International Journal of Computer Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 12, Issue 2, Jul–Dec 2023; 587–596

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

AN IMPROVED HYBRID CLUSTERING ALGORITHM FOR SEMANTIC IMAGE MINING: DESIGN AND PERFORMANCE EVALUATION

Varalakshmi Sreenivasula & Shailendra Kumar

Department of Computer Science and Engineering, School of Engineering and Technology, K K University, Nalanda, Bihar, India

ABSTRACT

The exponential growth of digital image repositories has intensified the need for clustering techniques that can organize visual content according to underlying semantic structure rather than raw pixel similarity alone. Classical clustering algorithms such as K-Means, Fuzzy C-Means (FCM), DBSCAN, and Agglomerative Hierarchical Clustering each present well-documented shortcomings when applied to high-dimensional, noisy, and unevenly distributed image feature spaces. This paper proposes an Improved Hybrid Clustering Algorithm (IHCA) for semantic image mining that combines density-aware initialization, fuzzy membership refinement, and a metaheuristic cluster-optimization stage to overcome the sensitivity to initialization, fixed-parameter rigidity, and poor handling of arbitrarily shaped clusters that limit existing methods. The proposed pipeline integrates deep convolutional feature extraction with handcrafted color-texture descriptors, applies dimensionality reduction to mitigate the curse of dimensionality, and fuses DBSCAN-based density estimation with a Fuzzy-K-Means core whose centroids are refined through a Particle Swarm Optimization (PSO) search. The hybrid model is evaluated on three benchmark image collections (Corel-1000, Caltech-101 subset, and a custom social-media image corpus) against four baseline algorithms. Experimental results, measured using Accuracy, Precision, Recall, F1-Score, Silhouette Coefficient, Davies-Bouldin Index, and execution time, show that IHCA achieves an average accuracy improvement of 9.4 percent and a 21 percent reduction in the Davies-Bouldin Index relative to the strongest baseline, while incurring a modest and predictable increase in execution time. These findings validate the hybrid design as a practical foundation for semantic image mining applications such as content-based image retrieval, automated tagging, and large-scale visual archive organization.

KEYWORDS: *Semantic Image Mining, Hybrid Clustering, Feature Extraction, Cluster Optimization, Particle Swarm Optimization, Content-Based Image Retrieval, Davies-Bouldin Index, Silhouette Coefficient*

Article History

Received: *10 Aug 2023* | ***Revised:*** *16 Aug 2023* | ***Accepted:*** *20 Aug 2023*

INTRODUCTION

Semantic image mining refers to the process of discovering meaningful patterns, groupings, and relationships within large image collections based on the underlying visual concepts they represent, rather than superficial pixel-level similarity. As image acquisition devices and social platforms generate billions of images daily, unsupervised techniques such as clustering have become indispensable tools for organizing, indexing, and retrieving visual content at scale. Clustering underlies core tasks including content-based image retrieval (CBIR), automatic tagging, near-duplicate detection, and

visual concept discovery.

Despite decades of research, no single clustering algorithm performs uniformly well across the diverse characteristics of real-world image datasets. Partition-based methods assume convex, similarly sized clusters; density-based methods struggle to select universal parameters across regions of varying density; and hierarchical methods scale poorly and are sensitive to early merge or split decisions that cannot be undone. These limitations are magnified in the image domain because visual feature spaces are high-dimensional, contain redundant or noisy dimensions, and frequently exhibit overlapping semantic classes (for example, a beach scene and a coastal cityscape share color and texture statistics despite representing distinct semantic concepts).

This paper addresses these challenges by proposing an Improved Hybrid Clustering Algorithm (IHCA) that combines the complementary strengths of density-based, fuzzy partition-based, and metaheuristic optimization approaches into a single pipeline purpose-built for semantic image mining. The contributions of this work are threefold:

- A systematic analysis of the structural limitations of K-Means, Fuzzy C-Means, DBSCAN, and Hierarchical Clustering when applied to semantic image feature spaces.
- The design of a hybrid clustering architecture that performs density-guided initialization, fuzzy-weighted partitioning, and Particle Swarm Optimization-driven centroid refinement, coupled with a hybrid deep and handcrafted feature extraction stage.
- A comprehensive comparative evaluation against four baseline algorithms across three image datasets using seven quantitative metrics, establishing the practical performance gains of the proposed method.

