Information Theoretic Approach on Supervised Feature Extraction for Tensor Objects

Chithra C Sekhar[1*], Jomy George[2], Meera Krishna G. H[3]
[1*]PG Scholar, Dept. of Computer Science and Engineering
[2]Asst.Professor, Dept. of computer science and engineering
[3]Asst.Professor, Dept. of computer science and engineering
TKM Institute of Technology, Kollam, India

***Abstract*—Image Classification is an important problem in several areas such as recognition of faces, handwritten digits, objects etc. In the existing methods for feature extraction and classification, the objects are considered as vectors but in modern applications the input data are usually treated as tensors. Extracting maximum discriminating features from the tensor objects and its classification are challenging problems in machine learning and pattern recognition. This work proposes a novel scheme for supervised feature extraction of tensor objects based on maximization of Tsallis mutual entropy. Several experiments show that the proposed approach results superior accuracy in both feature extraction and classification.**

***Index Terms*——Image Classification, Feature Extraction, Tensor decomposition, Tsallis Mutual information, KNN Classifier.**

I. INTRODUCTION

Classifying face images, handwritten digits and images of objects find immense application in several fields. Classical methods for feature extraction treat inputs as vectors. But this may lead to several problems such as increasing dimensionality, small sample size and computational burden. In most of the modern applications, image data are usually represented by multi-way arrays (tensors) [1]-[3]. In many applications the input image data may be too large and may consist of redundant information. In order to design optimal classifiers, we need to extract discriminating features from the input data. Several supervised feature extraction algorithms have been recently proposed for tensors [4]-[7].These algorithms are generalization of Linear Discriminant Analysis (LDA) to tensor objects and which uses only the second order statistics of the data. Extracting features by maximizing the mutual information (MMI) overcomes this problem and provides highly discriminating features [8, 9].

Shannon's definition of mutual information is used in [8] but it has some inherent limitations. The traditional use of Shannon entropy in Information theory may not be well applied in some situations. Tsallis has proposed a new concept of entropy which extends the preceding traditional Shannon's theory. This new concept, called non extensive entropy, was used recently for image segmentation and other related areas [10]-[12]. In this paper, our primary goal is to study the usefulness of the Tsallis entropy by comparing it to the classic Shannon entropy in the context of image classification. MMI based on Shannon entropy discussed by [8] provides a single discrimination measure for optimization. In order to provide a wide class of measures we propose a method called maximization of parametric Tsallis mutual entropy for extracting the most discriminative features from tensor objects. Shannon entropy is a particular case of Tsallis entropy and varying its parameter results different objective functions for optimization.

A series of experiments were carried out for the problem of classifying image patterns under different values of the entropy parameters. For classification we use KNN classifier for the sake of simplicity. Our goal is to assess how well the different entropies can be used for feature extraction and hence to determine the class of a new test sample. The experiments show that the Tsallis entropy has great advantages over Shannon entropy for pattern classification.

The rest of the paper is organized as follows. Section II will provide some notations and basic concepts of feature extraction by MMI. Section III describes the proposed method. Section IV provides performance analysis of the proposed method. Section V contains conclusions.

II. BASIC CONCEPTS AND RELATED WORKS

In this section, we will provide some basic notations of tensor objects and also introduce some methods such as maximization of mutual information for feature extraction using Shannon's entropy.

A. Notations

Tensors are geometric objects and it is a multi-way generalization of vector and matrix. The order of tensor is the dimensionality or number of indices needed to represent it. For example, tensor $\chi \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$ is an N-way tensor. A tensor χ can be decomposed by Tucker decomposition and can be expressed as $\chi \approx F \times \{A\}$, where A are factor matrices and F is the core tensor [8].

