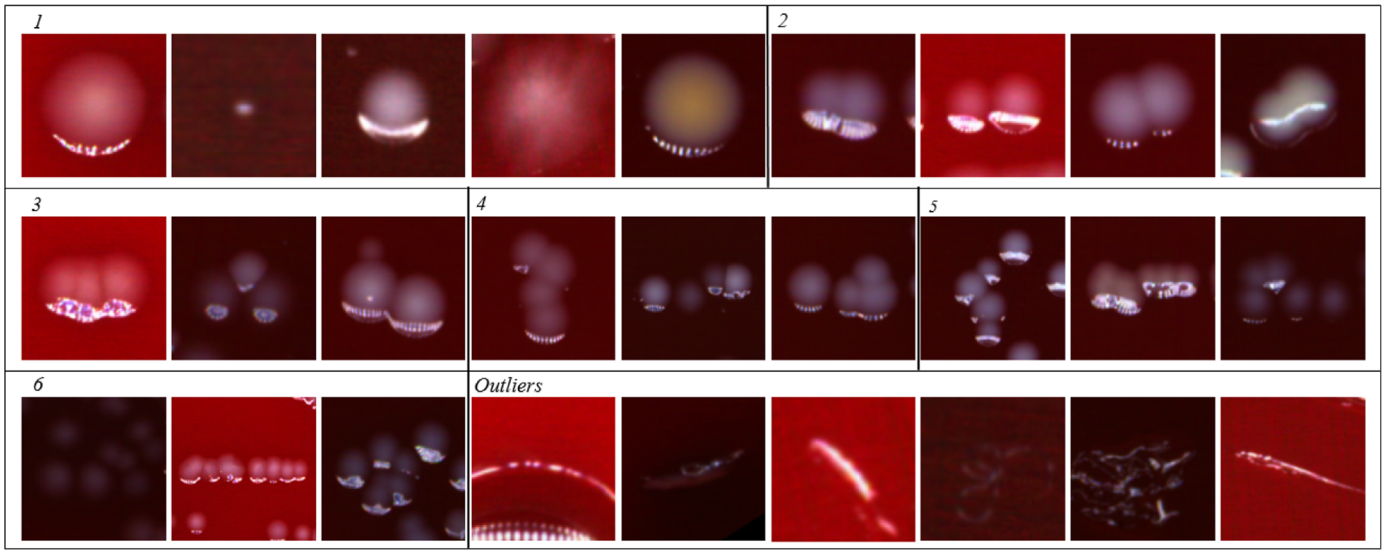

Fig. 2. Example of segments of different classes.

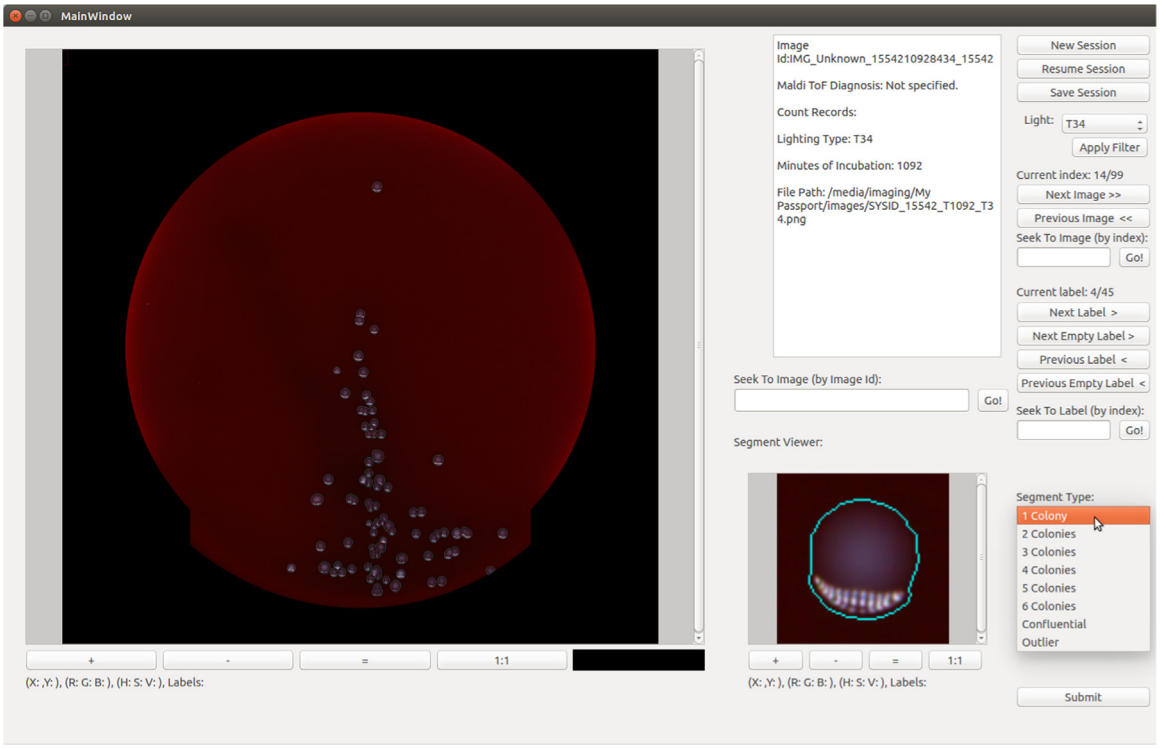

Fig. 3. Segment labeling GUI screenshot.

fixed sizes. Similar to [18] a cropping approach has been used. In this process, segments are cropped into a square image with side dimension longer than the greatest among the horizontal and vertical segment lengths, followed by a border extension of a fixed size and padding whether necessary. This way however, if the segment is elongated in one direction, nearby colonies from other segments may be included in the other direction. To solve this issue and try to remove the disturbing elements, the available segmentation mask information can be used. In doing so, two different output images can be easily produced, as depicted in Fig. 4, and detailed in the following. A bounding box of each segment has been created using the limits of the binary mask (Fig. 4(a)). Since the size of this inner bounding box (dashed line in Fig. 4(a)) is variable and not necessarily square, one possibility is to make a padding of the images, to restore the square dimension, with pixel values calculated as the average of the pixels around the bacterial colonies agglomerate. This way a *bounding box dataset* (Fig. 4(c)) has been created. This dataset better conserves the colony to agar interface but does not completely eliminate extraneous material than can still be present in the inner bounding box (see Fig. 4, third row). Therefore another *masked dataset* (Fig. 4(d)) has been also created by applying the binary mask (Fig. 4(b)) to the corresponding segment (setting the background to black zero values), in order to remove all the possible foreign colonies from the images. These two different baseline datasets have been tested in order to analyze how the network is influenced by the masking process and the flexibility of the network itself. All images have been resized to 128×128 pixels, since smaller sizes, by our tests,

