Hetero-Associative Memories for Voice Signal and Image Processing

Roberto A. Vázquez and Humberto Sossa

Centro de Investigación en Computación – IPN
Av. Juan de Dios Batíz, esquina con Miguel Othón de Mendizábal
Ciudad de México, 07738, México
ravem@ipn.mx, hsossa@cic.ipn.mx

Abstract. An associative memory AM is a type of neural network commonly used for recalling output patterns from input patterns that might be altered by noise. Most of these models have several constraints that limit their applicability in complex problems. Recently, in [13] a new AM based on some aspects of human brain was introduced, however the authors only test its accuracy using image patterns. In this paper we show that this model is also robust with other type of patterns such as voice signal patterns. The AM is trained with associations composed by voice signals and their corresponding images. Once trained, when a voice signal is used to stimulate the AM we expect the memory recall the image associated to the voice signal. In order to test the accuracy of the proposal, a benchmark of sounds and images was used.

Keywords: Associative memories, voice signal processing, image processing.

1 Introduction

The concept of associative memory AM emerges from psychological theories of human and animals learning [26]. These memories store information by learning correlations among different stimuli. When a stimulus is presented as a memory cue, the other is retrieved as a consequence; this means that the two stimuli have become associated each other in the memory.

An AM can be seen as a particular type of neural network designed to recall output patterns in terms of input patterns that can appear altered by some kind of noise. Several AMs have been proposed in the last 50 years, see for example [1], [2], [3], [4], [5], [6], [7], [8] and [9]. Most of these AMs have several constraints that limit their applicability in complex problems. Among these constraints we could mention their capacity of storage (limited), the type of patterns (only binary, bipolar, integer or real patterns), robustness to noise (additive, subtractive, mixed, Gaussian noise, deformations, etc). Recently, in [13] a new AM based on some aspects of human brain was introduced. Although authors show the robustness of the model applied to face and 3D object recognition [11] and [12], even if patterns are contaminated by different type of noises and transformations, they do not report results using other type of stimulus patterns such as voice signal patterns.

The storage of voice signal or other type patterns have sense because human memory is not only stores patterns acquired from the vision system such as objects or

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 659–666, 2008.

faces, but also stores patterns acquired from the auditory or olfactory system. From a sensory system point of view a hetero-associative memory is not only a memory that associates different patterns, but associates patterns obtained from different sensory systems.

In this paper we show that the AM described in [13] is capable of associating stimulus patterns acquired from different sensory systems. In order to train the AM, we will use associations composed by a voice signal and their corresponding images. Once trained, when a voice signal is used to stimulate the AM we expect that the memory recall the image associated to the voice signal. In order to test the accuracy of the proposal, a benchmark of sounds and images is used; each association was composed by a voice signal of the name of a flower or animal and an image that contains the flower or animal.

2 The Associative Model

Let $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} \in \mathbf{R}^m$ an input and output pattern, respectively. An association between input pattern $\mathbf{x}$ and output pattern $\mathbf{y}$ is denoted as $\left(\mathbf{x}^k, \mathbf{y}^k\right)$, where k is the corresponding association. Associative memory $\mathbf{W}$ is represented by a matrix whose components w_{ij} can be seen as the synapses of the neural network. If $\mathbf{x}^k = \mathbf{y}^k \forall k = 1,\ldots, p$ then $\mathbf{W}$ is auto-associative, otherwise it is hetero-associative. A distorted version of a pattern $\mathbf{x}$ to be recalled will be denoted as $\tilde{\mathbf{x}}$. If an AM $\mathbf{W}$ is fed with a distorted version of $\mathbf{x}^k$ and the output obtained is exactly $\mathbf{y}^k$, we say that recalling is robust.

The dynamic associative model DAM described in [13], it is not an iterative model as Hopfield's model [4]. The principal difference of this model against other classic models is that during recalling phase the synapses' values could change as a respond to an input stimulus. The formal set of propositions that support the correct functioning of this model and the advantages against other classical models can be found in [13]. This model defines several interacting areas, one per association we would like the memory to learn. Also integrate the capability to adjust synapses in change as a response to an input stimulus. Before an input pattern is learned by the brain, it is hypothesized that it is transformed and codified by the brain. This process is simulated using the procedure introduced in [7]. This procedure allows computing *codified patterns* from input and output patterns denoted by $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ respectively; $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are *de-codifying patterns*. Codified and de-codifying patterns are allocated in different interacting areas and *d* defines how much these areas are separated. On the other hand, *d* determines the noise supported by the model. In addition a simplified version of $\mathbf{x}^k$ denoted by s_k is obtained as:

$$s_k = s\left(\mathbf{x}^k\right) = \mathbf{mid}\ \mathbf{x}^k \tag{1}$$

where **mid** operator is defined as $\mathbf{mid}\ \mathbf{x} = x_{(n+1)/2}$.