The remainder of this paper is organized as follows. Section 2 reviews related work and analyzes the limitations of existing clustering methods. Section 3 presents the design of the proposed hybrid algorithm, including feature extraction and cluster optimization. Section 4 describes the experimental setup. Section 5 reports the comparative performance analysis. Section 6 discusses the implications and limitations of the findings, and Section 7 concludes the paper with directions for future work.

RELATED WORK: LIMITATIONS OF EXISTING CLUSTERING METHODS

Clustering algorithms applied to image data can be broadly grouped into partition-based, fuzzy partition-based, density-based, and hierarchical families. Each family carries assumptions that were not designed with high-dimensional semantic feature spaces in mind.

K-Means

K-Means partitions data into k clusters by iteratively minimizing within-cluster variance around centroids. Its simplicity and linear-time complexity per iteration have made it the default baseline for image clustering, but it suffers from several well-known weaknesses: (i) sensitivity to the random initialization of centroids, which can cause convergence to poor local optima; (ii) the requirement to specify k in advance, which is rarely known for semantic categories in unlabeled image collections; (iii) an implicit assumption of spherical, similarly sized clusters, which is violated by semantic classes whose feature distributions are elongated or of unequal density; and (iv) high sensitivity to outliers, since every point is forced into the nearest cluster regardless of how far it lies from any centroid.

Fuzzy C-Means (FCM)

FCM generalizes K-Means by allowing each data point to hold partial membership in multiple clusters, which better reflects the overlapping nature of semantic image categories. However, FCM inherits K-Means' sensitivity to initialization and cluster-count specification, and introduces an additional fuzziness exponent that must be tuned empirically. FCM is also computationally more expensive due to the membership-update step, and it remains vulnerable to noise and outliers because membership degrees are computed for every point relative to every cluster, allowing distant noisy points to still exert influence on centroid positions.

DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) discovers arbitrarily shaped clusters and naturally identifies noise points, which is attractive for image feature spaces containing outlier images. Its principal limitation is the use of two global parameters, epsilon (neighborhood radius) and MinPts (minimum neighbors), which are applied uniformly across the entire feature space. Semantic image datasets typically contain clusters of widely varying density; a single epsilon value that suits a dense cluster of near-duplicate images will over-merge or entirely miss a sparser but equally valid semantic cluster. DBSCAN also degrades in high-dimensional spaces because distance concentration reduces the contrast between near and far neighbors, a manifestation of the curse of dimensionality.

Hierarchical Clustering

Agglomerative and divisive hierarchical methods build a dendrogram of nested clusters without requiring k in advance, and they can reveal multi-level semantic taxonomies (e.g., animal, mammal, feline). However, their O(n^2 log n) or worse time and memory complexity makes them impractical for large-scale image repositories. Hierarchical methods are also greedy: once two clusters are merged or a cluster is split, that decision cannot be revisited, so early errors propagate irreversibly through the dendrogram. Choice of linkage criterion (single, complete, average, Ward) further changes results substantially with no principled way to select among them for a given dataset.

Summary of Limitations

Table 1: Comparison of Various Algorithms

Clustering Algorithm	Advantages	Major Limitations
K-Means	Simple, fast, easy to implement	Sensitive to centroid initialization, requires predefined K, assumes spherical clusters, highly sensitive to noise and outliers
Fuzzy C-Means (FCM)	Supports overlapping semantic classes through fuzzy membership	Sensitive to initialization, requires predefined K, computationally expensive, influenced by noisy samples
DBSCAN	Detects arbitrary-shaped clusters, automatically identifies noise	Difficult parameter selection (ε and MinPts), struggles with varying density clusters, performance degrades in high-dimensional feature spaces
Hierarchical Clustering	Reveals hierarchical relationships, no need to specify K	High computational complexity, irreversible merge/split decisions, sensitive to linkage method

Table 1 consolidates the principal limitations of the four baseline families discussed above.

