Biometric identification using Hierarchical Temporal Memory with face recognition as a case study

Dipan Kumar Pal
Dept. of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, India
dipan2912@gmail.com

Abstract - **This paper presents the application of Hierarchical Temporal Memory (HTM) to the face recognition problem domain of biometric identification. We test the HTM on three face recognition datasets having 1440, 3040 and 378 images respectively and the results are exceptional giving 100 percent accuracy on all three datasets, although with slightly modified parameters. The accuracy was achieved while keeping the number of training images at just one image per face. The HTM algorithm can also be used for classification of other biometric features and traits. This provides the possibility of development of a novel system of reduced complexity employing a single algorithm, but which still offers very robust identification over multiple biometric domains.**

Keywords – **Hierarchical temporal memory, Biometric identification, face recognition, HTM**

I. INTRODUCTION

Biometric identification systems uniquely recognize an individual's physical traits and allow them to become their own "passwords". These systems in general employ a variety of algorithms for identifying or classifying various biometric features. Each biometric identification domain such as face recognition, speaker identification, and hand geometry identification employs algorithms that are specifically designed for and tuned to the task. This creates a huge diversity in the nature of the algorithms since each problem appears to have almost nothing in common. But most of such biometric domain classification problems can be said to have been solved by biological systems such as the human neocortex. Furthermore, recent work has shown that most problems in the neocortex are solved using the same algorithm [2]. And the "knowledge" gained from processing different types of data is stored using a common mechanism. The mathematical model of the mechanism was developed by Hawkins and George and is called the Hierarchical Temporal Memory (HTM) [2]. The previous insight gives rise to the idea of a common biometric classification algorithm for multiple biometric features. This would reduce the complexity and cost of a system performing biometric identification over multiple domains.

We now briefly overview some work on face recognition, speaker identification and hand geometry identification. In [7], Sim et al propose a simple memory based technique for face recognition and show it to outperform other techniques such as PCA analysis or Eigenfaces. There also has been work on use of Gabor filters along with a supervised classifier to perform face recognition [5]. But the classifier they employed performed clustering in a single stage. A multistage clustering mechanism like that used by HTMs is shown to achieve better results. We will also use the Gabor filter, but only as a default part of the Numenta Vision Framework which we use to implement the HTM [3]. In [9], Grimaldi and Cummins use the AM-FM representation of a speech signal for classification whereas Reynolds and Rose promote the use of a Gaussian Mixture Model for robust text-independent speaker identification [8]. Previous attempts to optimize algorithms such as in [10], where the authors optimize vector quantization based speaker identification also have been effective. Successful work based on hand geometry identification can be found in [11], where the authors perform the task without feature extraction and through a general regression neural network. Whereas in [12], the authors focus on combining palm print features along with hand geometry features to achieve better recognition rates. As can be seen, the algorithms of different problem domains have huge diversity amongst themselves with considerable diversity within a problem domain itself on some occasions. This creates complexity in implementing a multi domain biometric identification system, but this problem can be solved with the use of a HTM.

In this paper, we test the classification performance of a HTM when applied to three face recognition datasets. The datasets are the faces95 [16], faces 96 [17] and grimace [18] and are all publicly available. The first two datasets have significant lighting and head scale variation, whereas the third one has large variation in expression. The datasets and the experiments conducted are explained in more detail in a later section. The HTM was used as a part of the Numenta Vision Framework of the NuPIC 1.7 release [3]. Also as part of the Numenta Vision Framework, a Gabor filter was used to receive raw visual input and to convert it into a sparse representation of the input, which is important because as we shall see, the HTM relies on sparse representations for efficient modeling of the data. Also, even though we do not test the HTM over other different types of data such as voice, iris and retina scans and hand geometries, we explain how their inherent hierarchical structure promises excellent results when classified using a HTM. As mentioned earlier, this could lead to systems performing robust biometric identification over multiple modalities or domains despite having reduced complexity as compared to combinations of traditional algorithms.

II. HIERARCHICAL TEMPORAL MEMORY IN BIOMETRIC IDENTIFICATION

A. *Description of the HTM*

The Hierarchical Temporal Memory (HTM) is an algorithm which tries to capture the mechanism of data modeling and processing capabilities of the human neocortex. HTM is similar to Bayesian networks which use belief propagation, but they are easier to handle. The algorithm essentially uses clustering mechanisms to achieve invariance in output when an input belonging to a particular class is presented to the network. It does this by forming a spatial temporal correlation between low level input patterns which appear to the network. Thus knowledge and understanding about the HTM environment is only gained with what the HTM perceives as input.

