International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 8029-8041

http://www.ripublication.com

# Feature Extraction and Feature Selection For Content – Based Image Classification Using Proclass Algorithm

**B. Muthukumar[a*], P.Sivakumar[b], P.Banumathi[c],**
*[a]Department of IT, Syed Ammal Engineering College, Ramanathapuram, India*
*[b]Department of ECE, GRT Institute of Engineering and Technology, Thiruttani, India*
*[c] Department of CSE, Kathir College of Engineering, Coimbatore, Tamilnadu, India*
*[*]Corresponding AuthorE-Mail id: muthu122@gmail.com*

## Abstract

In image processing, feature extraction is a special form of dimensionality reduction where the feature subset selection is normally used to remove irrelevant and redundant data dimensions. A set of image data features are extracted to found the subset of features which are most relevant to data mining task. A Feature Selection Algorithm (FSA) is proposed to evaluate both the efficiency and effectiveness points of view. In efficiency concern the time required to find a subset of features and effectiveness is related to the quality of the subset of features. The Spatial Gray Level Difference Method (SGLDM) feature extraction algorithm and Correlation based Feature Selection (CFS), Projected Classification algorithms (PROCLASS) are identified to propose the brain image data's as input and these experimental results are going to do compare these plug-in algorithms with conventional based feature selection algorithms.

**Keywords:** Feature Extraction, Feature Selection, CFS and PROCLASS

## Introduction

In recent, the computerized pictures are created universally and the numerous pictures on the internet may cause more problems for the clients.Images are the essential piece of input in our day to day life and the correspondent colossal measure of pictures digitally accessible is not reasonable by people any more. An individual viewin a database of 100 pictures will most likely to find that what he looks for quick by simply seeing the pictures or little forms of the pictures.

Nowadays, the computers have the ability to recovery the contents in the report by Internet Search motor called Google. Google offers a chance to scan pictures; however the method for inquiry is performed and it does not dependably prompt agreeable outcomes. One approach is to inquiry in pictures databases is to make a

literary portrayal of every last one of pictures in the database and utilize the systems from content based data recovery to pursuit dependent upon the printed depictions.

In this research work, the pre-processing stage is undertaking a human body image as picture recovery which is situated to extract the cerebrum in the brain to emphasize the features extraction and calculate the image features to determinate the concentrated features. Then the features are utilized as bunching pictures into comparable aggregations. As for the features extraction calculation, features extraction is an extraordinary manifestation of dimensionality lessening to calculate and the suspected data information will be changed into lessened representation set of features called features extraction.

Numerous features subset determination routines have been proposed and mulled over for machine taking in provisions and it might be partitioned into four general classes: the Embedded, Wrapper, Filter and Hybrid methodologies. These channel techniques are guaranteedto be specific FAST [12] and CFS [4] features subset selection calculations were proposed. FAST calculation on the feature extraction of picture information is to be neglected and select the features when contrast and content data is available. Also, the Correlation based feature selection will not able to focus the grouping picture in high-dimensional data information. For group investigation, Projected Classification (PROCLASS) calculations have been generally considered and it is utilized as a part of numerous provisions particularly at high dimensional information. In our work, an anticipated grouping module with correlation based features determination calculation is applied for mind pictures to emphasizethe picture investigation.

## Related Work

The different features extraction algorithms were proposed tocalculate the feature extraction and it can be adequately remove the features of pictures from the picture data. The Digital Terminal Model (DTM) features extraction will separate the features of pictures and it is able to achieve 95% of feature identification without human interfering [2] also and it is relevant for Global Positing System (GPS) picture. However, it is neglected to concentrate the features of medicinal pictures, for example, human mind, liver pictures. For the therapeutic picture examination, the SGLDM (Spatial Gray Level Difference Method) [13] for features extractions is identified and remove the picture features from the mind picture immediately.Features subset choice is the methodology of recognizing both irrelevant and repetitive features in two steps: (i) First, irrelevant featureswill don't facilitate the accuracy and (ii) Second, the redundant features will don't show the improvement indicator in the image features.