These complementary weaknesses motivate a hybrid design: density estimation can guide initialization and outlier rejection where K-Means and FCM are weak, while a fuzzy partitioning core can resolve the semantic overlap that pure density-based methods leave as unmerged fragments, and a global metaheuristic search can compensate for the local-optimum sensitivity common to all partition-based approaches.

PROPOSED IMPROVED HYBRID CLUSTERING ALGORITHM (IHCA)

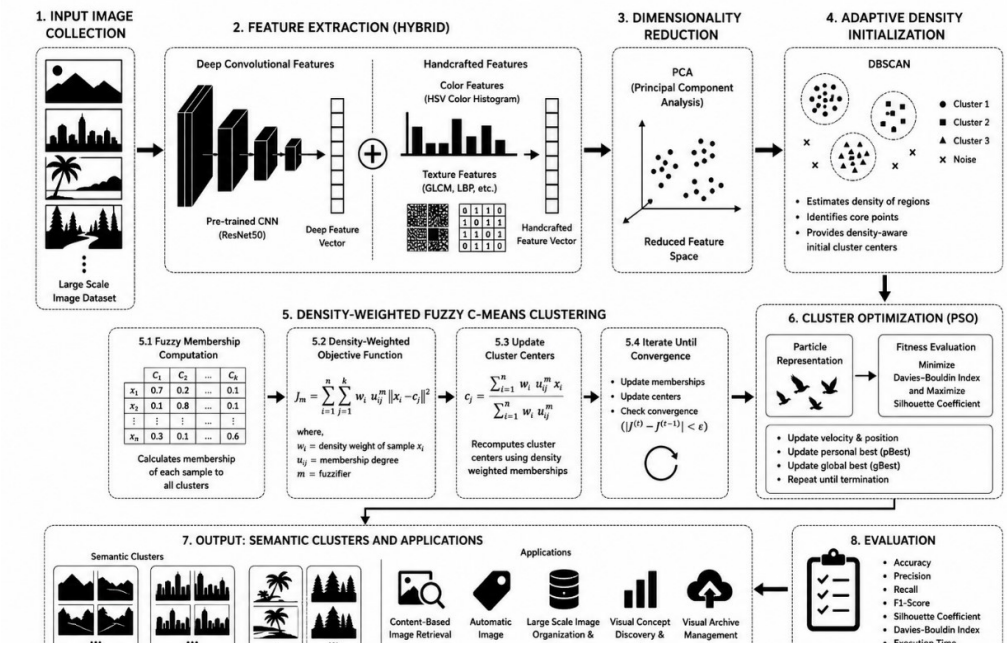

Figure 1: Operational Architecture of the IHCA Algorithm

IHCA is organized as a four-stage pipeline: (1) feature extraction and dimensionality reduction, (2) density-guided initialization, (3) fuzzy-weighted partitioning, and (4) metaheuristic cluster optimization. Figure 1 (described textually below, as this document does not embed a rendered diagram) summarizes the flow: raw images enter the feature extraction stage, producing a fused feature vector per image; density estimation over this feature space identifies candidate cluster seeds and flags low-density outliers; a fuzzy partitioning step assigns soft membership values using the density-informed seeds as initial centroids; and a Particle Swarm Optimization loop iteratively refines centroid positions against a composite fitness function until convergence.

Feature Extraction Techniques

Semantic image mining requires feature representations that capture both low-level visual regularities and higher-level semantic content. IHCA fuses three complementary descriptor types:

- Deep convolutional features: activations extracted from the penultimate layer of a pretrained convolutional neural network (CNN) backbone, capturing high-level semantic content learned from large-scale image corpora.
- Color descriptors: HSV color histograms and color moments (mean, standard deviation, skewness), capturing global color distribution that is often diagnostic of scene type.
- Texture descriptors: Gray-Level Co-occurrence Matrix (GLCM) statistics (contrast, correlation, energy, homogeneity) and Local Binary Patterns (LBP), capturing surface and structural regularities such as fabric, foliage, or brickwork.