HTMs in general are a tree structured multi-leveled hierarchy with each level consisting of a region of nodes. A typical 3 level HTM is shown in Fig. 1. An HTM can consist of any number of levels, but for most applications a 2 or 3 level node network suffices. Each level consists of a fixed number of nodes all of which perform the same algorithm. The bottom most level of the HTM is fed with the raw input data, which in this case is the output of a Gabor filter fed with a RGB color image. Each node performs clustering in overall three dimensions (two of space and one of time) and it does this in two stages. The first stage is called the spatial pooler and the second one is the temporal pooler.

As the name suggests, the spatial pooler pools or clusters data in the spatial dimension. Each pattern appearing at the input during learning of the spatial pooler is compared with the database of other patterns, if the distance between the input pattern and each is less than the maxDistance parameter, then the input pattern is considered same as the corresponding existing pattern, termed as a coincidence. If the previous condition does not satisfy, then the input pattern is “memorized” as a new coincidence. Thus the spatial pooler quantizes the input space and only remembers the patterns which appear. This helps it to capture coincidences from huge input spaces efficiently. The temporal pooler performs clustering over time and forms temporal groups of the coincidence patterns. These groups are formed on the basis of the statistical behavior of the input data, which is modeled using a Markov graph whose nodes are the coincidence patterns learned previously. Hence, the members of a temporal group are likely to follow one another. After training, a vector of probabilities of membership of the input pattern to each of the temporal groups is the input to the next level of nodes. Therefore, the overall effect of this approach causes the lower level nodes to remember and recognize patterns of lower complexities such as a line or corner. As we ascend the hierarchy, we find that the coincidences represent combinations of patterns of lower complexities. This increases the variance and complexity of data represented at higher levels. But in spite of the seemingly large input space at higher levels, the spatial pooler at higher levels only remembers patterns it encounters thereby improving efficiency. Furthermore, a sparse representation of the input reduces the input space and helps in efficient handling of data and this is one of the key features of the general theory of the HTM.

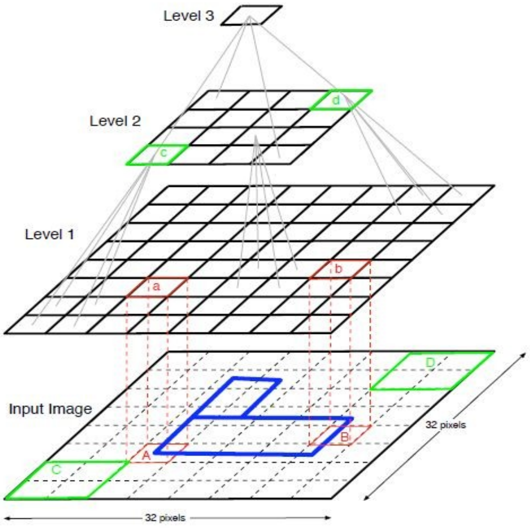

Fig. 1. General structure of a three level HTM

HTM levels can be individually run in two modes, the learning mode and the inference mode. In the learning mode, a level tries to find new coincidences and keeps updating the Markov graph as time progresses. In inference mode, the probability distribution of the membership of the input pattern is outputted to the next higher level. The learning mode provides no such output. During training of a particular level, all levels below it are run in inference mode and it itself is run in the learning mode.

For intelligent recognition, a system must be able to identify a known pattern from within novel inputs. We have seen how the HTM works to manage huge input spaces and has the ability to cluster various complex patterns under a single category. This property of the HTM allows it to achieve very good classification accuracy, since various images of the same object, with different orientations, noise, brightness etc, all get classified as same object. We must understand that this important characteristic of invariance of class at higher levels given a changing noisy pattern of the same class at lower level is achieved by clustering or quantizing and mapping the input space which results in robustness. The sparse representation property of the Gabor filter, we had mentioned earlier, now can be appreciated since it too acts as a kind of spatial pooler which clusters a complex visual scene into a simpler sparse one.

Also, one might argue as to the existence of comprehensible temporal sequences in flashing a sequence of unrelated images, but as we will see later during experiments, recognition still is possible due to the spatial pooler although its effectiveness is reduced. Using more extensive training algorithms such as ExhaustiveSweep and MultiSweep provided with the Numenta Vision Framework we can utilize the temporal poolers to boost performance.