B. Maximization of mutual information

Maximization of mutual information is considered as the more general criteria for extracting the most discriminative features from tensor objects [8]. Let χ denote a three-way random tensor and y denote its corresponding class label. Then χ can be represented through Tucker decomposition in terms of projection matrices and core tensor as

$$F = \chi \times_1 U^{(1)\mathrm{T}} \times_2 U^{(2)\mathrm{T}} \times_3 U^{(3)\mathrm{T}} \quad (1)$$

The elements of the core tensor are gives the features that can be used for classification. But our aim is to find out the most discriminative features for classification. In order to get such elements of the core tensor, we need to find out the projection matrices which maximize the mutual information measure. Jukic and Filipovic [8] proposed an iterative method for obtaining the mode-n projection matrix by solving the following optimization problem

$$U^{(n)} = \arg \max_{U^{(n)T} U^{(n)}=1} \tilde{I}_n(f, y) \quad (2)$$

where I_n is the mutual information based on Shannon measure of entropy

C. Estimation of Mutual Entropy via Shannon's entropy and Tsallis entropy

For a continuous random variable X with probability density function $f(x)$ with finite or infinite support $\mathcal{X}$ .The Shannon entropy H(X) of a random variable X is defined by

$$H(X) = -\int_{x \in \mathcal{X}} f(x) \log_2(f(x)) dx \quad (3)$$

The entropy measure H(X) quantifies the average uncertainty associated with the random variable X. The conditional entropy measures the average uncertainty associated with X, if we know the outcome of Y, which is defined as,

$$H(X|Y) = -\int_{x \in \mathcal{X}} \int_{y \in \mathcal{Y}} f(x,y) \log_2(f(x|y)) dxdy \quad (4)$$

where, $f(x,y)$ is the joint probability density and $f(x|y)$ is the conditional probability.

The mutual information (MI) between X and Y is defined by

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (5)$$

Mutual information quantifies the information gain or the shared information between X and Y.

Generalized Shannon entropy was given by Tsallis and can be expressed as [13]

$$H_\alpha^T(X) = \frac{1}{\alpha - 1}\left(1 - \int_{x \in \mathcal{X}} f(x)^\alpha dx\right) \quad (6)$$

where, α is the entropy parameter and when $\alpha = 1$ Tsallis entropy reduces to Shannon Entropy.

Mutual information can be generalized by Tsallis mutual entropy. The Tsallis mutual entropy is defined for $\alpha > 1$ as [14]

$$\begin{aligned} I_\alpha^T(X;Y) &= H_\alpha^T(X) - H_\alpha^T(X|Y) \\ &= H_\alpha^T(Y) - H_\alpha^T(Y|X) \\ &= H_\alpha^T(X) + H_\alpha^T(Y) - H_\alpha^T(X,Y). \end{aligned} \quad (7)$$

III. PROPOSED METHOD

In order to extract the most discriminating features in a tensor objects we have to use generalized mutual information criteria. This will provide a range of measures depending on the entropy parameter. In this paper we propose a supervised feature extraction algorithm for tensor objects by maximizing the Tsallis mutual information. Approach is similar to [8] but Tsallis entropy is used instead of Shannon entropy. Performance of the classification algorithm using the extracted features is examined by varying the entropy parameters including the Shannon counterpart.

Now we will discuss the estimation of Tsallis mutual information and its gradient of a scalar random variables. Negentropy of a random variable f is defined as

$$\mathcal{J}_\alpha^T(f) = H_\alpha^T(f_{Gauss}) - H_\alpha^T(f) \quad (8)$$

Tsallis Mutual information between scalar random variable f and y can be expressed as

$$I_\alpha^T(f,y) = \frac{(2\pi e\sigma_f^2)^{(1-\alpha/2)}}{1-\alpha} - \mathcal{J}_\alpha^T(f) - \sum_{k=1}^{c} P(y=k)\left[\frac{(2\pi e\sigma_{f|y=k}^2)^{(1-\alpha/2)}}{1-\alpha} - \mathcal{J}_\alpha^T(f|y=k)\right] \quad (9)$$