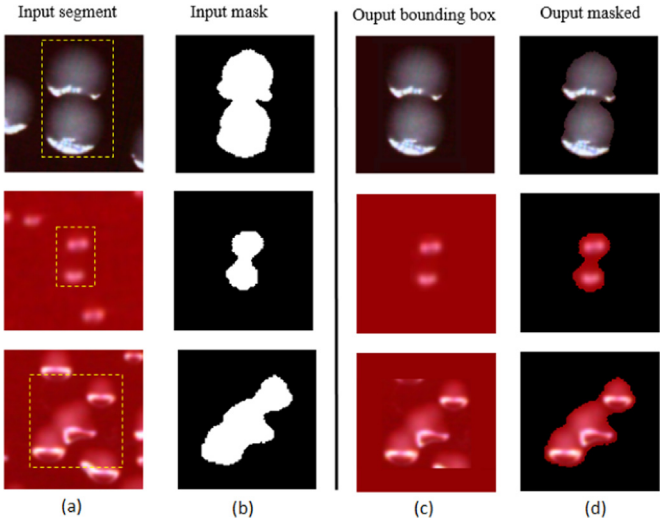

Fig. 4. Example of *masked dataset* and *bounding box dataset* creation from input segment and input mask. The dashed line represents the bounding box created using the limit of the input mask.

tend to decrease the overall performance, while higher sizes are not worth to be used after accurate computational load vs performance gain considerations.

4.3. Dataset enhancement

Dataset enhancements techniques, by means of class-preserving transformations, have been also investigated and tested to increase the training set cardinality for all classes and try to normalize the original images. Different datasets have been created (as depicted in Fig. 5) in order to test the performance of such transformations:

Dataset augmentation:

- A *horizontal flip* has been performed on the images, doubling the size of the training dataset (Fig. 5(a)). A vertical flip does not preserve the symmetry with respect to the reflections produced by the lighting system, so the corresponding augmentation is to avoid.
- Three different *artificial color distortions* on CIE-Lab color space have been applied. This color space is better suited to many digital image manipulations than the RGB space. 3 different magnitude values for these deviations space have been applied (an example in Fig. 5(b)). A dataset of color-distorted images can be useful to determine if the classification system is robust against this kind of distortion.
- Two different values of spatial *rescaling* before cropping (Fig. 5 (c)) has been implemented, in particular with values of 85% and 115% compared to the original size.
- Since, when considered independently, *horizontal flip* and spatial *rescalings* will provide the best performance improvements, their combination has also been considered, applying to the rescaled database a horizontal flip, increasing the cardinality of six times.

Dataset normalization:

- We tested a *Contrast Limited Adaptive Histogram Equalization* (CLAHE) [34] to enhance the local contrast (Fig. 5(d)).
- Another enhancement has been produced with the aim to achieve a *normalization with respect to the segment orientation*. Since colonies cluster together with different angles, this process may enforce some rotation invariance (Fig. 5(e)). Normalization was obtained by calculating the moments of the image converted to gray-scale by:

$$M_{ij} = \sum_x \sum_y x^i y^j I(x, y) \tag{1}$$

with $I(x, y)$ being the image pixel intensities. After the centroid has been calculated:

$$(c_1, c_2) = (M_{10}/M_{00}, M_{01}/M_{00}) \tag{2}$$

and with this value, the central moments can be calculated:

$$\mu_{pq} = \sum_x \sum_y (x - c_1)^p (y - c_2)^q I(x, y) \tag{3}$$

Given $\mu'_{20} = \mu_{20}/\mu_{00}$, $\mu'_{02} = \mu_{02}/\mu_{00}$ and $\mu'_{11} = \mu_{11}/\mu_{00}$ the segment orientation can been extracted by calculating its angle as follows:

$$\theta = \frac{1}{2} \arctan\left(\frac{2\mu'_{11}}{\mu'_{20} - \mu'_{02}}\right) \tag{4}$$

However, despite the randomness of the colony confluence patterns, they often tend to cluster in orientations already roughly aligned to the original streaking path. Moreover, colonies have reflections that have a clear horizontal orientation that is lost after rotation. Thus, enforcing rotation invariance can alter this property, leading to a negative effect in terms of classification performance.

All the enhancements have been applied only to the *masked* dataset, which has been chosen as the reference *baseline* for its demonstrated higher performance compared to the *bounding box* dataset.

5. Cardinality classification and outlier rejection

Three different methods for colonies enumeration inside segments are described in the following. First a CNN-based classification solution is designed to recognize segments cardinality and the presence of outliers. The second method consists in a classification pipeline based on handcrafted features extraction for the same purposes. Finally, a conventional watershed-based image processing approach is outlined.

5.1. CNN classification

The proposed CNN-based colonies classifier has been experimentally determined and configured deriving the best model by an accurate cross-validation process. Similar to [17], our CNN topology is composed of five learned layers, four convolutional and one fully connected, as shown in Fig. 6:

- 1st conv. layer, 20 feature maps with filter size 5×5;
- 2nd conv. layer, 50 feature maps with filter size 5×5;
- 3rd conv. layer, 100 feature maps with filter size 4×4;
- 4th conv. layer, 200 feature maps with filter size 4×4;
- fully connected layer, 500 hidden units;
- soft-max output layer.

The CNN has been implemented within the BVLC Caffe [35] framework of UC Berkeley which is conceived to facilitate the design, exploration and implementation of CNNs and other Deep Neural Networks, while providing computationally effective software solutions which make applications suitable for both investigative and industrial exploitations. Caffe models and optimization are defined by plain text schemes for ease of testing. In order to

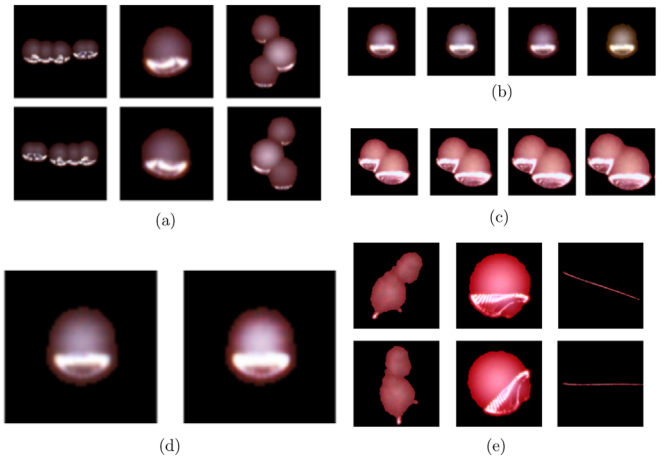

Fig. 5. Examples of dataset augmentation techniques, *horizontal flip* (a), *artificial color distortion* (b), *rescaling* (c), *Contrast Adaptive Histogram Equalization* (d), and *rotation invariance* (e).

