Optimum Path Forest Approach for Image Retrieval based on Context

Shrikant Pandurang Dhawale
Department of InformationTechnology
MAEER's MIT
Pune,India
shrikantdhwl7@gmail.com

Bela Joglekar
Assistant Professor
Department of Information Technology
MAEER's MIT
Pune, India
bela.joglekar@mitpune.edu.in

Abstract

CBIR System consist of large datasets with millions of image samples for statistical analysis, hence putting tremendous challenge for pattern recognition techniques, which needs to be more efficient without compromising effectiveness. The image samples are stored in a database in the form of feature vectors. Pattern Recognition Technique requires a high computational burden for learning the discriminating functions that are actually responsible to separate the samples from distinct classes. Many efforts have been taken to employee machine learning algorithm in a classification problem, such as support vector machine, Artificial Neuronal Network - Multi-Layer Perceptron and k-Nearest Neighbour, but all of them have usual problem of high computation burden for a training of dataset, also training becomes unrealistic due to huge training size. A novel approach is presented to reduce this problem by means of fast computation of optimum path forest in a graph derived from training samples. Each class is denoted by a multiple tree rooted at some representative samples. This Optimum Path Forest is a classifier which assigns to new sample the label of its most strongly connected root from representative samples.

Keywords- Optimum Path Forest; Minimum Spanning tree; Support Vector Machine; Pattern Recognition.

I. INTRODUCTION

Content Based Image Retrieval based on feedback tries to retrieve most relevant images in database. As the nature of problem is highly dynamic depending on relevance, and the users for same query varies vastly, these system generally stands on an active learning paradigm. In active learning paradigm, system first returns the small image set and user indicates their relevance at each iteration. Large image collections are available for Content Based Image Retrieval which requires user feedback, retraining and interactive time response during iterations. Existing classifier such as support vector machines and artificial neuronal network require more computational time for large datasets especially in training phase. The usual problem of high computation burden for training is there for these classifiers. Although the support vector machine based classifier have achieved high recognition rate in several applications, when training set becomes large its learning becomes unrealistic due to burden of huge training size over classifier. Other classifiers such as Artificial Neuronal Network with Multilayer Perceptron, Radial Basis Functions, Self Organizing Maps and Bayesian Classifiers have same problem[1].

It is necessary to have more efficient and effective pattern recognition method for large datasets. Keeping this in view, a new novel approach[5] for pattern classifier based on fast computation of an optimum path forest derived from training sample is proposed in order to overcome such challenges. Optimum Path Forest can obtain similar effectiveness to above mentioned machine learning techniques and can be faster in training phase.

Content Based Image Retrieval System consist of image database represented by feature vectors (points in a feature space) which encodes colour, texture and/or shape measure of images. Similarity can be measured between their feature vectors. For given query image, Content Based Image Retrieval System ranks more similar images based on their distance to query images. However a semantic gap usually occurs in retrieved result and users expectations. To reduce semantic gap problem, feedback based learning approach is presented. User indicates which images are relevant or irrelevant based on small set of returned images and then content based image retrieval system understands how to better rank and return more relevant images in further steps. This search process is repeated until the user is satisfied. So to retrieve more relevant images from large datasets Optimum Path Forest Classifier is used. This is illustrated in Fig 1 which shows typical CBIR system based on context. Here user first gives query image and based on similarity measure small set of images is returned, user marks those images about the relevance, hence forming the small set of labelled training set. This is given to OPF classifier model to retrieve most relevant images from image database.

