Indexed in Clarivate Analytics WoS, and ScienceOPEN

E-B Environment - Behaviour Proceedings Journal

ITMAHuB2023

https://kppim.uitm.edu.my/research/conferences/itmahub-2023

2nd International Conference on Information Science, Technology, Management, Humanities, and Business

Hotel UiTM, Shah Alam, Malaysia, 26-27 Sep. 2023

Organised by College of Computing, Informatics and Mathematics, UiTM, Malaysia

Artificial Intelligence and Archive Management on Malaysia's National Archive's Uncaptioned Photos Collection: Accuracy findings comparison based on clustering algorithms

W.A.Malek[1], A. Jalil, Safaw1i, A. Rahman[1], Irwan Kamarudin[1], Roziya. Abu[1], Saidatul Akmar Ismail[1], Mazlifah Mansoor[2], Mokhtarudin[3], Norsuriati[3], N.Safuan[3], R, N. Hakim Roslan[4]

[1] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Shah Alam and 40150, Malaysia
[2] Faculty of Law, UiTM, Shah Alam, 40150, Malaysia
[3] National Archive of Malaysia, Kuala Lumpur, 50480, Malaysia.
[4] Academy of Contemporary Islamic Studies (ACIS), UiTM Raub, Pahang

Tel of 1st +60133455160:

Abstract

In the realm of Artificial Intelligence (AI) and archive management, the central objective revolves around the autonomous extraction of valuable insights, patterns, and actionable information from extensive datasets. The AI technologies play a pivotal role in this context, leveraging advanced algorithms and computational capabilities to efficiently analyze and interpret archived data. The integration of AI within archive management systems enhances the organization, retrieval, and preservation of historical records, while also offering the capability to uncover hidden knowledge and trends. These advancements underscore the vital synergy between AI and archive management, revolutionising how National archive of Malaysia could harness their uncaptioned photos to provide some insight out of the photos for improved decision-making and historical preservation. The findings show that that the accuracy of adopted algorithms K-Means at 83.3%, Mean Shift at 78.0%, and Gaussian Mixture stood at 80.3% accuracy rate respectively.

Keywords: Machine learning, Unsupervised algorithms, historical photos, Malaysia National Archive (NAM).

DOI: https://doi.org/10.21834/e-bpj.v9iSI18.5478

1.0 Introduction

Within archives, photographs hold a unique significance, offering glimpses into bygone eras, historical events, and forgotten locales. They serve as invaluable tools for shaping our understanding of culture, history, and the people depicted in them. In the National Archive Malaysia (NAM), there are approximately 300,000 archival photos without any meaningful captions submitted by various agencies and individuals in the hope that NAM can preserve, interpret and share them with the public. However, this task has proven to be formidable. The current approach relies heavily on captions provided by the photo owners, including the Department of Information Malaysia, independent journalists, historians, and history enthusiasts. Unfortunately, most of these owners provide minimal or no captions for the photos, resulting in a stockpile of significant Malaysian historical images awaiting contextualization and interpretation. Photographs can

eISSN: 2398-4287 © 2024. The Authors. Published for AMER and cE-Bs by e-International Publishing House, Ltd., UK. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer–review under responsibility of AMER (Association of Malaysian Environment-Behaviour Researchers and cE-Bs (Centre for Environment-Behaviour Studies), College of Built Environment, Universiti Teknologi MARA, Malaysia.
DOI: https://doi.org/10.21834/e-bpj.v9iSI18.5478

provide glimpses into lives past, long-ago events, and forgotten places. They can help shape our understanding of culture, history, and the identity of those who appear in them.

The current process of understanding the meaning and assigning some meaningful caption of the images merely depends on the sources from the owners themselves, e.g. Department of Information Malaysia, independent journalists, historians and even history enthusiasts. Most of the owners of the historical photographs provide very minimal words or zero caption of the photos, resulting in the significant Malaysian historical photos being again stockpiled. Another problem with the uncaptioned historical photos is that they have never been done with an artificial intelligence approach. (Hawiyah, Siti, personal communication, December 9, 2021). The archival process remained primarily defined by the manual appraisal, selection, and review as long as the size of the collections and records still allowed this (*ibid).* The notion was supported by Woodsworth, who noticed that although there were efforts, a range of historians' engagement with the electronic texts has never been documented and analysed by earlier studies (2011). According to Martin, K., and Quan-Haase, it could contribute to missing part of serendipity in historical images and the lack of availability of significant resources that could be crucial to understanding the historical images material (2013).