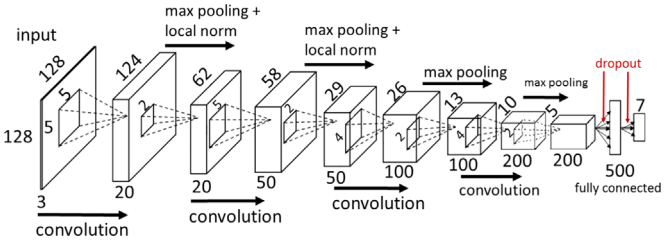

Fig. 6. The proposed Convolutional Neural Network topology.

speed up learning convergence, non-saturating non-linearities $f(x) = \max(0, x)$, also referred as ReLU activation function [36], are adopted on all layers. Deep convolutional neural networks with ReLU learn way faster than their sigmoid counterparts. The output of the convolutional layers after passing through ReLU non-linearities is normalized by means of Local Response Normalization [18], then downsampled with non-overlapping max-pooling. For reducing over-fitting on the two fully connected stages, a random dropout (cross-validated at the 75% of the weights) is adopted on them [37]. Networks weights are initialized following an initialization scheme, that in the Caffe framework is called *xavier* [38,39], which initializes each weight drawing their value from a uniform probability distribution on the interval $[-a, a]$ where

$$a = \sqrt{\frac{3}{\text{fan}_{\text{in}}}} \tag{5}$$

and fan_{in} is the number of input nodes, while the bias values are initialized as constants. The learning phase has been configured as follows:

- training is performed with Stochastic Gradient Descent (SGD) with batch size 64;
- for regularizing, weight decay is set to 0.0005;
- learning rate is initialized to 0.01 and it is decreased of the 0.01% at each iteration;
- training is performed at first applying momentum at 0.9.

Choosing the proper learning rate can be fairly difficult. One standard method that works well in practice is to use a small enough learning rate that gives stable convergence in the initial epoch (full pass through the training set) and then halve the value of the learning rate as convergence slows down. To prevent sharp fluctuations in the learning curves and then to stabilize the accuracy curves, the learning rate has been halved at 20,000 and 40,000 iterations. This demonstrated good convergence to a local minima of loss function. SGD can lead to very slow convergence particularly after the initial steep gains. Momentum $\gamma(0, 1]$ is one method for pushing the objective more quickly along the shallow ravine and it has been set to 0.9.

Performances evaluation have been done by randomly splitting the dataset into 70%/30% training/testing. Cross-validation has been done by retaining the 30% of the training set apart during model selection.

5.2. Handcrafted features classification

A conventional feature-based classification pipeline has been carefully designed, implemented and tested for a direct comparison of the results obtained, for the same task and on the same dataset, with the CNN approach. Such method is composed of a hand-crafted feature extraction phase, followed by a classification phase. The meaningfully selected features extracted on each image segment of our database are:

- Elliptic Fourier Descriptors (EFD) [40] of the segment shape up to the 50th order, for a total of 200 features;
- color histogram of the segment in the CIE-Lab pixel color space: eight bins for each channel for a total of 512 features (8^3);
- segments surface area.

Elliptic Fourier descriptors [18] are a Fourier representation of a chain code approximation of closed continuous contours [19]. Kuhl has proposed a method for deriving a chain code from a binary connected component [20]. Elliptic Fourier descriptors of a binary segment represent its shape. By taking the most significant orders of the descriptors, they will give a compact and approximate representation of the shape. The approximation error will depend upon the order of the descriptors that are retained. The higher the order, the better the approximation. Each order corresponds to 4 coefficients. Elliptic Fourier descriptors have some desirable properties:

- rotation invariance;
- scale invariance;
- shift invariance, directly from chain code representation;
- shape regularization (the approximation cleans small shape artifacts);
- compact representation (inherent reduction of the feature space dimensionality).

The original feature vector is 713-dimensional. Feature reduction was performed by PCA with the following arrangements. Elliptic Fourier descriptors coefficients are meaningfully comparable to each other, but their values are not comparable to color histogram features values, since they measure different properties, so they are not inter-scaled properly. For this reason, it would not make sense to process them together with PCA that is going to be unbalanced due to the differences in their scales, thus in their variances. Performing a features-wise scaling before PCA would disrupt relations of the features within the same data group type. Two separated PCA are performed on two types of data without prior scaling. Then, the obtained features are aggregated together and normalized with respect to their average and their variance, for avoiding a skewed features space, that can be harmful for the classifier learning. Random forests are used for features selection [23–25]. Despite random forests are mostly known as classifiers, since they build on the concept of features importance they can be

really effective also when exploited as feature selectors. Only features that have an importance bigger than a certain threshold, e.g. 5%, are kept, while the others are discarded [41]. Features preprocessed in the described way are then classified by a Support Vector Machine approach with Radial Basis Function kernels [42]. The optimal hyper-parameters are chosen by means of model selection on multiple iterations. Each iteration has a different shuffled dataset splitting between training and testing set, always keeping a 70%/30% proportion between training and testing sets respectively. A grid parameters exploration searches for the best σ of the basis function and C regularization parameters. Cross-validation is performed on a validation set, obtained from 30% of the training set.

5.3. Watershed based colony separation

For further comparison, a conventional morphological image processing method was also considered: a distance transform applied to the binary masks of the segments combined with watershed transform has been implemented to separate aggregated colonies. Watershed is a classical algorithm used for clumps splitting. Starting from the local maxima of the distance transform which are selected as markers, the watershed algorithm treats pixels values as a local topographic (elevation) map. The algorithm floods basins from the markers, until basins attributed to different markers meet on watershed lines. To allow a complete performance comparison, a segment will be considered as an outlier if the separated objects are more than six, i.e. classes not present in the dataset.

6. Results

The segments dataset has been randomly split into 70% for training and 30% for testing to evaluate the performance of the implemented systems. 30% of the training set has been initially retained for model cross-validation and, once the best topology has been defined, it has been included in the final training set. All the experiments shared the same testing set. For the learning based approaches, dataset enhancement techniques have been applied on the training set. The results in Section 6.1 are related to the CNN classification performance on the baseline datasets (i.e. the *masked* and the *bounding box* ones) and on the enhanced ones. In Section 6.1.1 the various dataset augmentations and normalizations have been applied and tested only to the *masked dataset*, given the better performance obtained with this one. The results for each test are presented in terms of accuracy (% of correctly classified samples) and loss function. In Section 6.1.2 we tested the use of a SVM classification stage in place of the fully connected neural network of the original CNN architecture. In Section 6.1.3 we obtained a learning curve of our system by testing different dataset splitting. For the best performing CNN case in terms of overall accuracy, we computed other figures of merit, i.e. confusion matrix, precision and recall [43] for each class and an original *per-colony error* metric. These are used in Section 6.2 for a performance comparison against the handcrafted feature classification system and watershed based colony separation. Also in these cases, results are presented considering the configurations that offer the best performance.