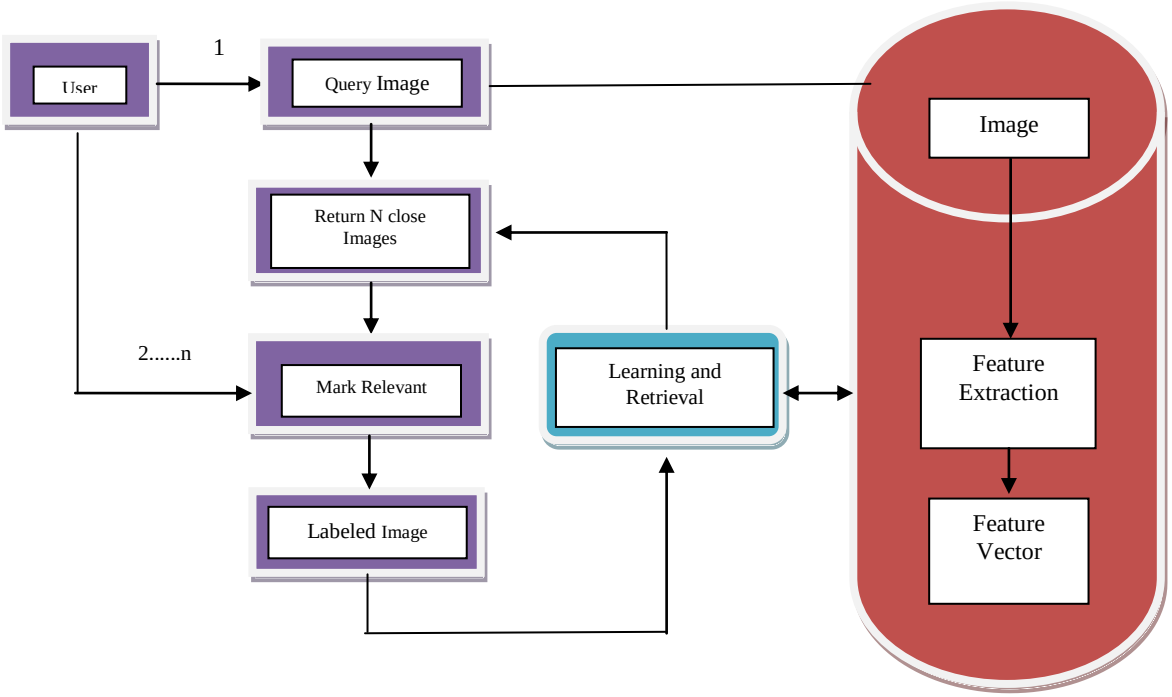


Figure 1. Content Based Image Retrieval based on context

This paper is organised as follows. Section II provides a survey of a work done in similar area, Section III provides the actual Content Based Image Retrieval based on context and Optimum Path Forest Classifier, Section IV provides Pattern Classification using Optimum Path Forest.

II. Related work

Efforts have been taken over the past decades to improve efficiency and effectiveness of the Content Based Image Retrieval System, Some example are IBM QBIC, UIUC, MARS, PicHunter, TinEye and Windsurf. The research on Content Based Image Retrieval starts from text retrieval. The main contribution is the precision×recall curve which is used to evaluate the effectiveness of content based image retrieval. Other contribution was feedback based learning which involves human intervention[3].

Efforts in Content Based Image Retrieval is divided into two goals of improvement, these are efficiency and effectiveness. Efficiency can be measured with the help of involved indexing structure helps accessing image in more efficient manner and also better scalability to huge image collections with reduction in a number of dimensions of a search space before indexing data. Effectiveness can be obtained by reducing semantic gap problem. Recently, focus is on Feedback based Active Learning which improves ranking of images, as well as Support Vector Machine, Artificial Neuronal Network, k-Nearest Neighbors, Optimum Path Forest, using several datasets and descriptor. Among all of them Optimum Path Forest is good classifier in computational time which is important in large dataset. Optimum Path Forest is more or less accurate than Support Vector Machine depending on the case, but accuracy is superior to those of Artificial Neuronal Network, k-Nearest Neighbor. Optimum Path Forest is simple, multi class and parameter independent which does not make any assumption of shape of the classes and can handle some degree of overlapping between classes[5].

Optimum Path Forest classifiers are being used in real applications: the supervised approach is used for oropharyngeal, dysphagia identification, laryngeal pathology detection and diagnosis of parasite from optical microscopy images. In all above examples Optimum Path Forest out perform in terms of accuracy and efficiency[5].So applications with large datasets definitely favours Optimum Path Forest approach with respect to support vector machine.

III. Framework for Content Based Image Retrieval based on context And Optimum Path Forest Classifier

Fig.3 shows learning by Optimum Path Forest Classifier using Feedback Technique. The simple descriptor D=(v,d) shows how nodes are spread in feature space illustrated in Fig.2. Here 'v' is feature extraction function which extracts the feature of images and d(s,t) the distance function between two image representations. In the first iteration, images from database 'Z' will be returned by similarity for given query image 'q'. System simply ranks some 'N' closest

images (t Є Z) in the increasing order of d(s,t) with respect to query image q. Goal is to return image list 'X' with 'N' most relevant images in Z with respect to query image 'q'. Problem is limitation of descriptor (v,d) which fails to represent the users expectation called as semantic gap, such that 'X' contains relevant and irrelevant images according to users opinion.