In this model, the most excited interacting area is call *active region* (AR) and could be estimated as follows:

$$ar = r(\mathbf{x}) = \arg\left(\min_{i=1}^{p} \left| s(\mathbf{x}) - s_i \right| \right) \tag{2}$$

Once computed the *codified patterns*, the *de-codifying patterns* and s_k we can compute the synapses of the DAM as follows:

Let $\left\{ \left(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k \right) \middle| k = 1, \ldots, p \right\}, \overline{\mathbf{x}}^k \in \mathbf{R}^n, \overline{\mathbf{y}}^k \in \mathbf{R}^m$ a fundamental set of associations (codified patterns). Synapses of DAM $\mathbf{W}$ are defined as:

$$w_{ij} = \overline{y}_i - \overline{x}_j \tag{3}$$

In short, building of the DAM can be performed in three stages as: First, transform the fundamental set of association into codified and de-codifying patterns by means of Procedure described in [7]. Then, compute simplified versions of input patterns by using equation 1. And finally, build $\mathbf{W}$ in terms of codified patterns by using equation 3.

Synapses could change in response to an input stimulus. There are synapses that can be drastically modified and they do not alter the behavior of the DAM. In the contrary, there are synapses that only can be slightly modified to do not alter the behavior of the DAM; this set of synapses is call *the kernel* of the DAM and it is denoted by $\mathbf{K}_\mathbf{W}$. Let $\mathbf{K}_\mathbf{W} \in \mathbf{R}^n$ the kernel of a DAM $\mathbf{W}$. A component of vector $\mathbf{K}_\mathbf{W}$ is defined as:

$$kw_i = \mathbf{mid}\left(w_{ij} \right), j = 1, \ldots, m \tag{4}$$

Synapses belonging to $\mathbf{K}_\mathbf{W}$ are modified according to the stimulus generated by the input pattern. This adjusting factor is denoted by Δw and can be computed as:

$$\Delta w = \Delta(\mathbf{x}) = s\left(\overline{\mathbf{x}}^{ar} \right) - s\left(\tilde{\mathbf{x}} \right) \tag{5}$$

where ar is the index of the AR and $\tilde{\mathbf{x}} = \mathbf{x} + \hat{\mathbf{x}}^{ar}$.

Finally, synapses belonging to $\mathbf{K}_\mathbf{W}$ are modified as:

$$\mathbf{K}_\mathbf{W} = \mathbf{K}_\mathbf{W} \oplus \left(\Delta w - \Delta w_{old} \right) \tag{6}$$

where operator $\oplus$ is defined as $\mathbf{x} \oplus e = x_i + e \ \forall i = 1, \ldots, m$. As you can appreciate, modification of $\mathbf{K}_\mathbf{W}$ in eq. 6 depends of the previous value of Δw denoted by Δw_{old} obtained with the previous input pattern. Once trained the DAM, when it is used by first time, the value of Δw_{old} is set to zero.

Once synapses of the DAM have been modified in response to an input pattern, every component of vector $\overline{\mathbf{y}}$ can be recalled by using its corresponding input vector $\overline{\mathbf{x}}$ as:

$$\overline{y}_i = \mathbf{mid}\left(w_{ij} + \overline{x}_j \right), j = 1, \ldots, n \tag{7}$$

In short, pattern $\overline{\mathbf{y}}$ can be recalled by using its corresponding key vector $\overline{\mathbf{x}}$ or $\tilde{\mathbf{x}}$ in six stages as follows: First, obtain index of the active region ar by using equation 2.

Then, transform $\mathbf{x}^k$ using de-codifying pattern $\hat{\mathbf{x}}^{ar}$ by applying the following transformation: $\tilde{\mathbf{x}}^k = \mathbf{x}^k + \hat{\mathbf{x}}^{ar}$. After that, compute adjust factor $\Delta w = \Delta(\tilde{\mathbf{x}})$ by using equation 5. Next modify synapses of associative memory $\mathbf{W}$ that belong to $\mathbf{K}_{\mathbf{W}}$ by using equation 6. Then, recall pattern $\tilde{\mathbf{y}}^k$ by using equation 7. Finally, obtain $\mathbf{y}^k$ by transforming $\tilde{\mathbf{y}}^k$ using de-codifying pattern $\hat{\mathbf{y}}^{ar}$ by applying transformation: $\mathbf{y}^k = \tilde{\mathbf{y}}^k - \hat{\mathbf{y}}^{ar}$.