6.1. CNN results

6.1.1. Baseline configurations and dataset enhancements

Training log loss and test accuracy curves with respect to the number of training iterations are shown in Fig. 7. Fig. 7(a) represents the training curves of the baseline *masked dataset* (with no enhancements), in this case 91.5% of accuracy is obtained. The *resized* dataset provides an accuracy gain of 0.5 percentage points (Fig. 7(b)) compared to the baseline. No significant accuracy gain is brought with the *artificial color distortions* augmentation (Fig. 7(c)), and the same holds for CLAHE (Fig. 7(d)). The *horizontal flip* dataset augmentation (Fig. 7(e)) produces the higher performance gain, achieving an overall accuracy of 92.1%. Conversely, a theoretically promising transform such as the *normalization with respect to the segment orientation* does not achieve the desired effect (Fig. 7(f)). This is due to the presence of reflection patterns on the colonies which result to be misplaced on the rotated images, producing a counterproductive source of variation. Also the combination of the best singularly performing augmentations, i.e. the *horizontal flip* and the two different values of spatial *resize*, does not produce any benefit, leading instead to a slight performance worsening (Fig. 7(g)). This is probably caused by an increased information 'noise' introduced in the classifier. The *bounding box dataset* (Fig. 7(h)) reaches 90.5% of accuracy, a lower value than in the previous cases (likely due to the disturbing residual traces of nearby colonies that may remain due to the simple bounding-box handling, see Fig. 4, 3rd line), nevertheless showing the robustness and flexibility of CNN in the addressed classification task.

In all the above settings the testing accuracy increases with the number of training iterations and flattens around 30,000 iterations, except for residual fluctuations due to the physiological variations of the network weights. The same holds for the loss function, which tends to zero. This means that the learned features are discriminative for this type of task. 50,000 iterations take approximately one hour on a Nvidia Titan X GPU for the considered dataset versions.

In Fig. 8, the confusion matrix of the best performing model, obtained from the masked dataset with horizontal flip, is presented. The confusion matrix shows that, even if the classes corresponding to cluster composed of 3–6 colonies are sometimes misclassified, the wrongly selected labels remain close to the main diagonal, e.g. cluster composed of 5 colonies are rather confused by clusters of 4 or 6 colonies. It has also been said that the discrimination of those aggregates is fairly often hard even for a trained technician. An illustrative and well representative selection of classification results produced by our reference model is shown in Fig. 9 which also comprises meaningful misclassification examples. The most frequent errors occur when a tiny micro-colony is attached to a normal sized single colony: the classifier can confuse the tiny micro-colony as clutter or noise and give as result a single colony. The CNN classifier effectively detected outliers, segments that are often really difficult to distinguish with standard image and feature analysis techniques.

As a final remark, observing the substantial symmetry of the confusion matrix, favorable error compensation effects are likely to occur in case the system want to be used as a core module within a whole plate CFU counting system.[2] A simple way to visualize classification on agar plates is shown in Fig. 10, where the different cardinality bacterial aggregates are assigned to their specific class color.

[2] In this work we do not consider the complete set of application requirements of whole plate colony counting, because in the frequent case of limited bacterial load (i.e. up to 80–100 CFUs, where quite well separated colonies and colony aggregates occur, as in Fig. 1) an accurate count is highly needed and the extension of the proposed system to whole plate counting is trivial, while, in case of higher bacterial load (i.e. over estimated 100 CFUs), highly confluent growth areas occur, where usually CFUs are no more distinguishable, and where completely different estimation criteria (usually driven by clinical guidelines) and less accurate CFU quantifications are required which are out of the scope of the present work.

6.1.2. Use of SVM classification in place of the fully connected NN

In order to test an alternative CNN+classifier configuration, instead of the 500 nodes fully connected NN, we put downstream of the last convolutional layer a small fully connected layer, with number of neurons equal to the number of classes and without non-linearities, feeding a *Hinge Loss* layer. Learning this architecture is equivalent to work with a linear SVM acting on features learned by the CNN, similar to what was proposed in [44]. The results in terms of accuracy are lower by about 10% as shown in Fig. 7(i). The cause of this gap can be found in the linearity of the SVM kernel (the only workable option due to the feature cardinality) against the depth and non-linearity of the original architecture which help to disentangle the complexity of the underlying manifold (a hinge loss function on top of convolutional features suffers of under-fitting).

6.1.3. Dataset splitting and learning curve

The 70/30 dataset splitting has been used for each augmentation/normalization test. Direct comparison among different splitting solutions has the drawback to change the test set, so a fixed testing set solution with variable training set dimensions has been tested. By observing the network accuracy for an increasing portion (20, 40, 60, 80 and 100%) of the training set (see Fig. 11(a)), we can understand to which extent a greater number of samples in the training set increase performance. The residual gap between test and training accuracies (we always used the best performing topology and 50k iterations) suggests the presence of overfitting [45] since the classifier performance on the training set do not perfectly generalize on the test set. A main reason of this could be the skewness of the classes in our dataset. This is also confirmed by looking at the asymptotic behavior of the learning curves (Fig. 11(b)) which predicts a little performance gain with respect to