where $\mathcal{J}_\alpha^T(f)$ is the negative entropy, σ_f^2 is the variance and $P(y=k)$ being the probability of y belonging to class k. Gradient of $I_\alpha^T(f,y)$ with respect to W is given by

$$\begin{aligned} \tilde{\nabla}_W I_\alpha^T(f,y) &= \tilde{\nabla}_W I_\alpha^T(W^T X, y) = \\ &\sum_{k=1}^{c} P(y=k)\tilde{\nabla}_W \mathcal{J}_\alpha^T(f|y=k) - \tilde{\nabla}_W \mathcal{J}_\alpha^T(f) - \\ &\sum_{k=1}^{c} P(y=k)\left[\frac{(2\pi e\sigma_{f|y=k}^2)^{(-(1+\alpha)/2)} C_{X|y=k} W}{1-\alpha}\right] \end{aligned} \quad (10)$$

where C is the covariance matrix estimated using the training data set. Let's derive expression for negative entropy and its gradient based on nonpolynomial approximation discussed by [15]. Following the similar steps of [15] we have

$$H(x) \approx \frac{1}{\alpha - 1}\left\{1 - \int (\hat{p}_x(u))^\alpha du\right\} \quad (11)$$

Cumulants in Equation (5.30) of [15] are very small and thus we can use an approximation

$$(1+\varepsilon)^\alpha = 1 + \alpha(\varepsilon - \varepsilon^2/2) \quad (12)$$

Following Equation (5.33) and (5.34) of [15] with Tsallis entropy we get

$$\mathcal{J}_\alpha^T(f) \approx \alpha \mathcal{J}(f) \quad (13)$$

$$\tilde{\nabla}_W \mathcal{J}_\alpha^T(f) \approx \alpha \tilde{\nabla}_W \mathcal{J}(f) \quad (14)$$

A. System Architecture

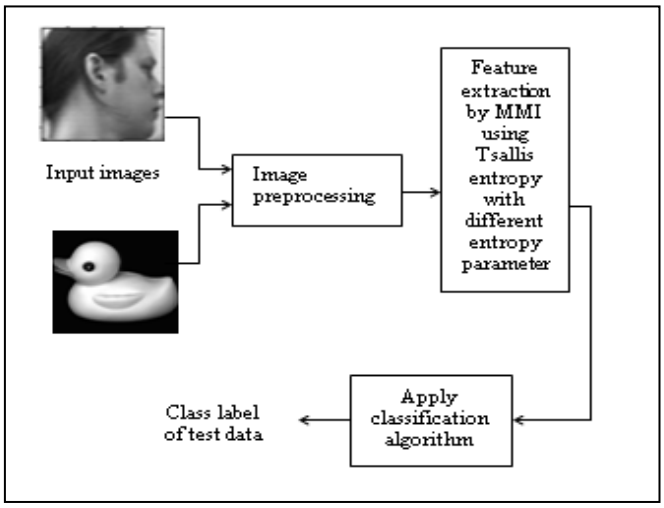

Fig. 1. Image classification

Algorithm 1: Feature Extraction

Input: 1. Set of K training samples,

$$\{\chi_k \in \mathbb{R}^{r_1 \times \ldots \times r_n}\}, k \in \{1, \ldots . C\}$$

2. Class labels y_k

Parameters: Features matrix (tensor) size for each mode($(p_1, \ldots, p_N)$), entropy parameter α with different values

Initialize $U^{(n)} \in \mathbb{R}^{r_n \times p_n}, n \in \{1, \ldots, N\}$

Repeat

For n = 1 to N

Compute

$$Z_k^{-n} = \chi \times_{-n} \{U\}^T$$

Find the mode-n matrix using the optimization procedure

$$U_{(n)} = \arg \max_{U^{(n)T} U^{(n)} = 1} I_\alpha^T(f, y)$$

End

Until (convergence)

Output: Projection matrices $U^{(n)} \in \mathbb{R}^{r_n \times p_n}, n \in \{1, \ldots, N\}$