As was shown in [11] and [12], the original DAM performs a low accuracy in complex problems such as face or 3d object recognition. In order to increase the accuracy of the DAM the authors suggest computing a simplified version of the DAM model by using a random selection of stimulating points. Some pixels (stimulating points) of pattern $\mathbf{x}^k$ are random selected, where *k* defines the class of the pattern. These stimulating points **SP** are used by the DAM to determine an active region and are given by $\mathbf{sp} \in \left\{\mathbb{Z}^+\right\}^c$ where c is the number of used **SP**. $sp_i = random(n), i = 1,\ldots,c$ where n is the size of the pattern.

To determine the active region, the DAM stores during training phase an alternative simplified version of each pattern $\mathbf{x}^k$ given by:

$$\mathbf{ss}^k = ss\left(\mathbf{x}^k\right) = \mathbf{x}^k\Big|_{\mathbf{sp}} = \left\{x^k_{sp_1},\ldots,x^k_{sp_c}\right\} \tag{8}$$

During recalling phase, each element of an input simplified pattern $\tilde{\mathbf{x}}^k\Big|_{\mathbf{sp}}$ excites some of these regions and the most excited region will be the active region. To determine which region is excited by an input pattern we use:

$$b = \arg\min_{k=1}^{p}\left|\left[ss\left(\mathbf{x}\right)\right]_i - \mathbf{ss}_i^k\right| \tag{9}$$

For each element of $\tilde{\mathbf{x}}^k\Big|_{\mathbf{sp}}$ we apply equation 9 and the most excited region (the region that more times was obtained) will be the active region.

Building of the DAM is done as follows: Let $\mathbf{S}_x^k$ and $\mathbf{S}_y^k$ an association between a voice signal with an image, and c be the number of stimulating points. First, take at random c stimulating point sp_i. Then, for each association transform the voice signal and image into a raw vector ($\mathbf{x}^k$, $\mathbf{y}^k$) and finally, train the DAM.

Image $\mathbf{S}_y^k$ can be recalled by using its corresponding voice signal $\mathbf{S}_x^k$ or distorted version $\tilde{\mathbf{S}}_x^k$ as follows: first, use the same c stimulating point sp_i. Then, transform the voice signal into a raw vector and finally, operate the DAM.

3 Accuracy of the Hetero-Associative Model

To test the accuracy of the DAM we performed several experiments. The first 5 experiments tested the accuracy of the model using sets of voice signal altered with additive, subtractive and mixed noise. The last experiments tested the model with a set of slightly distorted voice signals recorded at different tempo, volume, velocity and

tone (samples recorded under an uncontrolled environment). For whole experiments, each voice signal was associated with the image which best describe the voice signal, see Fig. 1.

Each voice signal was recorded into a wav file (PCM format, 44.1 KHz, 16 bits and mono). Before training the DAM, each voice signal has to be transformed into a voice signal pattern. To build a voice signal pattern from the wav file, we only read the wav information chunk of the file and then we stored it in an array (no preprocessing technique was applied to the wav information; we only used the raw information). On the other hand, each image was acquired and saved in a BMP file (True color, 24 bits). Before training the DAM, each image has to be transformed into an image pattern. To build an image pattern from the bmp file, the image was read from left -right and up-down; each RGB pixel (hexadecimal value) was transformed into a decimal value and finally, we stored the information into an array. Once trained and stimulated the DAM with a voice signal stimulus, each value of the recall pattern is transformed into a RGB value to restore the image.

Experiment 1: Recalling the fundamental set of associations: In this experiment we firstly trained the DAM with the set of voice signals and images. In all cases, each voice signal was associated with the image which best described the recorded voice signal (40 associations). Once trained the DAM, as described in section 2, we proceeded to test the accuracy of the proposal. First we verified if the DAM was able to recall the fundamental set of associations using set of voice signals. In this experiment the DAM provided a 100% of accuracy. Whole associations used to train the DAM were perfectly recalled.