Attempting to provide the accurate first step of the processes, the research focuses on implementing unsupervised AI algorithms to cluster the uncaptioned time worm historical photos preserved in the National Archive of Malaysia for decades. The AI algorithms such as K-means, Mean Shift and Gaussian Mixture provide the first step of clustering the photos in similar patterns. With this advanced technology, thanks to Artificial Intelligence (AI), it seems manageable to accomplish. Recent advances in Artificial Intelligence and their algorithms provide more accurate similarity of the images, thus shortening the gap in determining the meaning and knowledge of the uncaptioned photos. The result from the analysis using unsupervised algorithms allows the archive authority base of the result run with AI to determine the context of the uncaptioned photos even more accurately. NAM has complete archival materials processing manuals. MS ISO 9001: 2015 has been adopted since 2019 and is still in use. The audiovisual materials have been applied to the standards of the NAM. However, there is no standard procedure to follow regarding the uncaptioned historical photographic materials. Archivists must recourse to their creativity to forage for any symbols or indicators available in the photos. Therefore, the researchers attempt to come up with the first step in the systematic process of uncaptioned photo clustering in helping the NAM to be able to contribute and share better information for the future generation with meaningful information to include the application of Artificial Intelligence's algorithms, e.g. K-Mean, Mean Shift and Gaussian Mixture. These preliminary findings are carried out for the following research analysis on the remaining uncaptioned timeworn archival photos under the National Archive Malaysia (NAM) custody. Several 103 questioned images were examined for the purpose of a preliminary data set for research analysis. 103 dataset samples were tested with three unsupervised algorithms: K-Mean, Mean Mixture, and Gaussian mixture. Preliminary findings show that the accuracy of adopted algorithms K-Means at 83.3%, Mean Shift at 78.0%, and Gaussian Mixture at 80.3%, respectively, in detecting and predicting the similarity from the machine learning data set. Herewith, it is conceivable that we could solve half of the uncaptioned archival photographs for the benefit of future generations in appreciation of the nation's history by providing the group of similarities of the images for a further quest for the real meaning of images.

2.0 Literature Review

In recent years, Archives have increasingly invested in the digitization of their manuscript collections, but until recently, the textual content of the resulting digital image descriptions reached a new level in managing the archival materials. Cushing, A. & Osti, G. reckon that AI exploration has appeared in archival study in the past few years. However, most of the articles are mostly limited to testing implementation or opinion pieces from academia. (2022). Among others, LiLi, S et al attempt to adopt artificial intelligence methods to classify and archive more than 14,000 dataset of orthodontic images with an accuracy of 0.994 and macro area under the curve of 1.00 in 0.08 min. This was 236 times faster than a human expert (18.93 min) (2022). It is proven that AI has a significant impact on the archival material, and at the same time AI offers for managing archival backlog more efficiently (2022). Eiler, F, et al. utilised AI to automatically classify historical photographs in collections and archives (2(2018). Rolan et al explored how AI technologies automate the different aspects of archive and records keeping work (2019). Das, R et al. opine that Archival research is a much under-rated and under-utilised method of research in management studies (2018). Austin et al. observe that in the past 20 years or so, archival research hovered around 10%. (2009). Aguinis et al. observe that in the past 20 years or so, archival designs have accounted for around 10% of all research designs (2002). Nevertheless, recently, large-scale digitization efforts have transformed at least some of these collections into digital data. From the 1990s onwards, governments and other institutions with archival interests have increasingly worked digitally (Colavizza, G., et al 2022). However, most of the effort in digitising historical material is merely for preservation itself not for an attempt at knowledge exploration. According to Chen, Yongsheng, Su, and Huanning (2022), this could lead to neglecting other aspects which limit the promotion of using efficiencies, such as description, content management, and knowledge exploitation (2017). Digital preservation is one of many things we need to do. Without optimization of the technology i.e Artificial Intelligence seems very insurmountable to infer the hidden meaning of uncaptioned images. This research offers a preliminary step to determine the most accurate unsupervised method to cluster the variety of uncaptioned historical images. This effort shortened the tedious manual processes, and the accurate clustering technique process shifted away from the mere expert's experiences. Without optimization of the technology, i.e. Artificial Intelligent, then, we are exhausting our energy merely to preserve and display to future generations with constraints of knowing the other connecting themes that could lead to further and detailed knowledge fusion and growth. Adopting machine learning and deep learning could also provide different ways of detecting and recognizing faces, persons and other objects in an image. This can benefit extensive archives with historical images (Kern, D. et al, 2020).