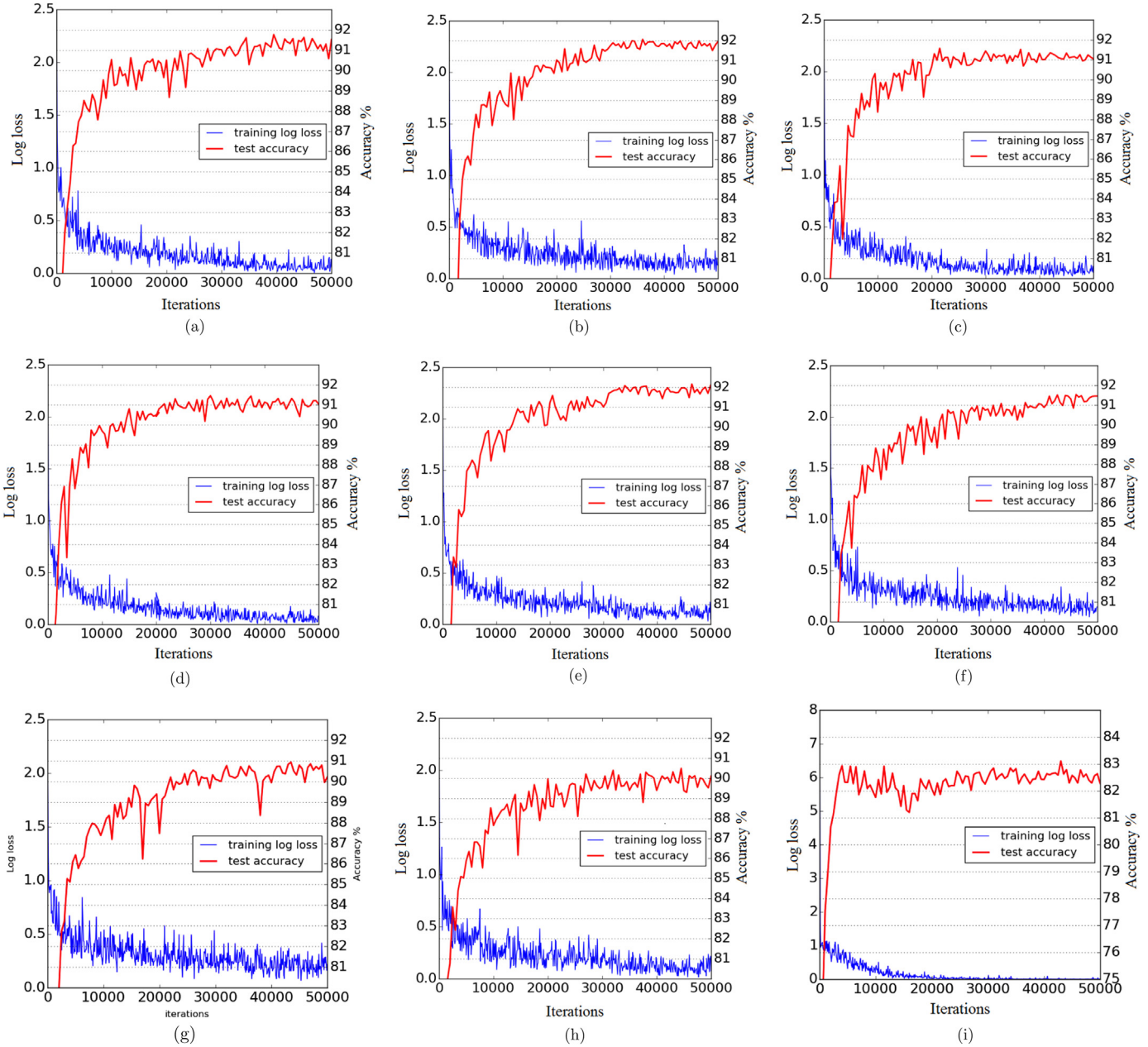

Fig. 7. CNN results: log loss function and test accuracy curve of the tested datasets: baseline masked dataset (a), resized dataset (b), lab color distorted dataset (c), baseline dataset with CLAHE contrast enhancement (d), horizontal flip dataset (e), rotation invariance dataset (f), combination of horizontal flip and resize dataset (g) and bounding box dataset (h). In (i) the same curves for the CNN+SVM configuration.

a class statistics preserving insertion of additional samples, therefore suggesting to add quality and informative new samples of less represented (because less probable) classes. This is exactly the reason why while extending our dataset from the initial one used in [5] we also tried to compensate the class skewness. This leads to a performance improvement, in fact we did not observe performance degradations despite the more challenging dataset due to an increment in higher rank samples. For this reason, despite the absolute classification performance that are already satisfactory, we will try to expand and update our database as soon as new bunch of data will be available.

6.2. Performance comparison

The obtained results are presented considering the configurations that offer the best performance. For hand-crafted features, this is the case of the dataset with combined flip and rescaling augmentations, while for the watershed counting method we only consider the binary mask of the segments (no learning is involved). The hand-crafted features classification pipeline (which confusion matrix is reported in Fig. 12) leads to an overall accuracy of 79.5%, corresponding to an overall error of 21.5%. This means that the CNN-based approach halves the error compared to its competing approach. Fig. 13 shows the confusion matrix for the watershed counting method that leads to an overall accuracy drop to 69%. In both cases the confusion matrix is no longer concentrated on the diagonal as it was for the CNN case. A more detailed analysis providing precision and recall coefficients for each class is presented in Table 1: again the performance gain in using the CNN approach is evident, with a particularly significant improvement for multiple CFUs aggregates. Eventually, a *per-colony* error metric is used to measure the counting accuracy and to further compare the considered techniques. This parameter is defined as

$$Err = \sqrt{\frac{1}{C}\sum_{s\in S}(c_s - \bar{c_s})^2} \tag{6}$$

where S is the test set, c_s the count for the sample s, considered empty if it is an outlier, $\bar{c_s}$ is the estimated count, C is the sum of the counts of all the aggregates in the test set. Results for the CNN technique on the best scoring database and for the two other techniques are shown in Table 2. Again it is evident how the proposed CNN approach drastically improves also this error metric.

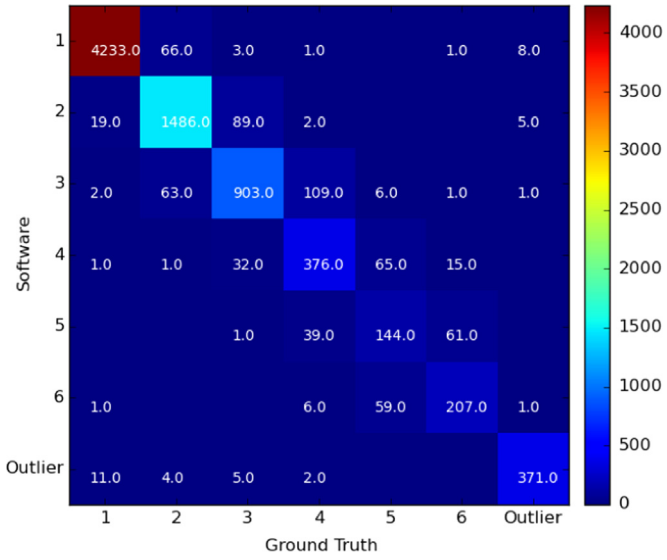

Fig. 8. CNN results: confusion matrix for the *horizontal flip* dataset.

7. Conclusion

While we are assisting to the thriving diffusion of Full Laboratory Automation in Clinical Microbiology which generates high expectations for advanced image interpretation solutions (capable to solve challenging visual tasks and to contribute to the early diagnosis and quantification of infectious diseases), at the same time Deep Learning solutions are indeed demonstrating disruptive performance for some challenging classes of image interpretation problems (also addressing many open issues coming

