Assessing the Feasibility of a Spatio-Temporal Approach for Recognizing Microorganisms in Image Sequences

João Pedro Ribeiro da Silva
University of São Paulo
São Carlos, SP, Brazil
joaopedrorib@usp.br

Antonio Rafael Sabino Parmezan
University of São Paulo
São Carlos, SP, Brazil
parmezan@usp.br

ABSTRACT

Data classification is one of the main challenges in data mining and knowledge discovery. It is widely explored in multiple applications, such as identifying microorganisms by analyzing images of cultures grown in Petri dishes. This work proposes to chronologically organize images of bacteria on solid media, captured at equidistant intervals over several days, treating them as video frames to improve the recognition of these microorganisms. To this end, we develop an approach that chains classification models, integrating spatial features from pre-trained convolutional neural networks with temporal information propagated through target meta-features, *i.e.*, classifier outputs. We experimentally compared our proposal with two intentionally designed baseline methods using a dataset with 240 images, 48 per class, and considering the macro-averaged F1 score. Results demonstrate that addressing chronological relations enhances the identification models' performance, even though baseline strategies may benefit from fewer examples.

KEYWORDS

Digital Image Analysis, Feature Extraction, Deep Learning, Machine Learning.

1 INTRODUCTION

Data classification plays a central role in data mining and knowledge discovery, aiming to map the characteristics of examples to known classes and predict categories for new data. Researchers have extensively investigated this task, developing methods ranging from explainable models, such as rule-based systems, to more complex deep learning algorithms, whose models are often referred to as black boxes. These techniques are applied across various fields, including microbiology, enabling the recognition of microorganisms from Petri dish images, thereby advancing progress in biotechnology and pharmacology [1].

The automatic identification of microorganisms through culture images challenges the scientific community due to the visual similarity among different colonies, which can exhibit subtle characteristics such as shape, color, and texture [1, 2]. Despite this difficulty, recent studies have reduced predictive error rates to below 15% for images containing more than five bacterial species, highlighting advancements in the application of machine learning techniques [3].

In this paper, we propose to chronologically organize images of actinobacteria on solid media, captured over several days at equidistant intervals, treating them as video frames to improve the recognition of these microorganisms. To this end, we develop an architecture based on the chaining of classification models, which integrates spatial features extracted from pre-trained convolutional neural networks with temporal information obtained by propagating target meta-features, such as classifier outputs. Using a dataset of 240 images—48 per class—and employing the macro-averaged F1 score, we experimentally compared our approach with two intentionally designed baseline methods.

Our work mainly contributes by investigating how the chronological processing of images as frame sequences enhances the classification of individual images. By handling a high-quality microbial dataset, we also ensure more conservative results, increasing the reliability of our findings.

2 RELATED WORK

Automatically identifying microorganisms from macroscopic images is a promising research area. However, the lack of standardized data-gathering protocols impacts not only image quality and consistency but also annotation accuracy. Although studies follow traditional machine learning pipelines, specific settings are often adopted to address data variability [1, 3, 4].

Collecting microorganism images is challenging. Conventional approaches are less effective than more advanced methods when the number of available images is limited. Temporal convolutional networks require large datasets to learn temporal dependencies, making them unsuitable for small image collections. Few-shot learning algorithms, designed to operate on a few examples per class, provide a viable alternative in data-scarce environments [5]. However, advanced methods like Mamba could be effective for complex analysis if more images—frames per video—were available [6].

This work differs from existing literature by proposing a classifier chaining architecture that learns spatio-temporal features from images treated as video frames. As shown throughout the text, we achieve significant results using simple strategies despite the limited availability of images.

3 PROPOSED APPROACH

The task motivating this paper is macroscopic microbial image classification. Our dataset includes actinobacteria images gathered over eight consecutive days of growth. In addition to the spatial features of each image, this study also explores the temporal relationships between images of the same bacteria. By explicitly addressing chronological information by treating images as video frames, we developed a model-chaining architecture that automatically learns spatio-temporal features from videos to classify their respective frames.

3.1 Image Recognition Structure

Firstly, we conceived an end-to-end frame recognition module. Given an image, this structure can (i) extract spatial features, (ii)

select the most important attributes for classification, and (iii) discriminate examples into predefined labels. Fig. 1 illustrates the sequential flow between the submodules.

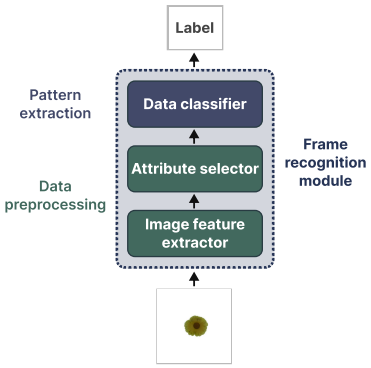

Figure 1: Recognition structure and its submodules.

The three submodules in Fig. 1 operate independently, allowing us to optimize each one individually for a given input image collection.