The numerous characteristic can adequately dispose an insignificant featureand however it will neglect to handle repetitive features. Yet some of others can kill the unessential while dealing with the excess features [15], [4], [1], [9]. FAST [12] calculation falls into the second aggregation. Generally, offer subset choice examination has kept tabs on scanning for applicable features. A well-known case is Relief [7], which weighs each one characteristic consistent with its capacity to

segregate examples under distinctive targets dependent upon separation based criteria capacity. On the other hand, Relief is inadequate at uprooting repetitive features as two prescient however profoundly related features are likely both to be remarkably weighted [7]. Easing augments Relief-F [8], empowering this system to work with boisterous and inadequate information sets and to manage multiclass issues, yet can't recognize excess features. Nonetheless, on top of unimportant features, excess features additionally influence the velocity and correctness of taking in calculations, and accordingly ought to be dispensed with too.CFS [4], FCBF [15], and CMIM [11] are illustrations that think seriously about the repetitive features. CFS, CMIM [6] iteratively picks features which amplify their shared data with the class to foresee, restrictively to the reaction of any characteristic recently picked. Unique in relation to these calculations, the FAST calculation utilizes the grouping based technique to pick features.

FAST [12], calculation utilizes least spreading over tree-based strategy to group features. Meanwhile, it doesn't expect that information focuses are bunched around focuses or differentiated by a standard geometric bend. Quick is breaking point to picture information when features determination caused, when contrast and the content information.CFS [10] is attained by the theory that a great features subset is one that holds emphasizes profoundly associated with target, yet uncorrelated with one another. FCBF [14], [15] is a quick channel strategy which can distinguish significant features and repetition around important features without pair savvy association investigation.

By the investigation of FAST [12] for picture information, CFS acquires the rank 1 and FAST ranks 3; so, we connected the CFS features determination calculation for medicinal picture dissection. With the end goal of investigating the relationship between the features determination calculation and closeness matches of pictures we proposed Projected Classification (PROCLASS) for the medicinal picture information.

## Proposed Methodology

The main focus of this proposed work is to classify the brain data set images into similar groups based on the given input image.Figure 1 show that the proposed framework architecture will use the Spatial Grey Level Difference Method feature extraction, Correlation based Feature selection and proposed Projected Classification (PROCLASS) algorithm used for image classification as explained as follows:

### A. SGLDM Feature Extraction

In order to imitate the method of radiologists read CT scans, the feature extraction is proposed and it is well–trained to perform and identify particular visual patterns to describe disease. A several methods toidentify the textural patterns in a language that a computer algorithm can understands so that these algorithms can extract features that radiologists look for to perform their diagnosis.

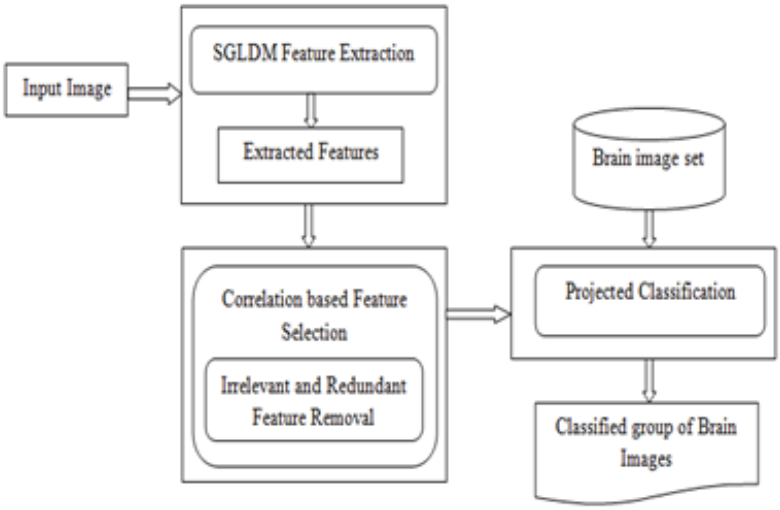


**Figure 1:** Framework architecture of proposed methodology

It is most important that the purpose of feature extraction is to enable an algorithm works like a radiologist and have a sense of humor in millions of pixels and extract only the important information from this veritable wealth of noisy information. This strategy is mainly dependent upon the estimation [5,9]tries to concentrate picture composition. Also, it offers the spatial circulation and spatial dependence around gray levels in neighborhoods of an image has been utilized within an assortment of medicinal picture transforming settings. In addition, the best techniques for surface segregation [1] can assess the effectiveness of the SGLDM and stretch out textural learning for utilization with pneumonic therapeutic picture, the ash levels are binned with the goal that the ensuing picture holds just a couple of light black levels. The main reason behind this as; the amount of gray level will improve the computational efficiency of this feature extraction approach.

