Clustering and Classification of Fungal Cells and PMMA Microparticles: Unsupervised and Supervised Learning using K-Means, PCA, Logistic Regression and Spiking Neural Networks on Event-Based Cytometry Datasets

MUHAMMED GOUDA[1,*], STEVE ABREU[2], AND PETER BIENSTMAN[1]

[1] Photonics Research Group, Ghent University/imec, 9000 Gent, Belgium
[2] Bernoulli Institute, University of Groningen, Groningen, The Netherlands
[*] muhammedgoudaahmed.gouda@ugent.be

Compiled July 4, 2024

Imaging flow cytometry (IFC) is a powerful analytical technique used for rapidly categorizing small cells and micro-particles. Unlike traditional microscopy, IFC instruments handle a high throughput of cells. Typically, photodetectors, photomultiplier tubes, or high-speed frame-based cameras are employed for this task. This study explores the potential of neuromorphic cameras, also known as event-based sensors (EVS), as a detection mechanism for cytometers. While previous work focused on PMMA microparticles, this investigation centers on fungal cells. Moreover, while earlier research leaned towards supervised learning algorithms like logistic regression and spiking neural networks, our approach in this paper employs k-means, an unsupervised learning paradigm. We demonstrate that training such a simple algorithm in conjunction with PCA achieves 100% **classification accuracy without relying on training labels.**

1. INTRODUCTION

Flow cytometry, a technology focused on discerning cell or microparticle populations within a fluid, holds relevance across diverse domains including medicine, cosmetics, and environmental engineering [1]. Precision is paramount across these fields, necessitating highly accurate classification devices. To address this need for a broad spectrum of cells and particles, we integrate a novel imaging sensor—referred to as an event-based camera —with unsupervised learning techniques (specifically k-means) alongside principal component analysis (PCA) for dimensionality reduction.

The rest of this paper is structured as follows.

- In section 2, we explore the theory underlying the event-based sensor employed in our study.
- Section 3 elucidates the operational principles of k-means, the unsupervised learning technique utilized in our research, alongside principal component analysis, employed as a pre-processing step for dimensionality reduction.
- In sections 4 and 5, we present both fungal cells and PMMA samples separately, each serving distinct purposes in flow cytometry applications.
- Section 6 outlines the optical setup utilized in our study, detailing various components.
- Section 7 showcases the results derived from the application of k-means for clustering fungal cells and PMMA microparticles.
- Finally, section 8 concludes the paper, highlighting avenues for future research.

2. EVENT-BASED VISION

Event-based cameras, unlike traditional ones, do not capture consecutive static images at fixed rates [2]. Introduced by [3], these cameras use dynamic vision sensors with pixels reacting independently to scene changes. Each pixel triggers an event when light intensity exceeds a set threshold (see Figure 1). Adjusting this threshold is crucial; too low captures unwanted background events, while too high misses targeted events. For our application, it must be high enough to avoid laser speckle noise yet sensitive to passing particle diffraction patterns.

3. UNSUPERVISED LEARNING

As the name suggests, unsupervised learning is the approach of training a machine to perform a task without any sort of supervision. This means that, in contrast to supervised learning paradigms, no labels are provided during the training phase of the model development [4]. In this regard, researchers utilize what is known as clustering methods, which is basically classification without relying on training labels. Several clustering algorithms exist such as k-means, hierarchical clustering,

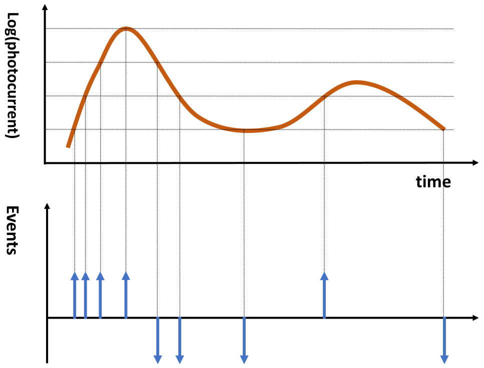

Fig. 1. Illustration of the working principle of an event-based pixel. The intensity of the light is sampled by the camera, and an event is fired whenever the intensity exceeds by a certain value. Based on whether there is an increase or a decrease in the intensity, a positive or a negative signal is recorded.

density-based clustering non-parametric algorithm (DBSCAN), Gaussian mixture models (GMM), self-organizing maps,.. .