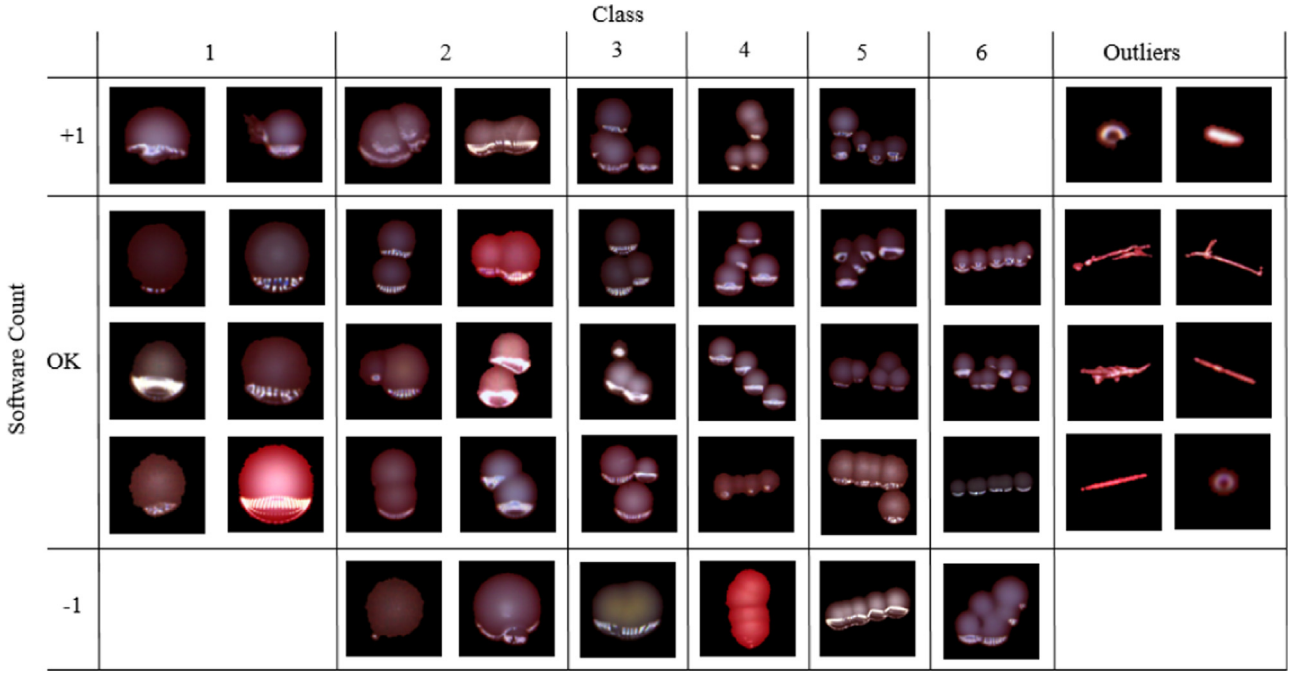

Fig. 9. CNN results: visual examples of counting issued by the CNN classification comprising correct and misclassified outcomes.

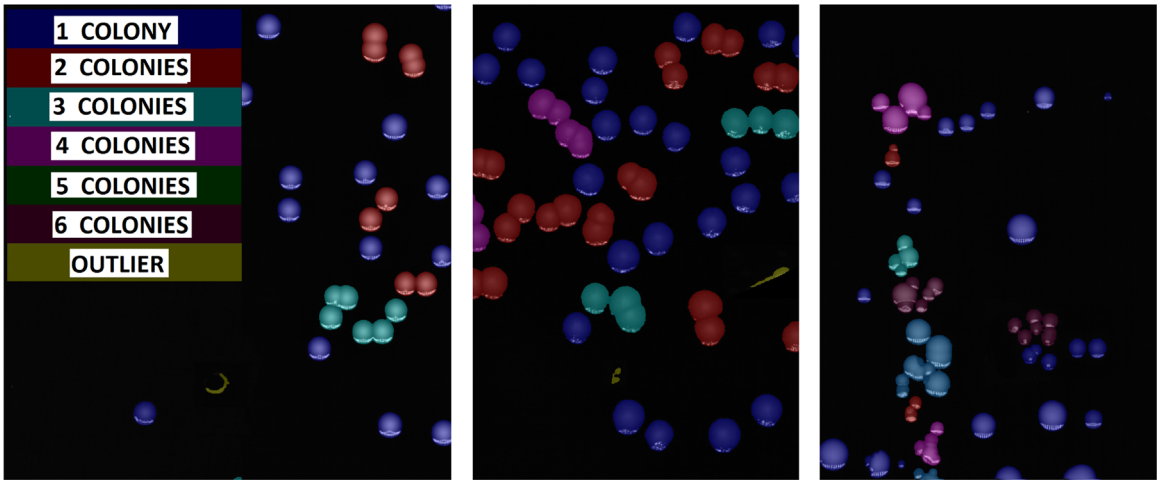

Fig. 10. Example of classification visualization.

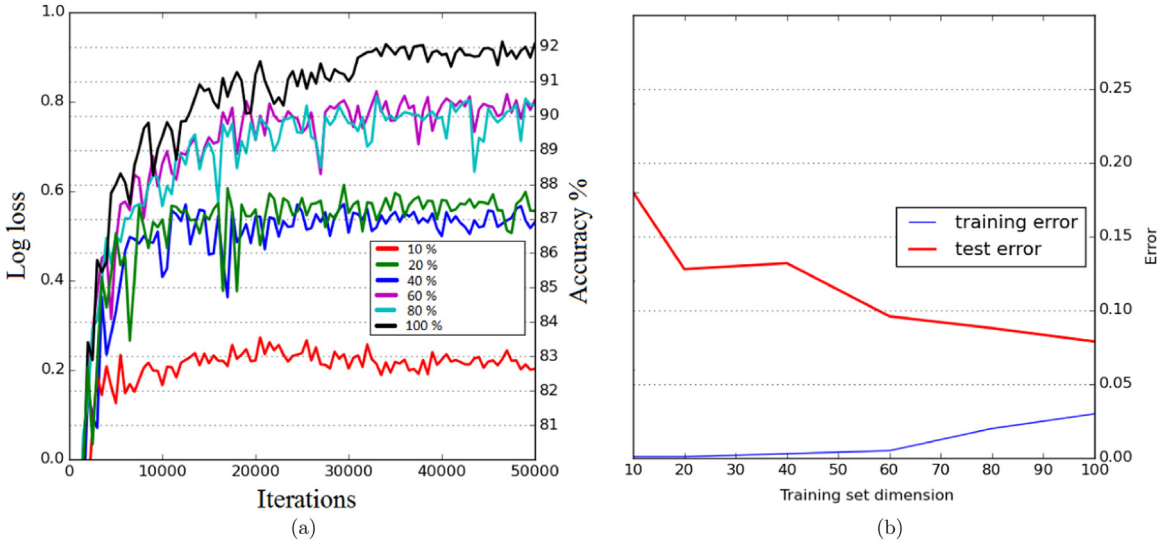

Fig. 11. Accuracy trends vs no. of iterations (a) and learning curve (b) for increasing portions of the split training set.