The amount of gray level G in the SGLDM evaluates the second request contingent bivariate pdf f(i, j|d ,θ), where θ ∈ Θ, Θ = {0, π/ 4, π /2, 3π/ 4 }, and d ∈ {1,...,s − 1} defines how vast a territory is recognized and 's' is the length of a side of the picture then given 'θ'and 'd', $j^{th}$, $k^{th}$ component in the G × G matrix f (θ, d) is the probability that that gray level 'j' is a distance 'd' away from gray level 'k' in the direction of 'θ'. Then,extract the accompanying eight co–occurrence features from each of these SGLDMs:

1) Contrast, which measures local contrast in an image

$$f_1^d = \sum_{j=1}^{G} \sum_{k=1}^{G} |j-k|^2 f_{jk}^d \tag{1}$$

2) Color-gray level, which measures correlation of pixel pairs on gray levels

$$f_2^d = \sum_{j=1}^{G} \sum_{k=1}^{G} \frac{\left(j-p_j^d\right)(k-q_k^d) f_{jk}^d}{p_j^d q_k^d} \tag{2}$$

3) Entropy, which measures the occurrence of repeated pairs in an image

$$f_3^d = \sum_{j=1}^{G} \sum_{k=1}^{G} (f_{jk}^d)^2 \tag{3}$$

4)Coarseness, which measures an image's smoothness

$$f_4^d = \sum_{j=1}^{G} \sum_{k=1}^{G} \frac{f_{jk}^d}{1+|j-k|} \quad (4)$$

5)Variance, which measures variety of the gray level circulation

$$f_5^d = \sum_{j=1}^{G} \sum_{k=1}^{G} (i - p^d)^2 f_{jk}^d \quad (5)$$

6)Intensity, which measures the average gray level in an image

$$f_6^d = \frac{1}{2} \sum_{j=1}^{G} \sum_{k=1}^{G} (j + k) f_{jk}^d \quad (6)$$

7)Gabour Maximum probability, which determines the predominance of the most predominant pixel pair

$$f_7^d = max_{jk} f_{jk}^d \quad (7)$$

8)Cluster prominence, which measures grouping of pixels with similar gray levels

$$f_8^d = \sum_{j=1}^{G} \sum_{k=1}^{G} (j - p_j^d + k - p_k^d)^4 + f_{jk}^d \quad (8)$$

We proceed by using equations (1) to (8), then the defined features of theSGLDMmatrices for t = 1,...,8 is as follows:

$$f = \frac{1}{\lfloor \frac{S}{2} \rfloor} \sum_{d=1}^{S/2} f^d \quad (9)$$

Then, the use of Spatial Gray Level Difference Method feature extraction is listed in the following Table 1.

**Table 1:** List of extracted features by SGLDM

| S. No | Extracted Features |
|---|---|
| 1 | Contrast |
| 2 | Entropy |
| 3 | Coarseness |
| 4 | Color-gray level |
| 5 | Intensity |
| 6 | Gabour |
| 7 | Invariant |
| 8 | Cluster prominence |

### *B. CFS*

After the picture features extraction, we need to select some target offers that is important for recovery of mind picture from the characterized aggregation of cerebrum picture dataset. Then, we need to utilize the Correlation-based Feature Selection for emulating the image features which are depicted in the Table 2 from the concentrated features by CFS.

**Table 2:** The list of targeted features by CFS

| S. No | Targeted Features |
|---|---|
| 1 | Coarseness |
| 2 | Color-gray level |
| 3 | Intensity |
| 4 | Gabour |

A feature may be repetitive and that could be inferred from an alternate feature or set of features. Some of the redundancies might be located by association investigation and given two features are examined to measure how emphatically one trait suggests the other and taking into account for the accessible information. The connection between a feature/attribute and the class is sufficiently high to make it significant to (or prescient of) the class and the relationship between it and whatever possible applicable features/attributes does not achieve a level so it could be anticipated by any of the other important features/attributes.In this sense, the issue of property choice obliges a suitable measure of correspondences between features and a sound strategy to select properties dependent upon this measure, the two methodologies can measure the relationship between two arbitrary variables. One is dependent upon established straight connection and alternate is dependent upon data hypothesis [4].