In the context of this research, we focused on basic clustering algorithms starting with k-means[5]. To illustrate the idea behind it, consider the training points in figure 2. In this figure, the objective is to divide the points into clusters. The approach is to partition the data space in such a way that data points within the same cluster are as similar as possible (intra-class similarity), while ensuring that data points from different clusters are as dissimilar as possible (inter-class similarity) [6].

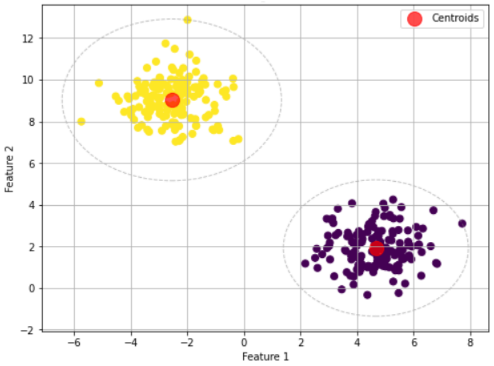

Fig. 2. Illustration of k-means. Two different clusters are shown. The goal is to minimize distances between samples in the same cluster and maximize the distances between different classes. K-means achieves this iteratively by choosing centroids and assigning samples to different classes.

In k-means, every cluster is represented by its centroid, i.e. the arithmetic mean of the data points assigned to that cluster. While a centroid denotes the center of the cluster (the mean), it doesn't have to be an actual member of the dataset. The algorithm iterates until each data point is closer to its cluster's centroid than to any other centroids, progressively minimizing intra-cluster distance [7]. K-means starts by selecting initial centroids arbitrarily and then iteratively adjusts them to converge on a final clustering of the data points:

- Initially, the algorithm randomly picks centroids for each cluster. For instance, with a "k" of 3, it selects 3 centroids randomly.
- K-means assigns each data point to the nearest centroid, effectively grouping them based on proximity.
- It recalculates centroids by averaging the points in each cluster, reducing intra-cluster variance. Since centroids change, points are reassigned to the nearest centroid.
- This process repeats until the total distance between data points and their respective centroids is minimized, reaching a maximum iteration limit, or when centroids no longer change.

Applying k-means in a high dimensional space could be challenging specially since k-means utilizes euclidean distances which works best on lower dimensional features. Therefore, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data. Its purpose is to map high-dimensional datasets into a lower dimensions, all the while preserving crucial patterns and trends [8].

Reducing variables in a dataset inevitably compromises accuracy, but the essence of dimensionality reduction lies in trading precision for simplicity. Shrinking data makes exploration and visualization easier, benefiting subsequent stages in the pipeline, which is k-means clustering in our case [9].

4. PMMA MICRO-PARTICLES

In our research, we initially utilized transparent PMMA (Polymethyl-methacrylate) microparticles spanning sizes from 2 μm to 20 μm from PolyAn GmbH [10]. These particles align with the size range of micro-plastics, offering an optimal testing ground for evaluating the viability of our system for this intriguing application.

5. BIOLOGICAL FUNGAL CELLS

We also considered experiments using fungal cells as an alternative to artificial PMMA spherical beads. These cells possess more complex shapes, presenting greater challenges for the developed machine learning algorithms. The following species of fungal cells have been employed, as depicted in Figures 3 and 4 in their respective dishes and individual cell forms under the optical microscope:

- *Aspergillus niger* (commonly known as black mold)
- *Yarrowia lipolytica* (a yeast used in industrial microbiology)

Both microorganisms are safe and straightforward to handle, requiring no specialized equipment. They were provided by CERTH (Center for Research and Technology Hellas [11]).

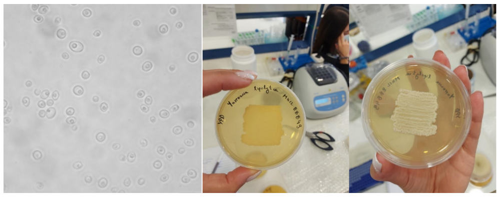

Fig. 3. Yarrowia lipolytica

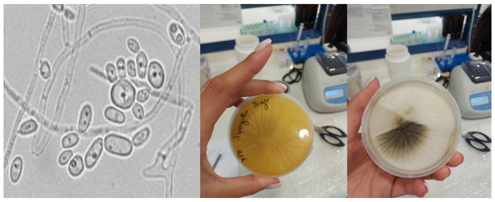

Fig. 4. Aspergillus niger

6. OPTICAL SETUP

The experimental setup is similar to that in [12]. It comprises a laser source that emits light with a wavelength of 632.8 nm. The light passes through a lens and a 25 µm pinhole before being focused onto a PMMA microfluidic channel.