B. Relevance of the HTM in biometric identification

HTM, as we saw, models the spatial and temporal correlations amongst the hierarchical components of data and uses it to perform inference over novel inputs. The data we refer to here is just general data. There have been no assumptions about it whatsoever in the development of the algorithm except as Hawkins explains, that the data must have been produced by a hierarchical system as well [2]. Fortunately, this is true of most world data. In our face recognition application, each image of a face consists of simpler structures which are the facial features. And each feature consists of even more simpler structures which have more subtle differences between them. The HTM captures those differences in its lower levels, whereas the higher levels capture differences on a larger scale such as the position of the facial features etc. Thus for our current application the lower levels store most of the 'knowledge' of the subject's faces.

But apart from faces, other biometric identification domains such as voice, hand geometry and handwriting recognition, retina and iris identification also have a hierarchical structure of its data. For example, a typical voice signal has many frequency components of varying magnitude. And the retina, iris and hand geometry recognition problems can be thought of as similar in nature to the face recognition problem in context of HTMs. In fact, there has been previous work on spoken digit recognition [14] and handwritten digit recognition [15] using HTMs, which might be seen as preliminary works towards complete speech and handwriting to text conversion. But these applications focus on what is being spoken and written rather than who is speaking or writing it. But Numenta provides a speech processing toolkit along with its NuPIC 1.7 (Numenta Platform for Intelligent Computing) release [13]. This toolkit has as a demo, a speaker identification problem solved with very good results. These show that we might change the questions we ask the HTM, for during training it will model the semantics of the dataset according to our supervision. Thus by changing the organization of the training dataset, we can change the semantics of the data the HTM tries to model.

Another point of interest as mentioned earlier is the possibility of integrating many different HTMs together into a single system for simultaneous identification of various biometric features. A more optimized system could have a single HTM which is provided with pre-processed biometric information about different physical traits, and whose outputs are concatenated together to form a complete biometric profile of the person. This concatenated profile can be again used to train the HTM which will then provide absolute confirmation of identity during inference. But other techniques could as well be used for the final identification. Nonetheless, HTMs as we shall see in a later section, provide robust classification of facial data and this might suffice as enough identity proof for many applications. Now, having understood the hierarchical nature of general world data and the mechanism that the HTM uses to model it, the claim of a universal biometric system employing a single algorithm can be made. A hardware optimized implementation of the HTM would result in very fast, robust and thorough identification systems.

III. EXPERIMENTAL DETAILS AND RESULTS

The experiments presented here were conducted using the Numenta Vision Framework. The three face recognition datasets tested on are explained in this section along with their respective results. Also mentioned are the HTM general parameter changes made in the Vision Framework. These are shown in Table I. We first explain the structure of the HTM used.

TABLE I
MAJOR PARAMETER CHANGES FOR THE HTM USING NUMENTA VISION FRAMEWORK

HTM parameters	Values
numCategories	10
seed	24
midLevelPatches	160
gaborNumOrients	(is varied)
gaborPhaseMode	'single'
gaborCenterSurroud	False
spatialPoolerAlgorithm	kthroot_product
maxDistance	0.3
temporalPoolerAlgorithm	maxProp
spatialPoolerTraining Algorithm	RandomFlash
temporalPoolerTrainingAlgorithm	MultiSweep

A. Structure of the HTM used

The HTM consisted of 5 levels. The first level consists of a Gabor filter with a receptive field which receives the input image from the image sensor provided with the Numenta Vision Framework. The Gabor filter can be said to work as a spatial pooler since it too clusters the input space and reduces it. It is also important because it provides a representation of the input images in a form that is more compatible with the HTM. The Gabor filter output is sent to a temporal pooler region. The spatial pooler and temporal pooler are considered to be different levels in the convention of the Vision Framework. The output of the temporal pooler is sent again to a spatial pooler region followed by a temporal pooler region up the hierarchy. This completes 4 levels of the hierarchy. But effectively the hierarchy forms only a 2 level HTM network. The top level is the classifier node which outputs a probability distribution of membership of the image in each of the categories.