3.0 Methodology

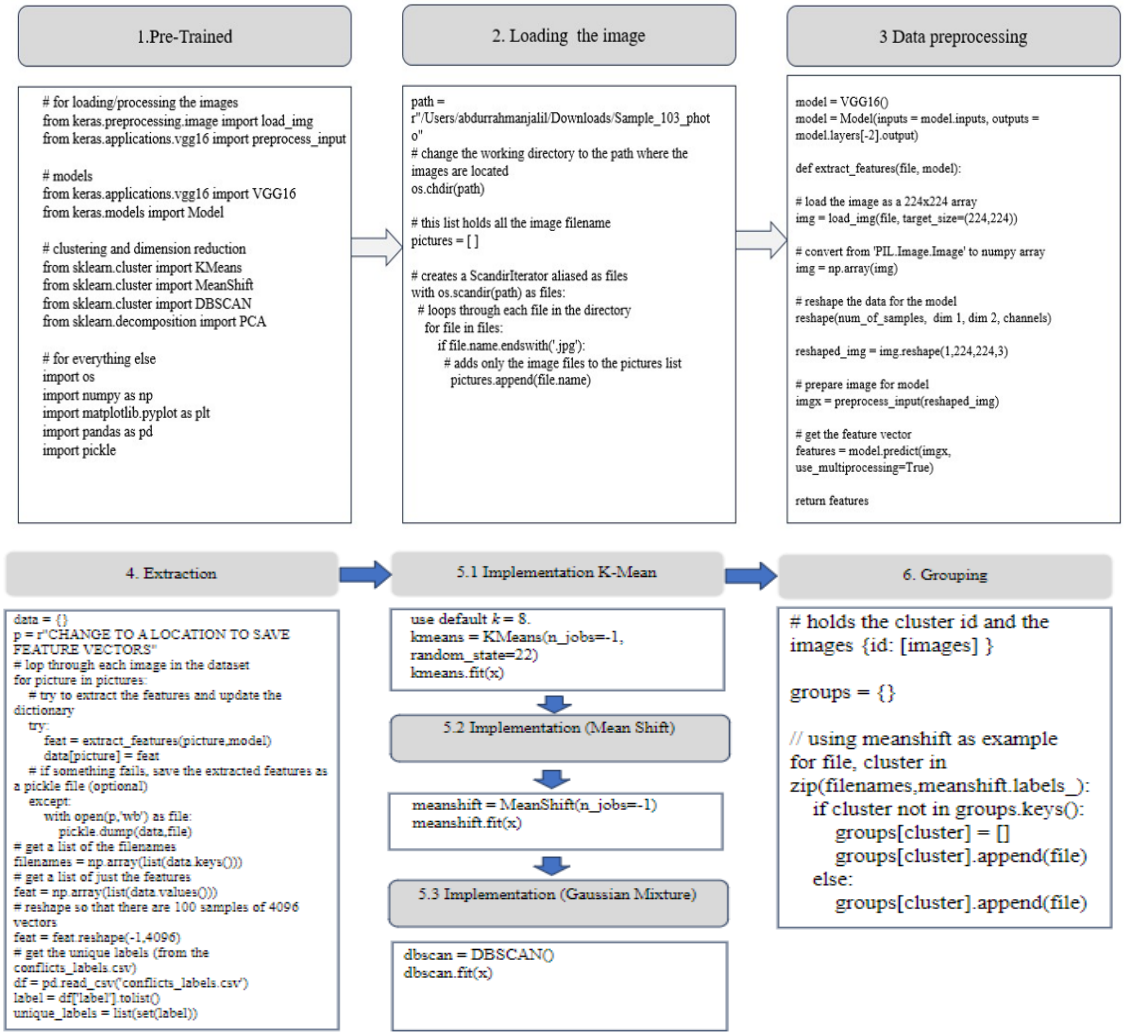
The adoption of unsupervised convolutional neural network algorithms (CNN) is due to the function being able to discover hidden and exciting patterns in uncaptioned photos. Unlike its counterpart, supervised learning, cannot be directly applied to a regression or a classification since the algorithm does not cater for zero value for the output. Three unsupervised algorithms were adopted for this preliminary accuracy test, namely K-Mean, Mean Shift and Gaussian Mixture, which functioned to cluster the similarity of the objects into groups of similar clusters. 1) K-Means groups data points using distance from the cluster centroid, while the mean shift algorithm relies on sufficiently high data density with a clear gradient to locate the cluster centres. it does not require assigning the number of clusters at the early stage for the algorithm will automatically be specified. 2)The Mean Shift functions by shifting