Optimization procedure

Input: Feasible initial projection matrix U^n

$k \leftarrow 0$

Repeat

Calculate gradient $\nabla_{U^{(n)}} \tilde{I}_\alpha^T(f, y)$

Calculate A, with $A := GU^{(n)T} - U^{(n)} G^T$

$$G = -\nabla_{U^{(n)}} \tilde{I}_\alpha^T(f, y)$$

Select the step size τ_k using curvilinear search

Update with $U^{(n)} \leftarrow Q(\tau_k) U^n$

$$Q(\tau_k) := (I + \frac{\tau}{2} \tilde{A})^{-1} ((I - \frac{\tau}{2} A)$$

Until $\left(\left\| \nabla_{U^{(n)}} I_\alpha^T(f, y) \right\|_F \leq \text{tolerence} \right)$

Output: New projection matrix $U^{(n)}$

Algorithm 2: Classification

Input:. $D = \{(x_1, c_1), \ldots, (x_N, c_N)\}$

$X_p = (x_{p1}, \ldots, x_{pm})$ new instance to be classified

1. Start

2..For each (x_i, c_i) calculate the Euclidean distance $d(x_i, x_p)$

3. Order $d(x_i, x_p)$ from lowest to highest $i = 1, \ldots, N$

4. Select k- nearest instances to X_p

5. Assign X_p into the most frequent class in D

6. Stop

Output: Class label of the most frequent class

IV. Performance Evaluation

This work mainly focuses on supervised feature extraction from tensor objects. Here we take images of objects and faces as inputs. Optimal features are extracted from these input images by maximization of mutual information criteria using Tsallis entropy. A comparative performance analysis of the feature extraction method is evaluated in the context of classification by varying the entropy parameter α from 1.25 to 3 with an increment of 0.25. Shannon entropy is the special case when the parameter α tends to 1. One of the simple well known classifiers such as KNN are used for classification purpose .Performance evaluation under different images and feature dimensions in object recognition and face recognition applications using KNN is given in Tables I to IV and Figures 4 and 5 .

In order to assess the performance of the proposed work several experiments are performed on the standard datasets with images of objects and face images. The Columbia University Image Library (COIL-20) dataset consists of gray scale images of 20 objects. Five objects out of 20 are used for the present study. Each object is represented by 72 gray scale images obtained by rotating the object with step of five degree. Each image is downsampled into 32X32 pixels and 16X16 pixels, and ten samples per class were randomly selected for training set with remaining samples forming the test set. The number of components in each mode was set to (R1,R2) ∈ {(5, 5), (10,10)} and no feature selection was performed on the extracted features.

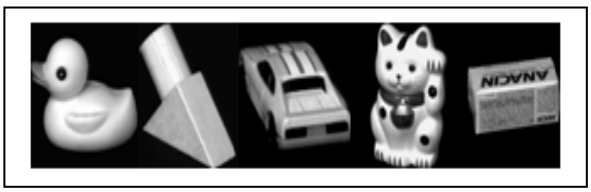

Fig. 2. Object images from COIL 20

Fig. 3. Face images from Sheffield face database

TABLE I .ACCURACY ESTIMATION OF OBJECT RECOGNITION USING KNN. IMAGE DIMENSION: 32X32, FEATURE DIMENSION: 10X10, 5X5

Object Rec	KNN	Object Rec	KNN
32x32	5X5	32x32	10x10
Alpha	Accuracy	Alpha	Accuracy
1	84.17	1	81.67
1.25	86.67	1.25	92.5
1.5	86.67	1.5	92.5
1.75	86.67	1.75	92.5
2	86.67	2	92.5
2.25	86.67	2.25	92.5
2.5	86.67	2.5	92.5
2.75	86.67	2.75	92.5
3	86.67	3	92.5