***a) Symmetrical Uncertainty***

Symmetric uncertainty (SU) [10] is derived from the mutual information by normalizing it into the entropies of feature values and target classes and it has been used to evaluate the goodness of features for classification by a number of researchersdiscussed [14,16,17]. A probabilistic model of an esteemed feature V might be shaped by assessing the distinct probabilities of the qualities v € V from the preparation information. This model is utilized to gauge the worth of V for a novel specimen drawn from the same circulation as the preparation information, and the entropy of the model and thus of the feature is the amount of bits it might take. Entropy is a measure of the doubt or eccentrics in a framework. The entropy of V is given,

$$H(V) = -\sum_{v\epsilon V} p(V) log_2(p(V)) \tag{10}$$

If the noticed qualities of V in the preparation information are divided as per the qualities of a second feature U and the entropy of V concerning the segments affected by U is less than the entropy of V before apportioning. Then there is a relationship between features V and U. Mathematical equation (10) gives the entropy of V in the wake of watching U.

$$H\left(\frac{V}{U}\right) = -\sum_{u\epsilon U} p(U) \sum_{v\epsilon V} p\left(\frac{V}{U}\right) log_2\left(p\left(\frac{V}{U}\right)\right) \tag{11}$$

The measure by which the entropy of V declines reflects extra data about V gave by U and is known as the data pick up, or, on the other hand, shared data. Data increase is given by

Gain=H(V)-H(V/U)
=H(V)-H(U/V)

=H(V)+H(U) – H(U, V) (12)

Data increase is a symmetrical measure that is; the measure of data picked up about V in the wake of watching U is equivalent to the measure of data picked up about U in the wake of watching V. Symmetry is an alluring property for a measure of feature inter-correlation to have. Data addition is pre-dispositional in favor of features with additional qualities. Moreover, the associations in equation (13) have to be standardized to guarantee that they are equivalent and have the same influence. Symmetrical doubt adjusts for data addition's inclination to features with additional values and standardizes its esteem to the extent [0, 1].

Symmetric uncertainty coefficient=$2.0 \times \left[\frac{gain}{H(V)+H(U)}\right]$ (13)

### *C. Projected Classification*

The classification problem is well known in the database literature for its numerous applications, such as segmentation, clustering and trend analysis. Unfortunately, all known algorithms tend to break down in high dimensional spaces because of the inherent points. One way of handling this is to pick the closely correlated dimensions and find classified group in the corresponding subspace. To achieve this, traditional feature selection algorithmswere used. The strengthens of this approach is that in typical high dimensional data mining applications different sets of points may classify better for different subsets of dimensions. The number of dimensions in each such classified group-specific subspace may also vary. Hence, it may be impossible to find a single small subset of dimensions for all the classified groups. We therefore discuss a generalization of the classification problem, referred to as the projected classification (PROCLASS) problem, in which the subsets of dimensions selected are specific to the classified group themselves. We develop an algorithmic framework for solving the projected classification problem, and its performance is tested on medicinal image data. The PROCLASS algorithm has three phases such as Initialization, Iteration and Refinement. The detailed description on each phase is as follows:

Algorithm 1: PROCLASS
Inputs: S ($f_1$, f2,…,$f_N$, I)-the selected features of input image, I and the given Brain image data set
Output: $G_i$–Classified group of retrieval brain images.

```
        //===== Phase 1: Initialization =====
1 for i=1 to N do
2       Choose the sample set S of features randomly
3       Choose of set of data points from S
4       Find the best medoid M_i in data point
5 end for
        //===== Phase 2: Iteration =====
6       Initialize testing input data: TestSample [image]
7       Assign Training data = Brain image set [B]
8 for each I ∈ Training data
9               Find the GI: gray-scale [I]
10 end for
11 for each data point ∈ S
12      for each GI ∈ Training data
13              Initialize dimension of GI
        //Manhattan segmental distance
14              d_D(x1, x2) = Σ_{i∈D} |x_{1,i} − x_{2,i}|/|D|
15              if(d_D<min dist)
16                      return the data point M_i
        //classified group of retrieval brain images
17      return G_i
18      end for
19 end for
        //===== Phase 3: Refinement =====
20 for each data points M_i in G_i
        //Average distance
21      Y_ij = d_D/| D_i |
        //Smallest Manhattan segmental distance
22      Δ_i = min_{j≠i} d_{D_i}(M_i, M_j)
23      if( Y_ij<Δ_i )
24      remove the training image as outlier from G_i
        // best classified group of retrieval brain images
25 return G_i
```

*Phase I:Initialization*

In this phase we have to initialize the features data point which was previously selected by using CFS feature selection algorithm. For N number of features subset choose the sample set, S of features randomly and choose a set of data point then find the best method, M in data points. The reduction to the sample set, S significantly reduces the running time of the initialization phase.