3.2 Model Chaining Architecture

We chain multiple frame recognition modules (Fig. 1), each specialized in frames from a specific video timestamp, to compose the proposed architecture (Fig. 2). For example, we train the first recognizer employing only the frames from the first timestamp. The structure extracts target meta-features from these frames and propagates them to the following module. This process continues until the last recognizer determines the class of all images of the sequence.

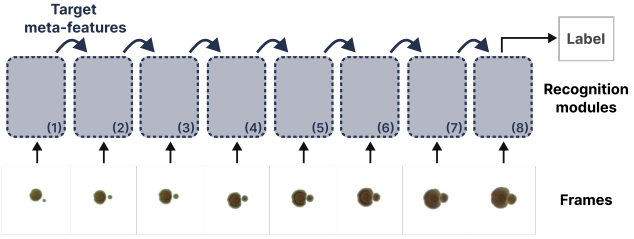

Figure 2: Proposed model-chaining architecture.

We explicitly embed the videos' sequential information by chaining various recognizers while capturing each frame's spatial characteristics. The extracted target meta-features are related to the classification of the frames, either through the labels assigned to them or the confidence scores given to each predefined class. Once generated, these target meta-features are propagated by concatenating them with the spatial attributes of the subsequent module. Finally, the last recognizer labels all frames, ensuring that classification accounts for the videos' chronological sequence.

4 MATERIAL AND METHODS

4.1 Dataset

This work uses a dataset consisting of macroscopic actinobacteria images. We obtained high-resolution photographs employing low-cost equipment [1]. We cultivated the bacteria in Petri dishes triplicate using the ISP2 culture medium. We photographed the samples' front and back sides from the 3rd to the 10th day after inoculation, corresponding to the activity period for actinobacteria [1, 2]. We employed five microbial species, generating a total of 240 images (5 species × 3 replicates × 8 days × 2 sides of the dishes). After capturing and labeling the photographs, we preprocessed and segmented them using methods developed by our research group [1]. The processed data supporting this study's findings are available from the first author upon reasonable request.

4.2 Baseline Methods for Comparison

We empirically compared our proposal with two other architectures: (i) a conventional classifier (Fig. 3) and (ii) an ensemble model (Fig. 4). Unlike our method, these baseline approaches do not extract temporal information. Therefore, the experiments aid in understanding how chronological relations within the videos impact frame classification.

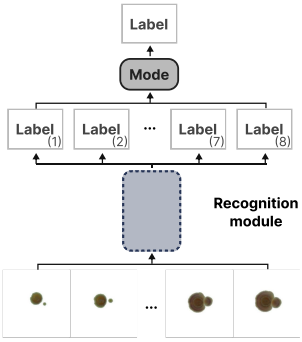

Figure 3: Conventional model.

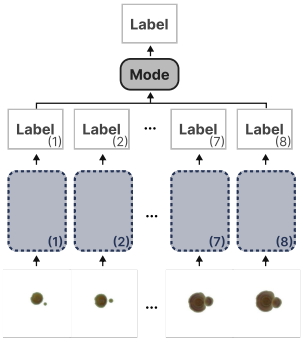

Figure 4: Ensemble model.

4.2.1 Conventional Classifier. This architecture employs a single recognition module to process all available frames without considering temporal information, either implicitly or explicitly. Thus, the recognizer focuses solely on the spatial features within the images, treating them as independent examples. Fig. 3 depicts the labeling process of a sequence of frames after fitting the model. Each image is classified individually, and the predicted labels are combined using a majority voting strategy.

4.2.2 Ensemble Model. This approach employs multiple recognizers, each assigned to a specific video timestamp. Hence, chronological information is implicitly addressed, with each module specializing in frames from a designated instant of time. Nevertheless, the temporal aspect is not explicitly presented to the learning algorithms, preventing them from extracting temporal patterns. Fig. 4 portrays the labeling process for this strategy. Each image is classified individually by its corresponding frame recognizer, and the sequence class is then determined using majority voting.

4.3 Experimental Setup

We organized the empirical evaluation into four steps: (i) dataset preparation, (ii) recognition module configuration, (iii) architecture building, and (iv) performance estimation.

4.3.1 Dataset Preparation. Building a high-quality image set proved to be labor-intensive and time-consuming. Consequently, our dataset lacked sufficient examples for directly extracting patterns employing popular learning algorithms within the context of this work. To enrich the dataset, we conducted two data augmentation strategies. First, we mirrored each image (i) horizontally, (ii) vertically, and (iii) both horizontally and vertically. This process resulted in three new instances per original record, although they remained similar to their unmodified versions. Second, we applied two transformations to the mirrored frame sequences: (i) random rotation by 45°, 135°, 225° and 315°, and (ii) random zoom-out with a scale factor between 1.2 and 1.5. We performed these modifications uniformly across all frames within a sequence to maintain continuity. We implemented these transformations using the PyTorch[1] library.