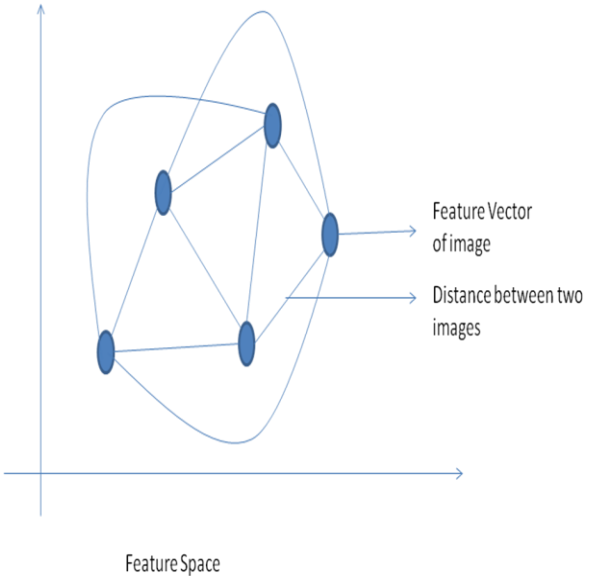


Figure 2. Feature Space with Image Descriptor (v,d)

The Active Learning approach circumvents semantic gap problem by taking user's feedback about the relevance of returned images during few iterations[3]. Here, user marks which images are relevant (or irrelevant) in 'X', thereby forming a labelled training set 'T' which gains new images at each iterations so T←T ∪ X[3].

On this training set we model our Optimum Path Forest Classifier. The process consist of estimating prototypes(adjacent node with different class labels) first. Then training process considers a complete graph whose nodes are all element in T and arcs between them is given by d(s, t), then grow minimum spanning tree over underlying graph is formed.

Every path in graph has a certain cost define by f_{max} and minimum cost paths are computed from prototypes to each image node tЄT such that classifier is Optimum Path Forest rooted in prototype. In this forest, the nodes which belongs to training set but not prototype are conquered and labelled by the prototype which offer the minimum cost path with terminus t. After this, classifier is used to evaluate the images in Z\T by computing the cost of optimum path from prototype to each image node tЄZ\T in incremental way.