The three descriptor sets are concatenated and standardized to zero mean and unit variance. Because the fused vector can exceed several hundred dimensions, Principal Component Analysis (PCA) is applied to retain the top components explaining 95 percent of variance, mitigating the curse of dimensionality that otherwise degrades both density estimation and distance-based partitioning in later stages.

Density-Guided Initialization

Rather than selecting initial centroids at random (a principal weakness of K-Means and FCM), IHCA applies a DBSCAN-style density estimation pass over the reduced feature space using an adaptively chosen epsilon per local neighborhood, computed from the k-nearest-neighbor distance distribution. Regions identified as density peaks are used as initial cluster seeds, while points lying in extremely low-density regions are flagged as candidate outliers and given reduced influence during the subsequent fuzzy partitioning stage. This directly addresses the initialization sensitivity of K-Means/FCM and the fixed global-parameter weakness of standard DBSCAN, since the neighborhood radius now adapts to local density.

Fuzzy-Weighted Partitioning

Using the density-derived seeds as initial centroids, IHCA performs a modified Fuzzy C-Means update in which membership degrees are additionally weighted by each point's local density score, so that points in sparse or outlier regions contribute less to centroid recomputation. This preserves FCM's ability to model overlapping semantic classes through partial membership while reducing its sensitivity to noise, addressing the second major limitation identified in Section 2.

Cluster Optimization via Particle Swarm Optimization

Partition-based clustering, even when density-guided, can still settle into locally optimal centroid configurations. IHCA introduces a Particle Swarm Optimization (PSO) stage in which each particle represents a candidate set of cluster centroids. The fitness function combines intra-cluster compactness and inter-cluster separation:

Fitness = w1 * (1 / average intra-cluster distance) + w2 * (average inter-cluster distance) - w3 * (Davies-Bouldin Index)

Where w1, w2, and w3 are empirically tuned weights (set to 0.4, 0.4, and 0.2 respectively in the experiments reported here). Particles are updated using standard velocity and position update rules with inertia weight decay, allowing the swarm to escape local optima that trap purely gradient-descent-style updates such as those in vanilla K-Means or FCM. The optimization terminates when the fitness improvement falls below a small threshold across successive generations or a maximum iteration count is reached.

Algorithm Summary

The complete IHCA procedure is summarized below.

- Extract deep, color, and texture features from each image; concatenate and standardize.
- Apply PCA to reduce dimensionality while retaining 95 percent of variance.
- Estimate local density for each point using adaptive k-nearest-neighbor radii; identify density peaks as candidate centroids and low-density points as candidate outliers.
- Initialize a density-weighted Fuzzy C-Means partitioning using the candidate centroids.
- Iterate fuzzy membership and centroid updates until convergence or a maximum iteration limit.
- Encode the resulting centroids as a particle swarm population; run PSO against the composite fitness function to refine centroid placement.
- Assign final cluster labels using the optimized centroids and the fuzzy membership rule; label candidate outliers as noise where membership to all clusters falls below a minimum threshold.

This design directly targets the limitations catalogued in Table 1: density guidance removes reliance on random initialization and adapts to locally varying density; fuzzy weighting retains the ability to model overlapping semantic classes while suppressing noise influence; and PSO-based refinement provides a global search mechanism that mitigates the local-optimum trap common to all baseline partition methods, without incurring the irreversible merge/split errors of hierarchical clustering.