The microfluidic channel utilized is from Chipshop (Fluidic 156) and is a straight-channel chip integrating four parallel channels where only one channel is used. The channel dimensions are 200 $\mu m \times 200\ \mu m \times 58.5\ mm$.

The channel allows the flowing of microparticles and biological fungal cells, which are pumped using a manual syringe pump connected to the upper port, while a liquid reservoir is connected to the other port, as shown in figure 5.

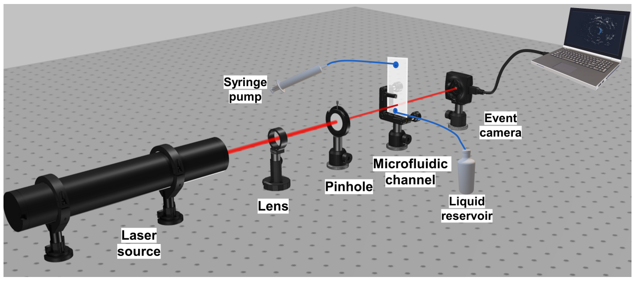

Fig. 5. The experimental setup built to generate the training and test datasets. Light coming from a 632.8 nm He-Ne laser is focused by a lens on a 25 μm pinhole. Behind the pinhole is a vertically-mounted PMMA microfluidic channel inside which microparticles are flowing downwards. The diffraction pattern caused by a flowing particle is captured by the event-based camera. The camera is connected to a laptop with dedicated software for recording the events fired at different pixels.

7. RESULTS

In this section we discuss the results obtained from applying k-means clustering on different tasks. The first task concerned measuring with the two classes of biological samples presented in Section 5 which are Yarrowia lipolytica and Aspergillus niger. We pumped the cells into the microfluidic channel and recorded the corresponding spikes generated by the event sensor over multiple measurement sessions. We followed a so-called intertwined measurement approach, with the goal to make sure that variations in the measurement conditions do not bias the machine learning algorithm. Therefore, the train dataset comes from three different sessions and the samples in testing dataset come from a fourth session. Those sessions were conducted at different times, thereby making sure that the measurement bias has minimal effect on the training. Histogram features were generated from the event data generated by the Inivation event-based camera citeinivation. This sensor has dimensions of 640X480 which results in a total number of features of 3,072 after downsampling by a factor of 10. We reduced the dataset to 1D, 2D and 3D spaces respectively. Figure 6 shows the results obtained for the fungal dataset after applying k-means on those lower dimensional spaces. As illustrated, the clusters were successfully recovered resulting in 100% accuracy. This shows that one can directly utilize our system even with cells that have unknown labels and accurately identify their groups.

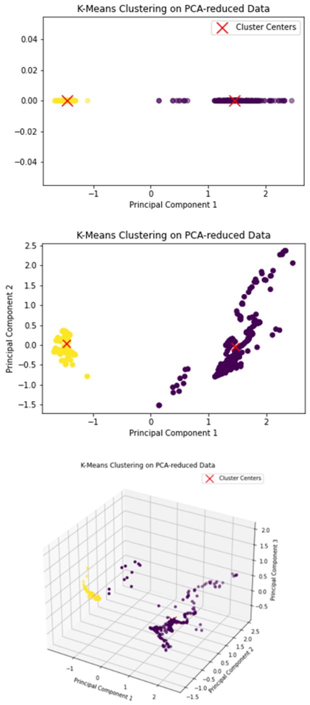

Fig. 6. Fungal cell testing dataset post k-means clustering, with Principal Component Analysis (PCA) reducing the dataset to 1D, 2D, and 3D spaces. Top to bottom: 1D, 2D, and 3D representations reveal accurately recognized clusters, achieving 100% accuracy without requiring training labels.

The task with microparticles instead of biological cells on the other hand was more challenging to solve. Figure 7 shows the samples from the original dataset after reducing them using PCA, similar to what has been done on the fungal cells dataset. As shown, the classes overlap significantly. This leads to k-means performing wrong assignments for a large number of samples as shown in figure 8.