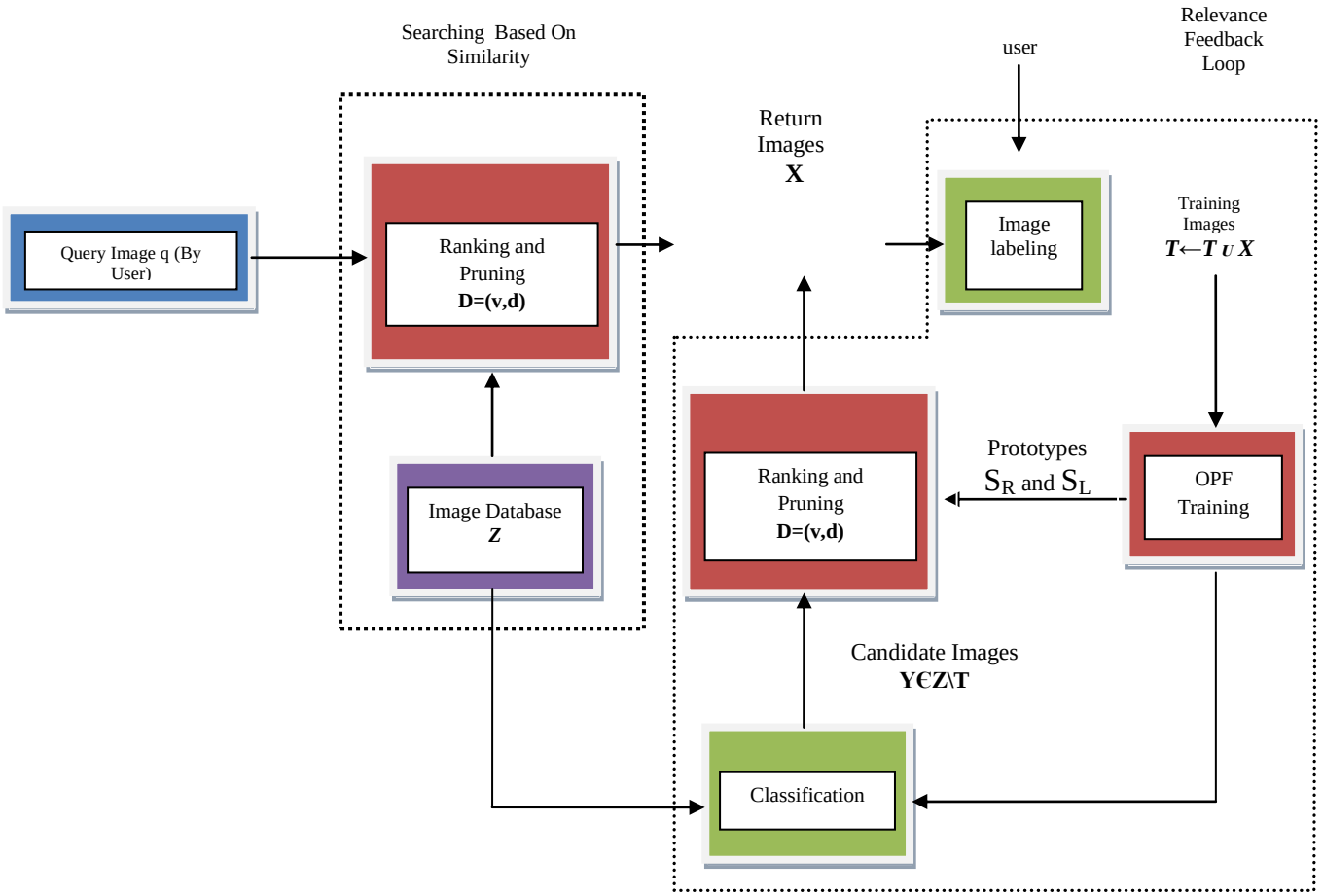


Figure 3. Framework for CBIR based on Context

IV. PATTERN CLASSIFICATION USING OPTIMUM PATH FOREST CLASSIFIER

Optimum Path Forest Classifier works by transforming the problem of classification as graph, partitioned in a given feature space. The nodes are represented by feature vectors and edges connect all pairs of them, defining a full connected graph. Competition process then partitions the graph between some key samples (prototypes) which offer optimum path to remaining nodes of the graph. The Optimum Path Forest can be seen as a generalization of well known Dijkstra's algorithm to compute optimum paths from source node to remaining nodes[15].The main difference relies on the fact that Optimum Path Forest uses a set of source nodes (prototypes) with any smooth path cost function [16].

A. Background Theory

Let Z=Z1∪Z2 be a dataset labelled with function λ, Z1 is training sets and Z2 is test sets used to train the classifier and to calculate its accuracy respectively. Let S, subset of Z1 is set of prototype samples. Now a smooth path cost function 'f' is defined as:

Given a sample t, there exist an optimum path π that is trivial or formulated as $\pi_s.<s, t>$, where

- $f(\pi_s) \leq f(\pi_t)$,
- π_s is optimum,
- and for some optimum path τ_s,
  $f(\tau_s.<s, t>)=f(\tau_t)$,

π is sequence of samples. Optimum Path Forest can be used with any metric since smooth path cost function is used. Here consider a path cost function f_{max} which is computed as:

$$f_{\max}(<s>) = \begin{cases} 0 & if \quad s \in S, \\ +\infty & otherwise, \end{cases}$$

$$f_{\max}(\Pi_s . \langle s,t \rangle) = \max\{ f_{\max}(\Pi_s), d(s,t)\}$$

here d(s, t) is distance between sample s and t and path π is defined as sequence of adjacent samples. Optimum path Forest Classifier is composed by two distinct phases:

- Training
- Classification

B. Training

By computing an Minimum Spanning Tree in the complete graph (Z_1,A), a connected acyclic graph whose nodes are all samples of Z_1 and the arcs are undirected and weighted by the distances d between adjacent samples is obtained. The spanning tree is optimum since the sum of its arc weights is minimum as compared to any other spanning tree in the complete graph. In the minimum spanning tree, every pair of samples is connected by a single path that is optimum according to f_{max} [17].

That is, the minimum spanning tree contains one optimum path tree for any selected root node. The optimum prototypes are the closest elements of the Minimum Spanning Tree with different labels in Z_1. After finding prototypes, the competition process is accomplished in order to build the Optimum Path Forest. Although the Minimum Spanning Tree can provide the optimum set of prototypes, they might not be unique. Additionally, if only one Minimum Spanning Tree is there (all arc-weights are different from each other, which is really difficult in practice), this set of prototypes would be unique, and thus we would have no classification errors in the training phase. Therefore, the computed optimum paths might not be unique.