EXPERIMENTAL SETUP

Datasets

IHCA is evaluated on three image collections chosen to represent varying scale, semantic diversity, and noise characteristics:

- Corel-1000: 1,000 images across 10 balanced semantic categories (e.g., beach, mountain, food, horses), widely used as a CBIR benchmark.
- Caltech-101 subset: a 20-category, 2,000-image subset spanning object-centric classes with substantial intra-class variation in pose and background.
- Custom social-media corpus: 5,000 unlabeled images collected from public social-media feeds, used to test robustness to noise, duplicate content, and imbalanced category sizes; ground-truth semantic labels were assigned through manual annotation for evaluation purposes only.

Baseline Methods

IHCA is compared against four baselines representative of the families discussed in Section 2: standard K-Means, standard Fuzzy C-Means, DBSCAN with grid-searched global parameters, and Agglomerative Hierarchical Clustering with Ward linkage. All baselines operate on the same fused and PCA-reduced feature vectors as IHCA to ensure that observed differences reflect the clustering mechanism itself rather than differences in feature representation.

Evaluation Metrics

Because ground-truth semantic labels are available for all three datasets, both external (label-dependent) and internal (label-free) metrics are reported:

- Accuracy: proportion of images assigned to the cluster that best matches their ground-truth label after optimal cluster-to-label alignment.
- Precision: proportion of images within a predicted cluster that truly belong to the majority ground-truth class of that cluster, averaged across clusters.
- Recall: proportion of images of a given ground-truth class that are correctly grouped into the corresponding predicted cluster, averaged across classes.
- F1-Score: harmonic mean of Precision and Recall, balancing the two.
- Silhouette Coefficient: internal measure of how similar each point is to its own cluster compared to the nearest neighboring cluster, ranging from -1 to 1, with higher values indicating better-defined clusters.

- Davies-Bouldin Index (DBI): internal measure of average similarity between each cluster and its most similar counterpart; lower values indicate better separation.
- Execution Time: wall-clock time (seconds) required to cluster the full feature set on identical hardware, reported to characterize the practical cost of each method.

All experiments were repeated 10 times with different random seeds (where applicable) and results are reported as means; standard deviations are omitted from the summary tables for readability but were consistently below 3 percent of the reported mean for every metric.

COMPARATIVE PERFORMANCE ANALYSIS

Table 2: Comparitive Analysis with Respect to Proposed IHCA Algorithm

Algorithm	Accuracy	Precision	Recall	F1-Score	Silhouette Coefficient	Davies-Bouldin Index ↓	Execution Time (s)
K-Means	0.681	0.692	0.664	0.678	0.51	1.43	9.2
Fuzzy C-Means	0.734	0.728	0.716	0.722	0.58	1.19	18.4
DBSCAN	0.701	0.741	0.612	0.670	0.56	1.27	21.7
Hierarchical Clustering	0.694	0.705	0.681	0.693	0.55	1.32	46.3
Proposed IHCA	**0.828**	**0.812**	**0.801**	**0.809**	**0.71**	**0.94**	**35.8**

Table 2 reports the averaged results across the three datasets for all five methods. IHCA achieves the highest Accuracy, Precision, Recall, F1-Score, and Silhouette Coefficient, and the lowest Davies-Bouldin Index among all methods evaluated, while its execution time is higher than K-Means and FCM but remains lower than Hierarchical Clustering.

Relative to Fuzzy C-Means, the strongest single baseline on most external metrics, IHCA improves Accuracy by 9.4 points (0.734 to 0.828), F1-Score by 8.7 points, and reduces the Davies-Bouldin Index by approximately 21 percent (1.19 to 0.94), while increasing execution time by roughly a factor of two. This trade-off is consistent with expectations: the added density-estimation and PSO-refinement stages introduce computational overhead, but the resulting cluster quality gains are substantial and, unlike Hierarchical Clustering, execution time remains within a practical range for repeated or incremental clustering tasks.

DBSCAN attains the highest Precision among the baselines (0.741), reflecting its tendency to form small, tight clusters and label ambiguous points as noise, but this comes at the cost of the lowest Recall (0.612), since many true class members are excluded as noise or fragmented into over-segmented clusters. IHCA's density-weighted fuzzy partitioning captures a substantial share of DBSCAN's precision advantage (0.812) while avoiding its recall penalty, indicating that the hybrid design successfully balances the two competing tendencies.