4.3.2 Recognition Module Configuration. Our frame recognition structure comprises three submodules (Section 3.1). As a spatial feature extractor, we adopted the convolutional neural network ResNet50, built in the PyTorch library and pre-trained on the ImageNet collection. We extracted 2048 scalar features from each image employing only its convolutional layers. We used a scikit-learn[2] ranking method for attribute selection based on the estimated mutual information between each feature and the target values [7]. This process led us to select the 512 best-ranked attributes, reducing the input vector dimensions by 75%. For pattern extraction, we considered five base classifiers: (i) K-Nearest Neighbors (KNN), (ii) Gaussian Naive Bayes (GNB), (iii) Random Forest (RF), (iv) eXtreme Gradient Boosting (XGB), and (v) MultiLayer Perceptron (MLP). Concerning RF, we adopted the default hyperparameter values recommended in [8]. For MLP, we employed a network with a single hidden layer containing 258 neurons, trained over 2000 epochs. We fitted the parameters via stochastic gradient descent with an initial learning rate of 0.03, which decayed to 0.003 at the 1500th epoch. We also used cross-entropy loss and a batch size of 32. For the remaining hyperparameters, we adhered to the values suggested by their respective implementation libraries: scikit-learn for KNN, GNB, and RF; XGBoost[3] for XGB; and PyTorch for MLP.

4.3.3 Architecture Building. We instantiated the proposed architecture (Section 3.2) and baseline approaches (Section 4.2). For our method, we defined the confidence scores associated with each class as the target meta-features and extracted them applying five-fold cross-validation.

4.3.4 Performance Estimation. We empirically evaluated the identification models through a cross-validation method similar to the leave-one-out technique. The process involved (i) setting aside all frames—both original and augmented, from the Petri dishes' front and back sides—associated with a microbial sample, (ii) training a model with the remaining examples, and (iii) evaluating the fitted model employing the original images from the set-aside. We settled the macro-averaged F1 score (F1) as the evaluation metric.

[1] https://pytorch.org (v. 2.5.0).
[2] https://scikit-learn.org (v. 1.2.2).
[3] https://xgboost.readthedocs.io/en/stable/python (v. 2.1.3).

5 RESULTS AND DISCUSSION

Table 1 summarizes the predictive performance of each researched approach alongside its respective base classifiers. We reported the average and standard deviation for each architecture (columns) and base classifier (rows).

Table 1: F1 for each combination of architecture and classifier.

Base classifier	Architecture			Average
	Proposed	Conventional	Ensemble	
KNN	0.701	**0.774**	0.650	**0.709** (0.051)
GNB	0.571	**0.729**	0.184	0.495 (0.229)
RF	0.663	**0.774**	0.545	0.661 (0.094)
XGB	0.601	**0.702**	0.671	0.658 (0.042)
MLP	0.593	0.540	**0.764**	0.632 (0.095)
Average	0.626 (0.048)	**0.704** (0.086)	0.563 (0.202)	

The conventional approach consistently delivered the best results, regardless of the base classifier (Table 1). The only exception occurred with MLP, where the ensemble architecture achieved the highest performance. Our proposal outperformed the ensemble approach on average but fell short of the conventional one. Regarding the base classifiers, KNN and RF achieved the best performances for the investigated task, with KNN demonstrating greater stability across different architectures.

6 CONCLUDING REMARKS

Comparing the performance of the proposed approach with the ensemble architecture, we conclude that explicitly incorporating chronological relations between frames enhances the models' average predictive ability. However, the conventional approach can achieve the best overall results by benefiting from fewer examples. As future work, we plan to address few-shot learning and alternative meta-feature propagation strategies to improve the proposed architecture's predictive power further.

ACKNOWLEDGMENTS

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Numbers #2024/07102-8, #2022/02176-8, and #2019/17721-9.

REFERENCES

[1] Antonio R. S. Parmezan, Danielle R. Gonçalves, and Solange O. Rezende. A unified framework for Petri dish image acquisition and processing to promote consistency in microorganism identification. In *arXiv*, 2025.
[2] Danielle R. Gonçalves, Antonio R. S. Parmezan, Lucianne F. P. Oliveira, Simone P. Lira, Roberto G. S. Berlinck, and Solange O. Rezende. Petri dish image-capturing guidelines for artificial intelligence-based microorganism recognition. In *CBM*, pages 615–1, 2023.
[3] Hedieh Sajedi et al. Actinobacterial strains recognition by machine learning methods. *Multimed. Tools Appl.*, 78:20285–20307, 2019.
[4] Alessandro Ferrari et al. Multistage classification for bacterial colonies recognition on solid agar images. In *IEEE IST*, pages 101–106, 2014.
[5] Flood Sung et al. Learning to compare: Relation network for few-shot learning. In *IEEE/CVF CVPR*, pages 1199–1208, 2018.
[6] Lianghui Zhu et al. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *ICML*, pages 62429–62442, 2024.
[7] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, 2004.
[8] João P. R. da Silva and Antonio R. S. Parmezan. Otimização de hiperparâmetros em modelos de classificação para dados de bactérias. In *SIICUSP*, pages 1–2, 2024.