B. Face recognition datasets

The datasets used were the faces95, faces 96 and the grimace [16], [17], [18]. The faces95 dataset had samples of 72 faces with 20 images per face sized 180x200. The faces96 dataset had 152 faces with 20 images per face with size 196x196, whereas the grimace dataset had samples of 18 faces with 21 images per face sized 180x200. Table II shows the results when the number of training images per face was kept at 1 and the accuracy was maintained at 100 percent by varying a parameter of the number of orientations of the Gabor filter. Each orientation of the Gabor filter gives a convolved output which has components from the image in only that orientation. Thus more the number of orientations, more the amount of information extracted from the image. The faces95 and faces96 datasets were the most tested due to their large size which would normalize the statistical effects a dataset of smaller size would have on the results. Faces95 was tested varying the number of orientations of the Gabor filter while keeping the number of training images per face at minimum i.e. 1. Table III depicts these results. Faces96 underwent a slightly different test where the number of training images per face was varied while keeping the number of orientations of the Gabor filter at minimum i.e. 2. The results of this test are shown in Table IV.

The faces95 and faces96 datasets had large scale variations, minor orientation variations and major lighting variations. The grimace dataset although had less number of samples, is technically the most difficult to classify due its large variation in expression, which is generally a challenge for any specialized face recognition algorithm. Some of the correctly classified faces along with variations within each subject's samples from the three datasets are shown in Fig. 2, Fig. 3 and Fig. 4. We see that the HTM performs very robust classification even with minimal number of training images per face. This is vital for scalability.

TABLE II
TOP RESULTS ON THE THREE DATASETS KEEPING THE NUMBER OF TRAINING IMAGES =1

Dataset	Number of Gabor orientations	Accuracy (%)	Level 2 coincidences (temporal)
Faces95	10	100	1680
Faces96	2	100	336
Grimace	2	100	336

TABLE III
RESULTS ON FACES95 KEEPING NUMBER OF TRAINING IMAGES =1

Number of Gabor orientations	Accuracy (%)	Level 2 coincidences (temporal)
2	99.4	336
5	99.9	840
10	100	1680
15	100	2520

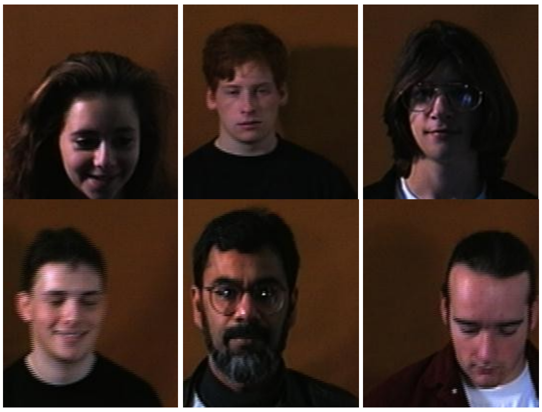

Fig. 2. Correctly classified samples from the faces95 dataset

V. CONCLUSION AND FUTURE WORK

We saw that a HTM implemented face recognition system can offer very high accuracies with minimal training samples. The database of faces can also be easily increased by training the system on additional samples. Thus the practical implementation of a biometric system employing HTMs would have reduced complexity, ease of training and use, and large scalability. Also, a biometric identification system with multi domain identification is possible within a single algorithmic framework, although separate pre-processing of biometric data would be required for compatibility with the HTM. This requirement can be illustrated by the use of the Gabor filter in the Numenta Vision Framework.

TABLE IV
RESULTS ON FACES96 KEEPING NUMBER OF GABOR ORIENTATIONS =2

Number of training images per face	Accuracy (%)	Level 2 coincidences (temporal)
5	99.9	336
3	99.9	336
1	100	336

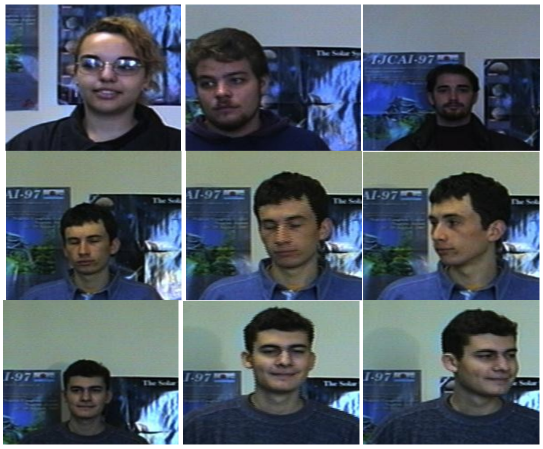

Fig. 3. Correctly classified samples from the faces96 dataset also showing variation of scale and orientation for each face