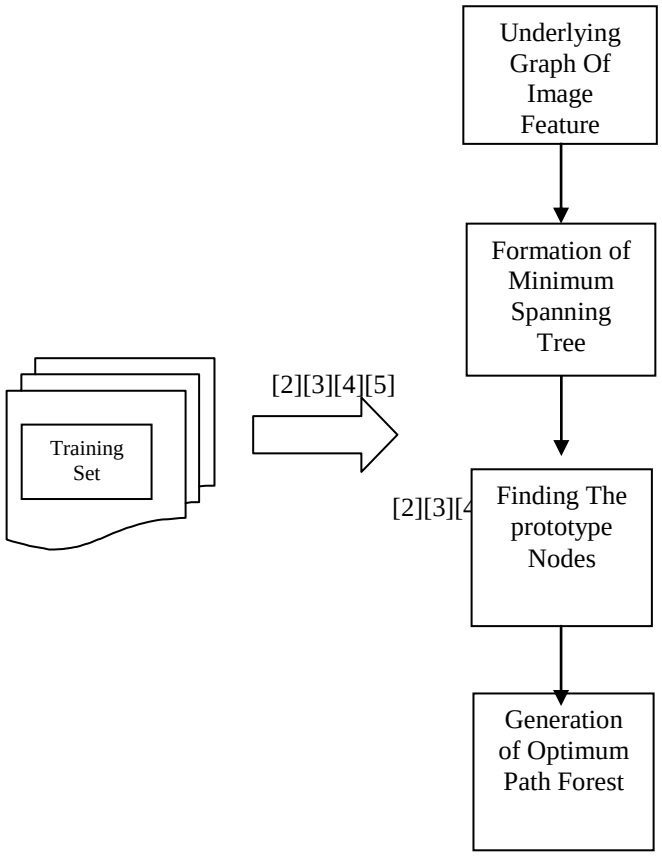


Figure 4. Optimum Path Forest Training phase

C. Classification

For any sample tЄZ2, and all arcs connecting t with sample sЄZ1 Considering all possible paths from prototype set S to t, the optimum path P(t) from prototype S and label t with the class λ(R(t)) of its most strongly connected prototype R(t)Є S can be found incrementally by evaluating the optimum cost C(t) as.

$$C(t) = \min\{\max\{C(s),\ d(s,t)\}\},\ \ \forall s \in Z1$$

Algorithm 1. Optimum Path Forest Algorithm[3]

INPUT: the input to algorithm is training set Z1, the set of relevant and irrelevant prototypes S_RC T and S_L C T and descriptor (v,d).
OUTPUT: Optimum path forest i.e. map containing the information of predecessor P, path cost map C, root map R, and the ordered list of training nodes
AUXILIARY: Priority queue Q and cost variables cst.

1. **for** each sample image (s) belongs to training(T) set but not prototype(S) **do**
2. set cost of sample C(s) to infinity.
3. **end**
4. **for** each sample image(s) belongs to prototype set(S) **do**
5. set cost of sample C(s)=0, predecessor of sample P(s)=nil,
6. root of sample R(s)= sample itself (s) and insert s into queue,
7. **end**
8. **While** queue Q becomes not empty **do**
9. Remove from priority Q a node s with minimum cost C(s) and insert s in T
10. **for** each t ЄT such that cost C(t) >C(s) **do**
11. compute cst=maximum (C(s),d (s, t))
12. **if** cst<C(t**) then**
13. P(t)=s, R(t)=R(s) and C(t)=cst.
14. **if** C(s)= infinity
15. insert t in Queue Q.
16. **else**
17. update the position of t in Q.
18. **end**
19. **end**
20. **end**
21. **end**