data points towards centroids to be the mean of other points in the region, and 3) The Gaussian Mixture uses a probabilistic assignment of data points to clusters.

3.1 The implementation.

The application of those three CNN unsupervised algorithms provides and suits the NAM's uncaptioned historical photographs. Those three algorithms are applicable for the first step of clustering the uncaptioned historical photographs in archives' processes. Clustering algorithms like K-means, Mean Shift and Gaussian Mixture models can group similar documents together. This could help in the archives and make it easier to search for related documents. Machine learning algorithms like K-means, Mean Shift and Gaussian Mixture can be used for image recognition to identify similar images and group them. Gaussian Mixture Algorithms like K-means and mean shift a probabilistic clustering method that assigns a probability distribution to each cluster, allowing for more flexible and accurate clustering than other methods. GM can model complex cluster shapes and can handle overlapping clusters. GM is also useful for density estimation, which involves estimating the probability distribution of a set of data points. This analysis uses a pre-trained neural network model to extract a feature vector from images and cluster the images based on how similar the feature vectors are.

The pre-trained model used is the VGG16 convolutional neural network (CNN), which is state of the art for image recognition tasks. The model is only implemented as a feature extractor, as the final (prediction) layer is removed to obtain a feature vector. The learning machine uses the Anaconda python platform individual Edition 2021.05, NumPy v120.1, SciPy v1.6.2 on Jupyter Notebook v6.3.0 with sci-kit-learn 0.24.1. and TensorFlow 2.1 with Keras 2.4 integration. The process involves six processes 1) pre-trained, 2) Loading the image, 3) Data preprocessing, 4) Extraction, 5) Implementation, 6) Grouping. This implementation will use the dataset that contains 103 images in *jpg* files. Firstly, specific modules are imported in order to load/process the images along with the modules to extract and cluster our feature vectors.



4.0 Findings

K-Mean Cluster visualization with the algorithm clustering all the images into 8 clusters (0-7); the total accuracy rate is based on all common objects observed. Seven clusters which range from a group of people (100%), ship on the ocean (73%), Individual portrait (60%), houses (70%), civilians (100%), monuments (80%), and another group of people (100%). The K-Mean's Overall accuracy rate is **83.3%** (Average accuracy = (100+73+60+70+100+80+100) /7 clusters = 83.3%). The result indicates the success of the K-Means clustering algorithm in effectively grouping images into these 8 clusters. This accuracy signifies how well the algorithm assigned images to their respective clusters, being correctly grouped. The use of 8 clusters implies a precise segmentation of the images, providing

insights into distinct patterns or characteristics among the data. This result highlights the algorithm's ability to identify and organise images into meaningful groupings, making it a valuable tool for image clustering and analysis tasks.