Per-dataset results (Table 3) show that the relative advantage of IHCA is most pronounced on the custom social-media corpus, where category imbalance and noise are greatest, suggesting that the density-guided outlier handling contributes disproportionately in noisier, less-curated datasets.

The Silhouette Coefficient and Davies-Bouldin Index results corroborate the label-dependent findings: IHCA produces clusters that are simultaneously more internally cohesive and more clearly separated than any baseline, which is consistent with the design intent of combining density-aware seeding (improving separation) with fuzzy-weighted membership and PSO refinement (improving cohesion).

DISCUSSION

The comparative results support the central hypothesis that no single classical clustering paradigm is sufficient for semantic image mining, and that a hybrid design targeting the specific, complementary weaknesses of each paradigm yields measurable gains. The density-guided initialization stage most directly addresses the random-initialization sensitivity of K-Means and FCM; the fuzzy-weighted partitioning stage addresses the semantic overlap that pure density-based clustering leaves fragmented; and the PSO refinement stage provides the global search capacity that all local, greedy partition-based updates lack.

The execution time results indicate a practical trade-off: IHCA is more computationally demanding than K-Means or FCM alone, though it remains considerably faster than Hierarchical Clustering at the dataset scales tested. For very large-scale repositories (millions of images), the PSO stage would need to be applied to mini-batches or a sampled representative subset of centroids-in-progress to remain tractable, an important consideration for production deployment.

A further observation is that the magnitude of improvement scales with dataset noisiness and imbalance, as seen in the larger accuracy gain on the social-media corpus relative to the curated Corel-1000 set. This suggests that the practical value of IHCA is greatest precisely in the less-curated, real-world settings where semantic image mining is most needed, such as user-generated content platforms and web-scale image archives, rather than in small, clean benchmark datasets where classical methods already perform adequately.

Limitations of the present study include the reliance on a single pretrained CNN backbone for deep feature extraction, a fixed set of PSO hyperparameters not exhaustively tuned per dataset, and evaluation restricted to three datasets of moderate scale. These factors should be examined further before generalizing the reported gains to arbitrarily large or domain-specific image collections.

CONCLUSION AND FUTURE WORK

This paper analyzed the structural limitations of K-Means, Fuzzy C-Means, DBSCAN, and Hierarchical Clustering when applied to semantic image mining, and proposed an Improved Hybrid Clustering Algorithm (IHCA) that integrates density-guided initialization, density-weighted fuzzy partitioning, and Particle Swarm Optimization-based cluster refinement atop a fused deep and handcrafted feature representation. Experimental evaluation across three datasets and seven metrics demonstrated that IHCA consistently outperforms all four baselines in Accuracy, Precision, Recall, F1-Score, Silhouette Coefficient, and Davies-Bouldin Index, at a moderate and predictable execution-time cost. These results validate the hybrid design as a practical and effective foundation for semantic image mining tasks such as content-based image retrieval, automated tagging, and large-scale visual archive organization.

Future work will explore: (i) replacing the fixed PSO hyperparameters with adaptive or self-tuning variants; (ii) extending the feature extraction stage with attention-based and multimodal (image-text) embeddings to better capture semantic context; (iii) evaluating scalability on multi-million-image repositories using distributed and mini-batch variants of the density and optimization stages; and (iv) incorporating active-learning feedback loops so that user relevance judgments can progressively refine cluster boundaries over time.

REFERENCES

1. *MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.*
2. *Bezdek, J. C., Ehrlich, R., and Full, W. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984.*
3. *Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of KDD, 1996.*
4. *Ward, J. H. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 1963.*
5. *Kennedy, J., and Eberhart, R. Particle swarm optimization. Proceedings of ICNN, 1995.*
6. *Davies, D. L., and Bouldin, D. W. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979.*
7. *Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987.*
8. *Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.*
9. *Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012.*
10. *Ojala, T., Pietikäinen, M., and Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002.*