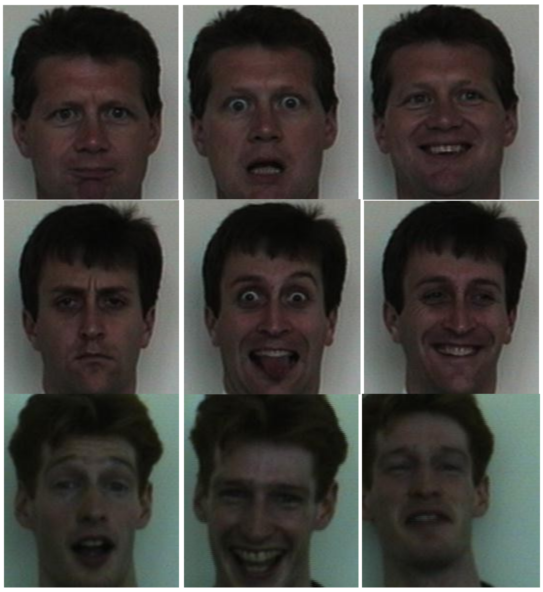

Fig. 4. Correctly classified samples from the grimace dataset showing large variation in expressions for each face

We intend to conduct further research on systems offering multi domain biometric identification. There might be two approaches to the problem. One is the method earlier described of using a single HTM for sequential recognition. The other is to use a larger HTM with multiple pre-processed biometric data streams being fed into different parts of its receptive field. The later is expected to provide better results as the post processing of the classified biometric data would also be effectively done by a HTM. In the first approach this post processing might be carried out by simpler, computationally less intensive techniques.

Numenta is also working on an improved version of the algorithm employing sub-cortical mechanisms. Its performance and behavior would be interesting to research upon with context to biometric identification.

REFERENCES

[1] D. George and B. Jaros, “The HTM Learning Algorithms”, Numenta Incorprated, March 1, 2007.

[2] J. Hawkins and D. George, “Hierarchical Temporal Memory: Concepts, Theory, and Terminology”, Numenta Incorprated , March 27, 2007.

[3] “VisionFrameworkGuide”, Numenta Incorprated, August 2009.

[4] “nupic_gettingstarted”, Numenta Incorprated, September 2008.

[5] T. Barbu."Gabor Filter-based Face Recognition Technique", Proc. of the Romanian Academy, Series A, Volume 11, Number 3/2010, pp. 277–283.

[6] J. R. Movellan, “Tutorial on Gabor filters”.

[7] T. Sim, R. Sukthankar, M. Mullin, and S. Baluja, “High-Performance Memory-Based Face Recognition for Visitor Identification,”, Tech. Rep. JPRC-TR-1999-001-1, 1999.

[8] D.A. Reynolds and R.C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models”. IEEE Trans. Speech Audio Process., 3 (1995), pp. 72–83.

[9] M. Grimaldi and F. Cummins, “Speaker Identification using Instantaneous Frequencies”, IEEE Transactions on Audio, Speech, and Language Processing, vol., 16, no. 6, August 2008.

[10] T Kinnunen, E. Karpov and P. Franti. “Real-time speaker identification and verification”. IEEE Transactions on Audio, Speech and Language Processing, 14(1):277-288, 2006.

[11] Ö. Polat, T. Yıldırım, “Hand geometry identification without feature extraction by general regression neural network”, Expert Systems with Applications, Volume 34, Issue 2, February 2008, Pages 845-849, ISSN 0957-4174, 10.1016/j.eswa.2006.10.032.

[12] A. Kumar, D. C. M. Wong, H. C. Shen, and A. K. Jain, “Personal Verification using Palmprint and Hand Geometry Biometric”, LNCS 2688, pp. 668-678 (2003).

[13] "Speech Processing with Hierarchical Temporal Memory", Numenta Incorporated, June 2008.

[14] J.V. Doremalen and L. Boves (2008): "Spoken digit recognition using a hierarchical temporal memory", In INTERSPEECH-2008, 2566-2569.

[15] B. Bobier, Handwritten Digit Recognition using Hierarchical Temporal Memory. Science 1-9 (2007).

[16] “faces95” face recognition dataset.

[17] “faces96” face recognition dataset.

[18] “grimace” face recognition dataset.