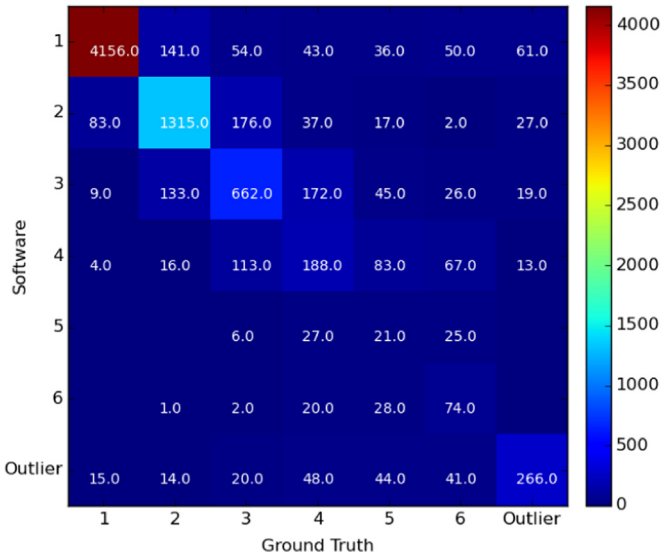

Fig. 12. Hand-crafted features results: confusion matrix for classification with hand crafted features.

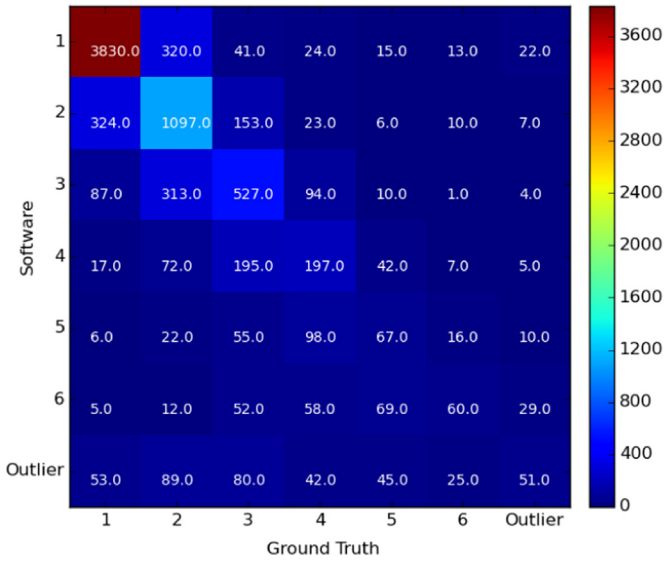

Fig. 13. Watershed results: confusion matrix for classification with watershed method.

Table 1
Precision and Recall comparison.

Class	CNN		HCF		Watershed	
	Precision	Recall	Precision	Recall	Precision	Recall
1 colony	0.98	0.99	0.91	0.97	0.89	0.87
2 colonies	0.93	0.92	0.79	0.81	0.67	0.57
3 colonies	0.83	0.88	0.62	0.71	0.51	0.48
4 colonies	0.77	0.70	0.38	0.35	0.37	0.37
5 colonies	0.59	0.44	0.26	0.08	0.24	0.26
6 colonies	0.71	0.73	0.59	0.26	0.21	0.45
Outlier	0.94	0.96	0.59	0.69	0.13	0.40

Table 2
Per-colony error comparison.

	CNN	HCF	Watershed
Per-colony error	**0.28**	0.80	0.88

from diversified biomedical imaging domains). The complexity of the colony counting task has been well recognized and addressed many times in the literature. Its nature is here well represented and documented by the dataset of labeled image segments we built and made available, containing a large number of isolated as well as aggregate CFUs coming from clinical bacterial cultures on blood agar plates related to urinary tract infection screenings. Building on this large database and on the configuration and exploitation of a CNN-based classification architecture we addressed bacterial colony counting obtaining a performance boost with respect to other carefully designed and reference solutions. Very good outlier rejection performances have also been achieved. Analyzing the obtained class confusion matrix, the level of nearby class confusion rarely exceeds the value of one step, while both classification accuracy and per-colony error results evidence a large margin superiority of the proposed deep learning solution. The proposed solution can be seen as a core component of a system which is capable to offer accurate counts with reliable outlier rejection. However, in a CML perspective, there is the need of a tool capable to provide a quantification on an extended range of bacterial load. Therefore what can be easily provided by our solution (i.e. an accurate count on plates with up to 80–100 CFUs) would need to be complemented by other analysis tools able to detect extended zones of confluent growth (typically where the CFUs are no longer visually discernible) and at least to handle coarse bacterial load quantifications on these areas, according to clinical requirements and guidelines. This is left to more implementation oriented works.

Acknowledgment

This work was partially supported by the Italian Ministry of Education, Universities and Research (MIUR) under the “Smart Factory Cluster” initiative, Adaptive Manufacturing Project: CTN01 00163 216730. The authors would also like to express their sincere thanks to the scientific and technical staff of Copan SpA (Brescia, Italy) for their essential support in providing material for the creation of the image database.