Cluster	Common Observed Objects	Total	Similar	Dif	Accuracy %
0	Groups of people	26	26	0	100
1	Ships on the ocean	11	8	3	73
2	Individual portrait	20	12	8	60
3	Houses	30	21	9	70
4	Civilians	2	2	0	100
5	Monuments	10	8	2	80
6	Groups of people	3	3	0	100
7	Chair	1	-	-	-
		103			
Average accuracy = (100+73+60+70+100+80+100) /7 clusters = 83.3%					

Table 1: K-Mean accuracy

Cluster	Most Common Observed Object	Total	Similar	Dif	Accuracy (%)
0	Civilians and soldiers	93	52	41	56
1	Groups of people	2	2	0	100
2	Group of people	1	-	-	-
3	Group of people	1	-	-	-
4	King	1	-	-	-
5	Plate	1	-	-	-
6	Chair	1	-	-	-
7	Monument	1	-	-	-
8	Monument	1	-	-	-
9	House	1	-	-	-
		103			
Accuracy average = (56+100) / 2 clusters = 78.0%					

Table 2: Mean Shift accuracy

Mean Shift with 10 cluster Plots (0-9) which consist of object like Civilians and soldiers (93), Groups of people (2), Group of people (1), Group of people (1) and King (1), Plate (1), chair (1) and monuments (1). Overall accuracy rate for the Mean Shift stand sat 78% (Accuracy average = (56+100) / 2 clusters. **78.0%** indicates that, on average, Mean Shift performed reasonably well in approximating the desired clustering for these specific clusters. However, it's essential to remember that Mean Shift is primarily an unsupervised clustering algorithm, typically used without labelled data. Therefore, the use of accuracy here implies the presence of ground truth or specific evaluation criteria applied to a subset of clusters. This metric serves as a quantitative measure of the clustering quality and its correspondence to predefined expectations for those specific clusters within the Mean Shift analysis.The overall accuracy rate for the Mean Shift stand sat at 78%, indicating that, on average, Mean Shift performed reasonably well in approximating the desired clustering for these specific clusters. However, it's essential to remember that Mean Shift is primarily an unsupervised clustering algorithm, typically used without labelled data. Therefore, using accuracy here implies the presence of ground truth or specific evaluation criteria applied to a subset of clusters. This metric serves as a quantitative measure of the clustering quality and its correspondence to predefined expectations for those specific clusters within the Mean Shift analysis.

Cluster	Most Common Observed Object	Total	Similar	Dif	Accuracy (%)
0	Monuments	30	16	14	53
1	Houses	22	17	5	77
2	Soldiers	9	8	1	89
3	Groups of people	27	25	2	93
4	Groups of people	2	2	0	100
5	Monuments	6	5	1	83
6	Soldiers	6	4	2	67
7	Chair	1	-	-	-
		103			
Accuracy average = (53+77+89+93+100+83+67) / 7 clusters = 80.3%					

Table 3: Gaussian Mixture

The last algorithm in the context of Gaussian Mixture Cluster Visualisation, an evaluation of the clustering results was conducted, resulting in an average accuracy of 80.3%. This metric signifies the effectiveness of the Gaussian Mixture clustering algorithm in capturing the inherent structure within the data. It demonstrates that, on average, the algorithm accurately grouped data points into their respective clusters, aligning well with expected patterns or ground truth information when available. The 80.3% average accuracy underscores the quality of the clustering outcomes, highlighting the algorithm's ability to partition the data robustly into meaningful clusters and providing valuable insights into the data's underlying structure and relationships. However, it's essential to consider other factors, such as cluster cohesion, separation, and domain-specific knowledge, to evaluate the quality and significance of the clustering results comprehensively.