TABLE II. ACCURACY ESTIMATION OF OBJECT RECOGNITION USING KNN IMAGE DIMENSION: 16X16, FEATURE DIMENSION: 10X10, 5X5

Object Rec	KNN	Object Rec	KNN
16x16	5x5	16x16	10x10
Alpha	Accuracy	Alpha	Accuracy
1	84.17	1	91.67
1.25	90	1.25	89.17
1.5	90	1.5	89.17
1.75	90	1.75	89.17
2	90	2	89.17
2.25	90	2.25	89.17
2.5	90	2.5	89.17
2.75	90	2.75	89.17
3	90	3	89.17

The Sheffield Face database (SFD) consists of 575 images of 20 individuals with mixed race gender and appearance. Four individuals with mixed combinations are considered in the present study. Each individual shown in a range of poses from profile to frontal views with each image cropped to 112 X 92 pixels with 8 bit gray levels per pixels. Prior to feature extraction all images were down sampled to 28 X 23 pixels, and raw images were used as input for feature extraction. Training set was formed by randomly selecting six samples for each class with remaining images forming the test set. The number of components in each mode was set to (R1,R2) $\in$ {(5, 5), (10,10)} and no feature selection was performed on the extracted features.

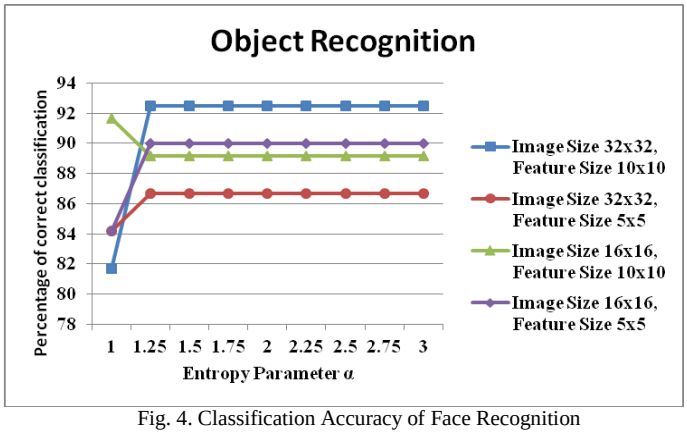

Fig. 4. Classification Accuracy of Face Recognition under different values of entropy parameter α.

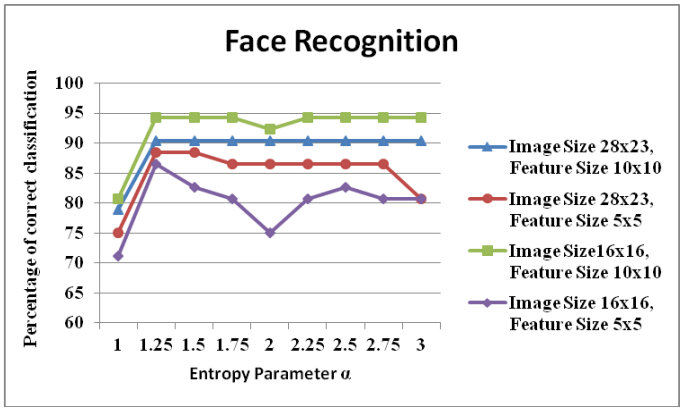

Fig. 5. Classification Accuracy of Object Recognition under different values of entropy parameter α.

TABLE III. ACCURACY ESTIMATION OF FACE RECOGNITION USING KNN. IMAGE DIMENSION: 16X16, FEATURE DIMENSION: 10X10, 5X5

Face Rec	KNN	Face Rec	KNN
16x16	10x10	16x16	5x5
Alpha	Accuracy	Alpha	Accuracy
1	80.77	1	71.16
1.25	94.23	1.25	86.55
1.5	94.23	1.5	82.69
1.75	94.23	1.75	80.76
2	92.31	2	75
2.25	94.23	2.25	80.76
2.5	94.23	2.5	82.69
2.75	94.23	2.75	80.77
3	94.23	3	80.77