*Phase II: Iteration*

Before performing the iteration, we have to prepare each image in the brain dataset for gray scale processing is shown in the Figure 2. Then initialize the testing input image data which contains only the target features and then assign the brain image data set as training data.Iteratively for each data point from the feature selector and for each training image data first, find the dimension $D_i$ of the training data second, calculate the Manhattan segmental distance

$$d_D(x1, x2) = \sum_{i \in D} |x_{1,i} - x_{2,i}| / |D| \quad (14)$$

Then, for each data point, assign it to the method $M_i$if its Manhattan Segmental Distance for dimension D is minimumthen the point will be assigned to $M_i$and return the classified group of images $G_i$.

*Phase III: Refinement*
In iterative phase, we don't handle the outliers, and now we will handle it.For each method,$M_i$ with the dimension, $D_i$ in the group,$G_i$ find the average distance, $Y_{i\,j}= d_D/|D_i|$ and find the smallest Manhattan segmental distance,$\Delta_i$ to any of the other medoids with respect to the set of dimensions, $D_i$. $\Delta_i = min_{j\neq i} d_{D_i}(M_i, M_j)$. Then if the average distance,$Y_{i,\,j}$ is less than that of the smallest Manhattan distance, then specify that training image as outlier and remove it from the classified group $G_i$.Finally, returnbest classified group of brain images.

## Experimental Result

### *A. Data Source*

For the purpose of identifying classified group of medicinal images for the given input image we have to use the Medicinal Brain image data set. Existing methodology FAST [1] concentrated on human face image data set.

### *B. Experimental Procedure*

An input image which is going to be processed is chosen from the brain image set and the input image is given to SGLDM Feature extraction phase, where the image preprocessingand feature extraction takes place. In this preprocessing phase, given input image is converted in to gray scale image and the noises from the gray scale image such as salt and pepper dots are removed. Then the automatic feature extraction takes place. The list of features which were extracted from the given image by the SGLDM method and it is shown in Table 1.After the feature extraction, the image with extracted features is given to correlation based feature selection-CFS feature selector phase is shown in the Figure 3 and specify the target features which are going to be targeted, mentioned in Table 2.Then the image with selected features and the brain image set are given to our proposed projected classification (PROCLASS) phase. Finally, the PROCLASS gives the classified group of retrieval brain images for the given input brain image. For the above example input image the first classified group contains the exact match of images then other classified group has some similar images which are related to the given input image.

## Performane Analysis

### A. *Proportion of selected features*

In Table3shows the proportion of selected features of existing FAST [12], FCBF [14], [15] algorithms and our proposed work PROCLASS with CFS. For image data the proportion of selected features of FCBF and FAST varies from 0.2 to 19.In this work, the proportion of 0.5 for the feature selection is achieved. For image data, when

compare with text and microarray data FAST ranks 3 with the proportion of selected features.

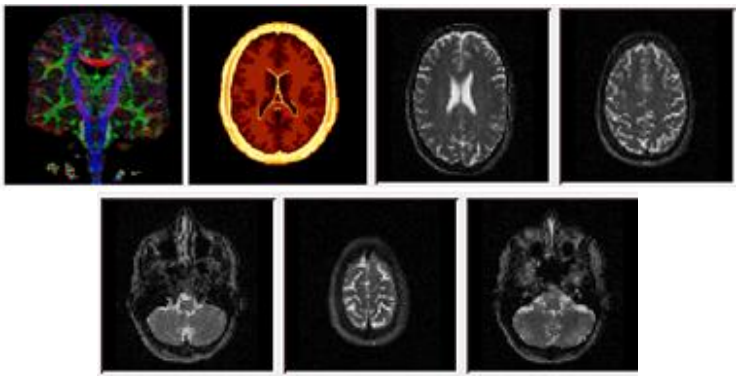

**Figure 2:** An example medicinal brain images present in the brain image dataset.

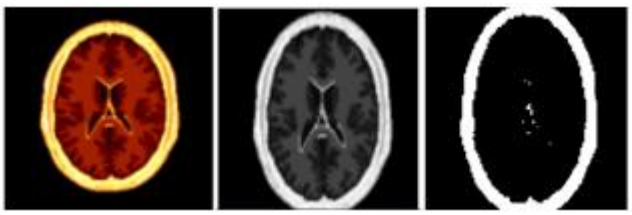

**Figure 3:** An Example of input colored brain medicinal image chosen from the brain image data set (Left), an image after gray-scale finding (middle), an image after feature extraction and selection (right)