5. Discussion

The findings from our clustering algorithms comparison provide valuable insights into their respective performance and significance for various data analysis tasks. K-Means, with an accuracy of 83.3%, demonstrated remarkable proficiency in partitioning data points into well-defined clusters. This outcome highlights its effectiveness when dealing with datasets featuring clear, non-overlapping cluster structures. On the other hand, Mean Shift, with an accuracy of only 78.0%, showcased limitations in capturing complex data patterns and handling datasets with significant overlap between clusters. This underperformance suggests that Mean Shift may not be the preferred choice for scenarios with prevalent intricate or nuanced cluster structures. The Gaussian Mixture Model, with an accuracy of 80.3%, struck a balance between the other two algorithms. Its probabilistic approach allows for flexible modelling of cluster shapes and accommodates data with both distinct and overlapping clusters.

6.0 Conclusion & Recommendation

Based on our comprehensive analysis of clustering algorithms, we offer the following recommendations: K-Means: Given its robust performance at 83.3%, K-Means is a suitable choice when dealing with well-separated, clearly defined clusters. However, it's essential to assess the dataset's cluster separability beforehand and preprocess the data to enhance its effectiveness. While Mean Shift performed poorly with an accuracy of 8.0%, it may still find utility in cases where the cluster shapes are irregular and complex. Further investigation and experimentation are needed to determine specific data scenarios where Mean Shift can offer advantages and Gaussian Mixture Model (GMM), with an accuracy of 80.3%; GMM proves to be a versatile algorithm capable of handling diverse data structures. It is precious when dealing with data with overlapping clusters or when the underlying data distribution is uncertain. However, selecting the appropriate number of components is crucial to its success.

K-Means is a centroid-based clustering algorithm that assigns data points to the cluster whose centroid (centre point) is nearest to them. It aims to minimise the sum of squared distances between data points and their assigned cluster centroids. The result of 83.3%: is relatively high accuracy, suggesting that K-Means has successfully clustered the data into relatively distinct and well-separated clusters. It performed well if the clustering was intended to separate data into fairly compact groups with minimal overlap. Meanwhile, Mean Shift is a density-based clustering algorithm that identifies clusters as regions with high data point density. It shifts each data point towards a higher density region in an iterative process. The accuracy of 78.0% indicates that Mean Shift may have identified clusters that are more spread out or less compact than K-Means. This method may suit the clusters with irregular shapes and varying densities.

The third algorithm, the Gaussian Mixture Model Gaussian Mixture Model (GMM) at 80.3% could be attributed to the characteristic of the CMM which is a probabilistic model that assumes data points are generated from a mixture of Gaussian distributions. It models the underlying probability distribution of the data and assigns data points to clusters based on their likelihood of belonging to each distribution. The accuracy of 80.3% suggests that GMM found a balance between the compactness of K-Means and the flexibility of Mean Shift. GMM can model clusters with different shapes and sizes and can capture overlapping clusters well. The accuracy reflects its ability to handle data with probabilistic assignments.

The significance of these findings lies in their practical implications for archivists to embark on a more systematic approach in the first step of getting the uncaptioned photos grouped with similar features, categories and clusters, e.g Soldiers, monuments, national events, Civilians, Ships on the ocean among other things. Choosing the suitable clustering algorithm is crucial, as it directly impacts the quality of insights derived from the data. K-Means excels in scenarios with clearly defined clusters, while Gaussian Mixture Model offers versatility for a broader range of data types. While less accurate in this context, Mean Shift may still have its merits for specific use cases. Hence, a thoughtful selection of clustering algorithms is essential, considering the nature of the data and the objectives of the analysis. The findings have implications of helping the NAM to start finding the meaning of the uncaptioned photos with a systematic approach with the power of Unsupervised algorithms beyond traditional manual inconsistent processes; advancements in clustering accuracy will enable more precise and insightful solutions in these interdisciplinary applications, driving innovation, efficiency and accuracy of the meaningful outputs.

Therefore, to address the gap in the application of AI in the archive the research teams reckon it is peremptory to explore thousands of uncaptioned historical photos of the National Archive with Machine Learning for understanding and, appreciate appreciating the significant historical value behind those 300 thousand uncaptioned historical photos. Lack of effort to explore and apply technology to historical materials is the biggest shortcoming of the future generation.