TABLE IV. ACCURACY ESTIMATION OF FACE RECOGNITION USING KNN. IMAGE DIMENSION: 28X23, FEATURE DIMENSION: 10X10, 5X5

Face Rec	KNN		Face Rec	KNN
28x23	10x10		28x23	5x5
Alpha	Accuracy		Alpha	Accuracy
1	78.85		1	75
1.25	90.38		1.25	88.46
1.5	90.38		1.5	88.46
1.75	90.38		1.75	86.54
2	90.38		2	86.54
2.25	90.38		2.25	86.54
2.5	90.38		2.5	86.54
2.75	90.38		2.75	86.54
3	90.38		3	80.77

V. CONCLUSION

This work proposes a novel approach for supervised feature extraction for tensor objects by MMI criteria using Tsallis entropy. The projection matrices are obtained by maximizing an approximation of mutual information between the extracted features and class labels. More discriminative features can be obtained by using higher order statistics of the data rather than using only second order statistics. Several experiments show that the proposed approach can be used to significantly improve discriminative ability of the features extracted from tensor objects. Various linear and non-linear tensor based classifiers can be used to analyze the performance of the proposed method and an effective comparative study can be done in future.

REFERENCES

[1] Nie, F., Xiang, S., Song, Y., Zhang, C., 2009. Extracting the optimal dimensionality for local tensor discriminant analysis. Pattern Recogn. 42, 105–114.

[2] Wang, S. J., Chen, H. L., Yan, W. J., Chen, Y. H., & Fu, X. (2014). Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine.Neural processing letters, 39(1), 25-43.

[3] Lu, G., Halig, L., Wang, D., Chen, Z. G., & Fei, B. (2014, March).Spectral-spatial classification using tensor modeling for cancer detection of hyperspectral imaging. In SPIE Medical Imaging (pp. 903413-903413).International Society for Optics and Photonics.

[4] Yan, S., Xu, D., Yang, Q., Zhang, L., Tang, X. Z.-J., 2005. Discriminant Analysis with Tensor Representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA, pp. 526–532.

[5] Tao, D., Li, X., Wu, X., Maybank, S.J., 2007. General tensor discriminant analysis and Gabor features for gait recognition. IEEE Trans. Pattern Anal. Mach. Intell. 29 (10), 1700–1715.

[6] Zhang, W., Lin, Z., Xiaoou, T., 2009. Tensor linear Laplacian discrimination (TLLD) for feature extraction. Pattern Recogn. 42, 1941–1948.

[7] Phan, A.H., Cichocki, A., 2010. Tensor decompositions for feature extraction and classification of high dimensional datasets. IEICE Nonlinear Theory Appl. 1, 37–68.

[8] Ante Jukic, Marko Filipovic, 2013. Supervised feature extraction for tensor objects based on maximization of mutual information, Pattern Recognition Letters 34, 1476–1484.

[9] Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. Neural Computing and Applications, 24(1), 175-186.

[10] Marius Vila , Anton Bardera, Miquel Feixas and Mateu Sbert,2011. Tsallis Mutual Information for Document Classification.,Entropy, 13, 1694-1707.

[11] Ricardo Fabbri, Wesley N. Goncalves, Francisco J. P Lopes, Odemir M. Bruno, 2012.Multi-q Analysis of Image Patterns,PhysicaA.,p1-10.

[12] Sluga, D., & Lotric, U. (2013). Generalized Information-Theoretic Measures for Feature Selection.In Adaptive and Natural Computing Algorithms (pp. 189-197). Springer Berlin Heidelberg.

[13] Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. J.Stat.Phys.1988,52,479–487.

[14] Furuichi, S. Information theoretical properties of Tsallis entropies.J.Math.Phys.2006,47,023302.

[15] Hyvärinen A, Karhunen, J., Oja, E., 2001. Independent Component Analysis.Wiley,NewYork,USA.