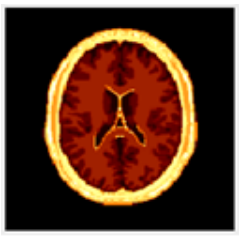

**Figure 4:** A classified group 1 of similar image for the given input image

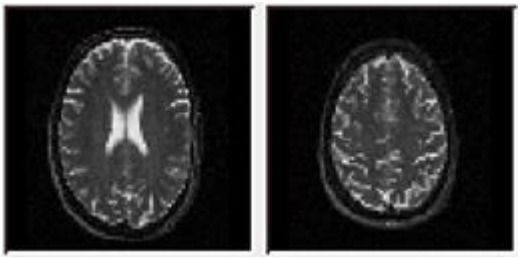

**Figure 5:** An classified group 2 of similar images for the given input image

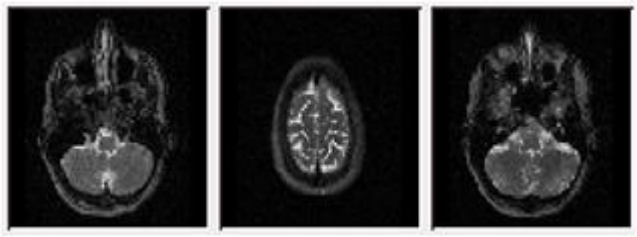

**Figure 6:** A Classified group 3 of similar images for the given input image

To solve the above mentioned problem in the proposed work CFS with PROCLASS is used to classify the image retrieval and it is shown in Figure 4, 5 and 6. The Figure 7shows that the comparison description of proportion of selected features.

**Table 3:** The comparison of proportions of selected features

| Data set(Image, Face) | FAST | FCBF |
|---|---|---|
| Mfeat-fourier | 19.48 | 49.35 |
| AR10P | 0.21 | 1.04 |
| PIE10P | 1.07 | 1.98 |
| PIX10P | 0.15 | 3.04 |
| ORL10P | 0.30 | 2.61 |
| Data set(Image, Brain) | Proposed work | |
| Brain | 0.5 | |

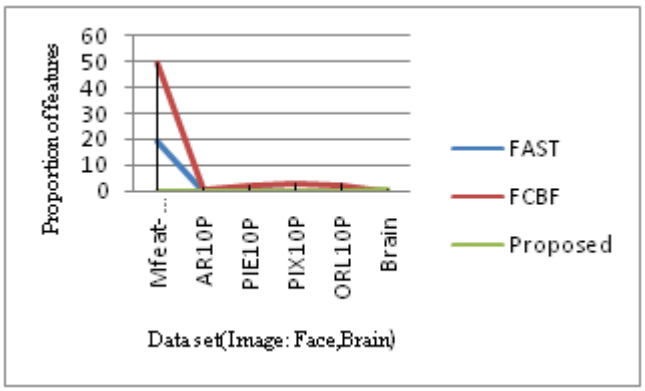


**Figure 7:** A Comparison of proportion of selected features.

**B.** ***Run time***

In Table4 shows the runtime of our method and the previous feature selection algorithms such as FAST and FCBF. The runtime for feature extraction and feature selection of our method works on quit less than or average run time whencompare

with existing methods. The Figure 8 describes the comparison of run time with existing methodologies.