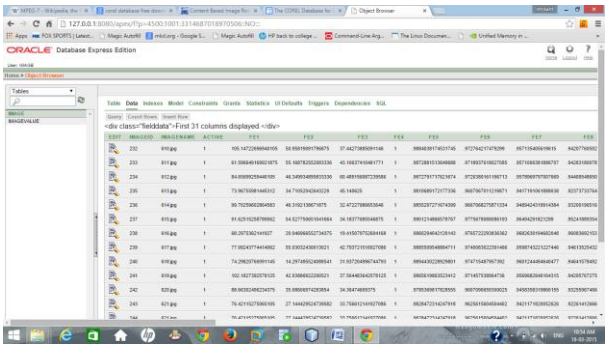
Description:
line 1-7 initialize maps and tries to insert prototypes in queue. line 8--21 main loop computes optimum path from S(prototype set) to every sample s(belongs to training but not prototypes) in increasing order of minimum cost. S* is an optimum set of prototype when algorithm minimizes classification error in training set. S* can be found by exploiting the theoretical relation between MST [6] and optimum path tree for f_{max}[2]. by computing the minimum spanning tree in complete graph of training sample, a connected acyclic graph whose nodes are all sample of training set and the arc are undirected and weighted by the distance d between adjacent sample is obtained. The spanning tree is optimum in the sense that sum of its arc weights is minimum as compared to any other spanning tree in complete graph. In the formed MST, every pair of sample is connected by a single path which is optimum according to f_{max}. That means MST contains one optimum path tree for any selected root node. The Optimum prototype are the closest element of MST with different labels in the training set. By removing the arc between different classes, their adjacent samples become prototypes in S* and this algorithm tries to compute Optimum Path Forest in training set, also it is worth to note that a given class may be represented by multiple prototypes(i.e., optimum path trees) and at least one prototype per class must exist. [2][3][4][5]

Complexity Analysis:
OPF can be divided in two phases, training and classification.
The most part of computational efforts is expend in the training, in which essentially just Optimum Prototypes are computed (closest sample in MST) and then OPF is run to form Optimum Path forest. Now consider |n| and |e| are the number of samples and edges, respectively, in the training set represented by the underlying graph of samples. MST is computed using Prim's algorithm with complexity O (|e| log |n|) and to find the prototype in O (|n|). The OPF algorithm main (line 8) and inner for loop (line 10) runs in O (|n|) times each one because of the complete graph. So OPF run in O ($|n^2|$). Overall training steps complexity can be executed in O (|n| log |e|) +O (|n|) +O ($|n^2|$), which is dominated by O ($|n^2|$) [6].
The classification step can be done in O ($|Z_2|.|n|$), in which $|Z_2|$ is the test set size. a sample tЄZ_2 to be classified is connected to all samples of training set and the optimum path cost is evaluated. Since OPF classifier can be understood as dynamic programming algorithm [6], there is no need to execute the OPF again in test phase, because each node has already stored its optimum path value, there is need to evaluate paths from each node of training until t. Finally, the estimated OPF complexity is O ($|n^2|$) + O ($|Z_2|.|n|$) [6].

V. IMPLEMENTATION DETAILS

In order to analyze the effectiveness of the proposed method, the experiment is performed using test dataset containing 1000 images [19] with diverse set of samples. Presently fifty images are taken into consideration for testing first module of project i.e. Feature Extraction. Result of this module is shown below assuming .jpg images for feature extraction. For implementation of first module, a comparative study of low level feature extraction algorithms for the said images is done. F-SIFT (Fast Scale Invariant Feature Transform) algorithm is chosen which is fast and has good performance than others like SURF, SIFT, PCA-SIFT [18].

Feature Vectors extracted using F-SIFT:



As shown above for the feature extraction module, descriptors are obtained for each image in the form of feature vector, quantized to the nearest approximate value. This enables easy retrieval which is essential for scaling in large databases. The performance parameters like Precision and Recall will be the major decision factors in testing the effectiveness of image retrieval.

VI. CONCLUSION

Supervised OPF classifier computes an optimum-path forest on a training set and classifies samples with the label of their most strongly connected root in the forest. A supervised learning algorithm, usually improves performance of the classifiers without increasing the training set. The advantage of OPF over the others in computational time is significant, which is crucial in the case of large datasets while retrieval. It can be more or less accurate than SVM, depending on the case, but its accuracy is usually superior to those of ANN-MLP and k- NN. OPF also presents some interesting properties. It is fast, simple, multi-class, parameter independent, does not make any assumption about the shape of the classes, and can handle some degree of overlapping between classes. The performance of OPF may be reduced for small training sets, if the number of samples are not enough to represent the classes. In SVM, this may also be a problem, but as it estimates a decision hyper plane, it has a chance to divide the feature space with separation between classes. Too much overlapping between classes may also represent an advantage for SVM with respect to OPF, because its transformation to a higher dimensional space may separate the classes, solving the problem.