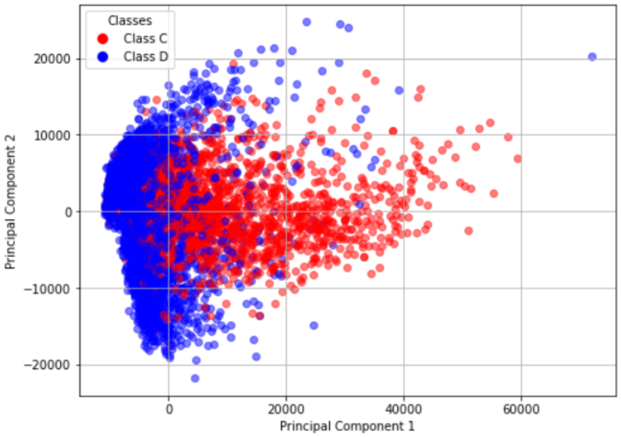

Fig. 7. True class labels for the original training dataset of event-based PMMA microparticles reduced to 2D prior to clustering. The significant overlap of clusters presents a challenging task for k-means clustering.

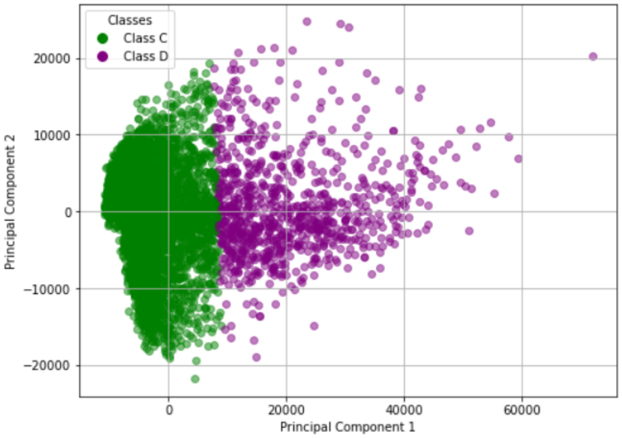

Fig. 8. Predicted classes labels after applying k-means clustering to the event-based PMMA microparticles training dataset. Unlike the biological cells task, the k-means approach failed to accurately reconstruct the clusters present in the training set for this particular task.

8. CONCLUSION

In conclusion, we explored various methods for classifying and clustering event-based datasets involving micro-particles and fungal cells. Unsupervised methods like k-means clustering and principal component analysis (PCA) demonstrated effectiveness only with fungal cells, but achieving 100% accuracy, due to the large difference in shape.

9. ACKNOWLEDGMENT

This work was performed in the context of the European projects Neoteric (grant agreement 871330), Prometheus (grant agreement 101070195), and in the Flemish FWO project G006020N and the Belgian EOS project G0H1422N.

REFERENCES

1. K. M. McKinnon, Curr. protocols immunology **120**, 5 (2018).
2. C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck, Proc. IEEE **102**, 1470 (2014).
3. P. Lichtsteiner, C. Posch, and T. Delbruck, IEEE journal solid-state circuits **43**, 566 (2008).
4. K. P. Sinaga and M.-S. Yang, IEEE access **8**, 80716 (2020).
5. P. Balakrishnan, M. C. Cooper, V. S. Jacob, and P. A. Lewis, Psychometrika **59**, 509 (1994).
6. C. Ding and X. He, “K-means clustering via principal component analysis,” in *Proceedings of the twenty-first international conference on Machine learning,* (2004), p. 29.
7. C. Boutsidis, P. Drineas, and M. W. Mahoney, Adv. neural information processing systems **22** (2009).
8. H. Abdi and L. J. Williams, Wiley interdisciplinary reviews: computational statistics **2**, 433 (2010).
9. L. Li, Towards Data Sci. (2019).
10. “Polyan gmbh,” https://www.poly-an.de/. Accessed: YYYY-MM-DD.
11. “Centre for Research and Technology Hellas (CERTH),” https://www.certh.gr/root.en.aspx (Accessed 2024).
12. M. Gouda, A. Lugnan, J. Dambre, *et al.*, IEEE J. Sel. Top. Quantum Electron. **29**, 1 (2023).