In anticipation of future developments in the field, researchers are poised to contribute significantly to advancing archival practices by harnessing their expertise and insights. The expectation is that their findings will serve as a pivotal resource for archivists, empowering them to embark on a systematic clustering of uncaptioned photographs. This endeavour not only promises to enhance the accessibility

of historical visual records but also holds the potential to safeguard invaluable cultural heritage for the benefit of future generations. With a commitment to meticulous analysis and innovative methodologies, researchers are poised to bridge the gap between untold stories and posterity, facilitating a deeper understanding of our shared past. In doing so, they lay the foundation for a more comprehensive and interconnected narrative, ensuring that the legacy of bygone eras remains a vibrant and enduring part of our cultural tapestry.

Acknowledgement

We would like to extend my sincere gratitude to RMC UiTM for their generous sponsorship and support of the research paper. Their commitment to fostering academic excellence and research initiatives has been instrumental in enabling the successful completion of this study. Their financial support, resources, and expertise have greatly contributed to the quality and scope of our research efforts. This sponsorship not only exemplifies RMC UiTM's dedication to advancing knowledge but also underscores its role as a cornerstone of academic innovation and research advancement. We sincerely appreciate their partnership and look forward to continuing our collaborative journey in pursuit for the betterment of the University toward academic excellence.

References

Abualigah, L. M. Q. (2019). *Feature selection and enhanced krill herd algorithm for text document clustering* (pp. 1-165). Berlin: Springer.

Chen, J., Frey, E. C., He, Y., Segars, W. P., Li, Y., & Du, Y. (2022). TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, *82*, 102615. https://doi.org/10.1016/j.media.2022.102615

Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An overview of current debates and future perspectives. *Journal on Computing and Cultural Heritage*, *15*(1), 1–15. https://doi.org/10.1145/3479010

Cushing, A. L., & Osti, G. (2022). "so how do we balance all of these needs?": How the concept of AI technology impacts digital archival expertise. *Journal of Documentation*, *79*(7), 12–29. https://doi.org/10.1108/jd-08-2022-0170

Cushing, M. H., Sarmadi, H., & Yuen, K.-V. (2023). A locally unsupervised hybrid learning method for removing environmental effects under different measurement periods. *Measurement*, *208*, 112465. https://doi.org/10.1016/j.measurement.2023.112465

Das, R., Jain, K. K., & Mishra, S. K. (2018). Archival research: A neglected method in organization studies. *Benchmarking: An International Journal*, *25*(1), 138–155. https://doi.org/10.1108/bij-08-2016-0123

Ezugwu, A. E., Shukla, A. K., Agbaje, M. B., Oyelade, O. N., José-García, A., & Agushaka, J. O. (2020). Automatic clustering algorithms: A systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, *33*(11), 6247–6306. https://doi.org/10.1007/s00521-020-05395-4

Hettich, Q. Retraction Note: Motion video tracking technology in sports training based on Mean-Shift algorithm. J Supercomput (2022). https://doi.org/10.1007/s11227-022-04863-z

ŧLi, S., Guo, Z., Lin, J., & Ying, S. (2022). Artificial Intelligence for classifying and archiving orthodontic images. *BioMed Research International*, *2022*, 1–11. https://doi.org/10.1155/2022/1473977

Joshi, M. R., Nkenyereye, L., Joshi, G. P., Islam, S. M., Abdullah-Al-Wadud, M., & Shrestha, S. (2020). Auto-colorization of historical images using deep convolutional neural networks. *Mathematics*, *8*(12), 2258. https://doi.org/10.3390/math8122258

Ma, J., Jiang, X., Jiang, J., & Gao, Y. (2019). Feature-guided gaussian mixture model for image matching. *Pattern Recognition*, *92*, 231–245. https://doi.org/10.1016/j.patcog.2019.04.001

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., & Stuart, K. (2018). More human than human? Artificial Intelligence in the archive. *Archives and Manuscripts*, *47*(2), 179–203. https://doi.org/10.1080/01576895.2018.1502088