REFERENCES

[1] A. T. Silva, A. X. Falcão, L. P. Magalhães, Active learning paradigms for CBIR systems based on Optimum path Forest classification, Pattern Recognition, Elsevier, 44 (2011) 2971–2978.

[2] J.P. Papa, A.X. Falcão, V.H.C. Albuquerque, J.M.R.S. Tavares, Efficient supervised optimum path forest classification for large datasets, Pattern Recognition, Elsevier, 45 (2012) 512–520.

[3] André Tavares da Silva, Jefersson Alex dos Santos, Alexandre Xavier Falcão, Ricardo da S. Torres, Léo Pini Magalhães, Incorporating multiple distance spaces in optimum-path forest classification to improve feedback-based learning, Computer Vision and Image Understanding, Elsevier, 116 (2012) 510–523.

[4] A.S. Iwashita, J.P. Papa, A.N. Souza, A.X. Falcão, R.A. Lotufo, V.M. Oliveira,Victor Hugo C. de Albuquerque, João Manuel R.S. Tavares, A path- and label-cost propagation approach to speed up the training of the optimum-path forest classifier, Pattern Recognition Letters, Elsevier, 40 (2014) 121-127.

[5] J.P. Papa, A.X. Falcão, C.T.N. Suzuki, Supervised pattern classification based on optimum path forest, Int. J. Imaging Syst. Technol. 19 (2) (2009) 120–131.

[6] Greice M. Freitas, Ana M. H. Avila, J. P. Papa, A. X. Falcao, Optimum Path Forest Based Rainfall Estimation, 2009, IEEE.

[7] J. P. Papa, A. X. Falcao, A. M. Levada, D. Correa, D. Salvadeo, N. D. A. Mascarenhas, Fast and accurate holistic face recognition through optimum path forest, in: Proceedings of the 16th International Conference on Digital Signal Processing, Santorini, Greece, 2009, pp.1–6.

[8] J. P. Papa, A. X. Falcao, Fabio A. M. Cappabianco, Optimising Optimum Path Forest Classification for Huge Datasets, International conference on pattern recognition, IEEE,2010.

[9] Roberto Souza, Roberto Lotufo, Leticia Rittner, A Comparison between Optimum Path Forest and k-Nearest Neighbors Classifiers, Conference on Graphics, Patterns and Images, IEEE, 2012.

[10] Luis C. S. Afonso, J. P. Papa, Apaarecido N. Marana, Ahmad poursaberi and Svetlana N. Yanushkevich, A fast scale iris database classification with optimum path forest technique: A Case Study,WCCI, IEEE,2012.

[11] L. C. S. Afonso, J. P. Papa, A. N. Marana, A. Poursaberi, S.Yanushkevich and Gavrilova, Optimum Path Forest Classifier for large scale biometric Applications, third international conference on Emerging Security Technologies,IEEE,2012.

[12] R. Pisani, K. Costa, R. Nakamura, C. Pereira, G.Rosa, J. Papa, Automatic Land Slide Recognition through Optimum Path Forest,IEEE,2012.

[13] Luis A. M. Pereira, J P Papa, Jurandy Almeida, Ricardo da S. Torres, William Paraguassu Amorism, A multiple labeling based Optimum Path Forest for Video Content Classification, conference on graphics, pattern and images,IEEE,2013.

[14] William Paraguassu Amorism, Marcelo Henriques de Cavalho, Valguima V. V. A. Odakura, Face Recognition Using Optimum Path Forest Local Analysis, Brasilian Conference On Intelligent Systems,IEEE,2013.

[15] E.W. Dijkstra, A note on two problems in connexion with graphs, Numerische Mathematik 1 (1959) 269–271.

[16] Falcão, J. Stolfi, R.A. Lotufo, The image foresting transform theory, algorithms, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 19–29.

[17] Alléne, J.-Y. Audibert, M. Couprie, R. Keriven, Some links between extremum spanning forests, watersheds and min-cuts, Image Vision Comput. 28 (2010) 1460–1471.

[18] M.M. El-gayar, H. Soliman, N. meky, A comparative study of image low level feature extraction algorithms, Egyptian Informatics Journal, Elsvier (2013) 14,175-181.

[19] James Z. Wang, Jia Li, Gio Wiederhold, ``SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol 23, no.9, pp. 947-963, 2001.