References

[1] R.E. Williams, R.E. Trotman, Automation in diagnostic bacteriology, J. Clin. Pathol. s2–3 (1) (1969) 8–13.
[2] P.P. Bourbeau, N.A. Ledeboer, Automation in clinical microbiology, J. Clin. Microbiol. 51 (6) (2013) 1658–1665.
[3] S. Novak, E. Marlowe, Automation in the clinical microbiology laboratory, Clin. Lab. Med. 33 (3) (2013) 567–588.
[4] C.D. Doern, M. Holfelder, Automation and design of the clinical microbiology laboratory, in: Manual of Clinical Microbiology, 11th edition, ASM Press, Washington, DC, 2015, pp. 44–53.
[5] A. Ferrari, S. Lombardi, A. Signoroni, Bacterial colony counting by convolutional neural networks, in: 37th Annual international Conference of the IEEE Engineering in Medicine and Biology Society EMBC 2015, 2015, pp. 7458–7461.
[6] S. Brugger, C. Baumberger, M. Jost, W. Jenni, U. Brugger, K. Mühlemann, Automated counting of bacterial colony forming units on agar plates, PLoS One 7 (3) (2012) e33695.
[7] C. Zhang, W. Chen, W. Liu, C. Chen, An automated bacterial colony counting system, in: IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC 2008), 11–13 June 2008, Taichung, Taiwan, 2008, pp. 233–240.
[8] H.P. Mansberg, Automatic particle and bacterial colony counter, Science 126 (3278) (1957) 823–827.
[9] H.E. Kunitschek, Electronic counting and sizing of bacteria, Nature 182 (4630) (1958) 234–235.
[10] N. Alexander, D. Glick, Automatic counting of bacterial cultures—a new machine, IRE Trans. Med. Electron. PGME-12 (1958) 89–92.
[11] D. Mukherjee, Bacterial colony counting using distance transform, Int. J. Biomed. Comput. 38 (1995) 131–140.
[12] G. Corkidi, R. Diaz-Uribe, J. Folch-Mallol, J. Nieto-Sotelo, Covasiam: an image analysis method that allows detection of confluent microbial colonies and colonies of various sizes for automated counting, Appl. Environ. Microbiol. 64 (4) (1998) 1400–1404.
[13] A. Liu, Z. Liu, L. Song, D. Han, Adaptive ideal image reconstruction for bacteria colony detection, in: E. Zhu, S. Sambath (Eds.), Information Technology and Agricultural Engineering, Advances in Intelligent and Soft Computing, vol. 134, Springer, Berlin, Heidelberg, 2012, pp. 353–360.
[14] A.B.G.L. Masala, U. Bottigli, Automatic cell colony counting by region-growing approach, Il Nuovo Cimento C (2008) 633–644.
[15] Q. Geissmann, Opencfu, a new free and open-source software to count cell colonies and other circular objects, PLoS One 8 (2) (2013) e54072.
[16] A. Ferrari, A. Signoroni, Multistage classification for bacterial colonies recognition on solid agar images, in: 2014 IEEE International Conference on Imaging Systems and Techniques (IST), 2014, pp. 101–106.
[17] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[18] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., Red Hook, NY, 2012, pp. 1097–1105.
[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, CoRR abs/1409.4842 (2014), http://dx.doi.org/10.1109/CVPR.2015.7298594.
[20] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, CoRR abs/1311.2524 (2013), http://dx.doi.org/10.1109/CVPR.2014.81.
[21] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, P. Barbano, Toward automatic phenotyping of developing embryos from videos, IEEE Trans. Image Process. 14 (9) (2005) 1360–1371.
[22] D. Ciresan, A. Giusti, L. Gambardella, J. Schmidhuber, Deep neural networks segment neuronal membranes in electron microscopy images, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., Red Hook, NY, 2012, pp. 2843–2851.
[23] D. Ciresan, A. Giusti, L. Gambardella, J. Schmidhuber, Mitosis detection in breast cancer histology images with deep neural networks, in: K. Mori, I. Sakuma, Y. Sato, C. Barillot, N. Navab (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2013, Lecture Notes in Computer Science, vol. 8150, Springer, Berlin, Heidelberg, 2013, pp. 411–418.
[24] D.B.A. Shkolyar, A. Gefen, H. Greenspan, Automatic detection of cell divisions (mitosis) in live-imaging microscopy images using convolutional neural networks, in: 37th Annual international Conference of the IEEE Engineering in Medicine and Biology Society EMBC 2015, 2015, pp. 743–746.
[25] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, CoRR abs/1505.04597 (2015), http://dx.doi.org/10.1007/978-3-319-24574-4_28.
[26] H. Su, F. Liu, Y. Xie, F. Xing, S. Meyyappan, L. Yang, Region segmentation in histopathological breast cancer images using deep convolutional neural network, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015, pp. 55–58.
[27] Y. Anavi, H.K. Greenspan, I. Kogan, E. Gelbart, O. Geva, A comparative study for chest radiograph image retrieval using binary, texture and deep learning classification, in: 37th Annual international Conference of the IEEE Engineering in Medicine and Biology Society EMBC 2015, 2015, pp. 797–800.
[28] J. Arevalo, F.A. Gonzalez Osorio, R. Ramos-Pollan, J.L. Oliveira, M.A. Guevara Lpez, Convolutional neural networks for mammography mass lesion classification, in: 37th Annual international Conference of the IEEE Engineering in Medicine and Biology Society EMBC 2015, 2015, pp. 797–800.
[29] H. Roth, J. Yao, L. Lu, J. Stieger, J. Burns, R. Summers, Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications, in: Recent Advances in Computational Methods and Clinical Applications for Spine Imaging, Lecture Notes in Computational Vision and Biomechanics, vol. 20, Springer International Publishing, Switzerland, 2015,

pp. 3–12.
[30] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, M. Nielsen, Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network, in: Medical Image Computing and Computer-Assisted Intervention MICCAI 2013, Lecture Notes in Computer Science, vol. 8150, Springer, Berlin, Heidelberg, 2013, pp. 246–253.
[31] M.L. Wilson, L. Gaido, Laboratory diagnosis of urinary tract infections in adult patients, Clin. Infect. Dis. 38 (8) (2004) 1150–1158.
[32] R.C. Gonzalez, R.E. Woods, Digital Image Processing, 3rd edition, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.
[33] A. Ferrari, A. Signoroni, Multistage classification for bacterial colonies recognition on solid agar images, in: 2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings, 2014, pp. 101–106.
[34] K. Zuiderveld, Contrast limited adaptive histogram equalization in: P.S. Heckbert (Ed.), Graphics Gems IV, Academic Press Professional, Inc., San Diego, CA, 1994, pp. 474–485.
[35] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, CoRR abs/1408.5093 (2014), http://dx.doi.org/10.1145/2647868.2654889.
[36] V. Nair, G. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21–24, 2010, Haifa, Israel, 2010, pp. 807–814.
[37] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, CoRR abs/1207.0580 (2012).
[38] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: JMLR W& CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010), vol. 9, 2010, pp. 249–256.
[39] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.
[40] F. Kuhl, C. Giardina, Elliptic fourier features of a closed contour, Comput. Graph. Image Process. 18 (3) (1982) 236–258.
[41] G. Louppe, P. Geurts, Ensembles on random patches, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24–28, 2012, Proceedings, Part I, Springer, Berlin, Heidelberg, 2012, pp. 346–361.
[42] B. Schölkopf, K.-K. Sung, C.J. Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik, Comparing support vector machines with Gaussian kernels to radial basis function classifiers, IEEE Trans. Signal Process. 45 (11) (1997) 2758–2765.
[43] D.M.W. Powers, Evaluation: from precision, recall and f-measure to roc., informedness, markedness & correlation, J. Mach. Learn. Technol. 2 (1) (2011) 37–63.
[44] Y. Tang, Deep Learning with Linear Support Vector Machines, In: Workshop on Representational Learning, ICML 2013, Atlanta, USA, 2013, http://deeplearning.net/icml2013-workshop-competition.
[45] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Comput. 4 (1) (1992) 1–58.

Alessandro Ferrari graduated in Telecommunication Engineering (University of Brescia, Italy, in 2012). In 2016 he got a Ph.D. degree. in Telecommunication Engineering (University of Brescia) in apprenticeship with Futura Science Park, Copan Italia S.p.a, Brescia. His research interests include machine learning, deep learning and computer vision applied to clinical microbiology.

Stefano Lombardi received a M.Sc. degree in Communication Technologies and Multimedia (University of Brescia, Italy, in 2015). He is a Research Fellow with the Information Engineering Department (University of Brescia). His research interests include biomedical image analysis and machine learning.

Alberto Signoroni received M.Sc. in Electronic Engineering '97 and Ph.D. in Information Engineering '01 from the University of Brescia, Italy. Currently he is an Assistant Professor within the Signal Processing and Communications group, Information Engineering Department at the University of Brescia. His research interests include biomedical image analysis, multidimensional and hyperspectral image processing, 3D computer vision and geometry processing.