**Table 4:** The comparison of runtime of framework

| Data set(*Image, Face*) | FAST | FCBF |
|---|---|---|
| Mfeat-fourier | 1472 | 716 |
| AR10P | 706 | 458 |
| PIE10P | 678 | 1223 |
| PIX10P | 2957 | 9056 |
| ORL10P | 2330 | 12291 |
| Data set(*Image, Brain*) | Proposed work | |
| Brain | 5033 | |

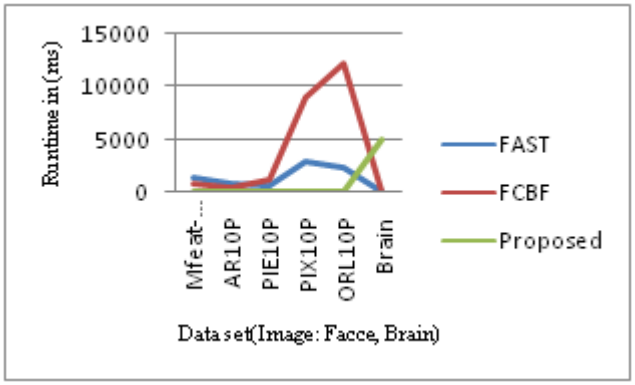


**Figure 8:** A Comparision of runtime of frameworks.

## Conclusion

In this paper a feature extraction and feature selection for content-based image classification-PROCLASS was presented, investigated, and experimentally evaluated. This work gives a review of features proposed for image retrieval and refines several of them. Our experiments is going to conclude that the optimal set of features for medicinal brain image classification tasks were carried out and the characteristics of the different features were analyzed using an empirical correlation analysis. The experimental result shows both color images of medicinal brain images as well as the gray images of medicinal images. We also compared the performance of the proposed work with previous methodologies such as FAST and FCBF. We believe that application of our work will be used for the medical analysis for classifying the similar images for learning purpose.In future work, the proposed method is planned to explore different types of medicinal images as well as some other application domains and study some formal properties of image features.

## References


[1] Battiti, R., 1994, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Transactions on Neural Networks, 5 (4), pp. 537-550.

[2] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A., 2010, Miscellaneous functions of the department of statistics (e1071), TU Wien.

[3] Fleuret, F., 2004, "Fast Binary Feature Selection with Conditional Mutual Information," J. Machine Learning Research, 5, pp. 1531-1555.

[4] Hall, M.A., 1999, "Correlation-Based Feature Subset Selection for Machine Learning," Ph.D. dissertation, University of Waikato.

[5] Haralick, R.M., and Shapiro,L.G., 1992, "Computer and Robot Vision," Addison– Wesley Publishing Co., Boston, MA.

[6] Haralick, R.M., Shanmugam, K., and Dinstein, I., 1973, "Textural features for image classification," IEEE Transactions on Systems, Man and Cybernetics, 3, pp. 610–621.

[7] Koller, D., and Sahami, M., 1996, "Toward Optimal Feature Selection," Proc. International Conference on Machine Learning, pp. 284-292.

[8] Kononenko, I., 1994, "Estimating Attributes: Analysis and Extensions of RELIEF," Proc. European Conf. Machine Learning, pp. 171-182.

[9] Liu, H., and Setiono, R., 1996, "A Probabilistic Approach to Feature Selection: A Filter Solution," Proc. 13th Internaitonal Conference on Machine Learning, pp. 319-327.

[10] Mir,A.H., Hanmandlu, M., and Tandon, S.N., 1995 "Texture Analysis of CT images," IEEE Engineering in Medicine and Biology, November/December:781–786.

[11] Press, W.H., Flannery,B.P., and Teukolsky, S.A., 1988, "Vetterling, Numerical Recipes in C,". Cambridge University Press.

[12] Qinbao Song, Jingjie Ni, and Guangtao Wang, 2013, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," IEEE Transactionson Knowledge and Data Engineering, 25 (1), pp. 1-14.

[13] Shoshana Rosskamm, B.S., 2008, "Computer–Aided diagnosis of cystic fibrosis and pulmonary sarcoidosis using texture descriptors extracted from CT images," Ph.D. dissertation, University of Colorado Denver.

[14] Yu, L., and Liu, H., 2004, "Efficient Feature Selection via Analysis of Relevance and Redundancy," J. Machine Learning Research, 10 (5), pp. 1205-1224.

[15] Yu, L., and Liu, H., 2003, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th International Conference on Machine Leaning, 20(2), pp. 856-863.

[16] Zhao, Z., and Liu, H., 2009, "Searching for Interacting Features in Subset Selection," J. Intelligent Data Analysis, 13 (2), pp. 207-228.

[17] Zhao, Z., and Liu, H., 2007, "Searching for Interacting Features," Proc. 20th International Joint Conference on Artificial Intelligence.