Experiment 2: In this experiment, we verified if the DAM was able to recall the image associated to the voice signal used as input pattern, even if the voice signal is altered by additive noise (AN). To do this, each voice signal previously recorded was contaminated with AN altering from 2% until 90% of the information. This new set of voice signals was composed of 3560 samples. The accuracy of the proposal using this set of voice signals and 1 SP was of 42.5%, when the number of SP was increased the accuracy of the DAM increased. By using 100 SPs, the accuracy of the proposal increased to 100%.

Experiment 3: In this experiment, we verified if the DAM was able to recall the image associated to the voice signal used as input pattern, even if the voice signal is altered by subtractive noise (SN) as equal as in experiment 2. This new set of voice signals was composed of 3560 samples. The accuracy of the proposal using this set of voice signals and only 1 SP was of 44.4%, a little bit better than in experiment 2. By using 100 SPs, the accuracy of the proposal was also 100%.

Experiment 4: In this case, we verified if the DAM was able to recall the corresponding image even if the voice signal is altered by mixed noise (MN). Altering signals with MN, we built a set of voice signals composed of 3560 samples. The accuracy of the proposal using this set diminished to 36.6%. By using 100 SPs, the accuracy of the proposal was 99.9% and by using 110 SPs the accuracy was increased to 100%.

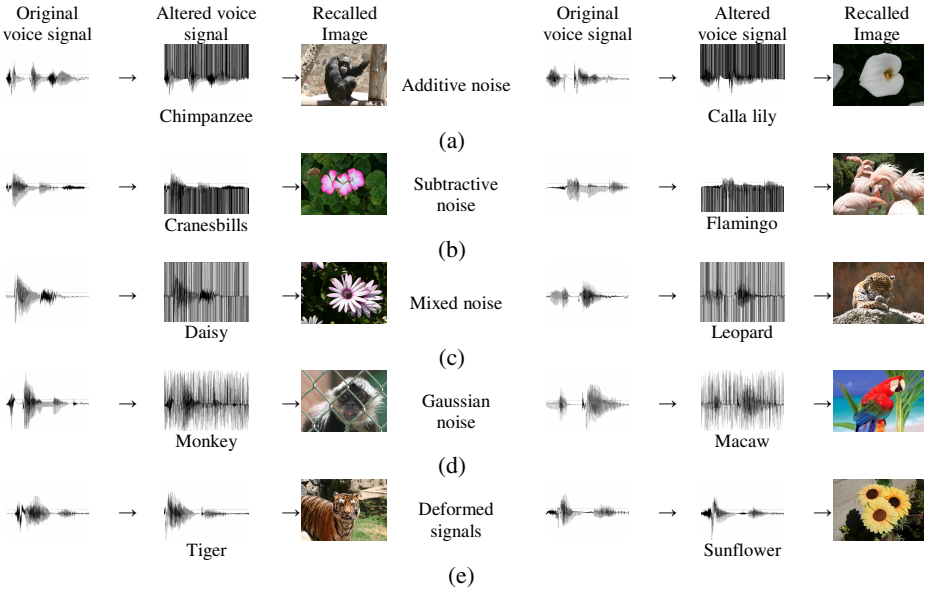

Fig. 1. Some images recalled using voice signals altered by different type of noises

Experiment 5: In this experiment, we verified if the DAM was able to recall the image associated to the voice signal used as input pattern, even if the voice signal is altered by Gaussian noise (GN). To do this, each voice signal previously recorded was contaminated with Gaussian noise again by altering from 2% until 90% of the information. This new set of voice signals was composed of 3560 samples. The accuracy of the proposal using this set of voice signals and only 1 SP was of 46.5%, greater that all previous experiments. By using 100 SPs, the accuracy of the proposal was also 100%.

Experiment 6: In this case, we verified if the DAM was able to recall the image associated to the voice signal used as input pattern, even if the voice signal suffer slightly deformations such as voice signals recorded at different tempo, volume, velocity and tone. To do this, each voice signal previously recorded was again recorded 10 times under an uncontrolled environment. This new set of voice signals was composed of 400 samples. In contrast to previous experiments, the accuracy of the proposal using this set of voice signals was of 10.5%. Furthermore, the accuracy of the proposal slightly increased (19%) when the number of SP was increased.

3.1 General Discussion

The accuracy increases when the number of stimulating points increases. After using a number of stimulating points greater than 100, the obtained accuracy was of 100%. In addition, as you can appreciate from Fig. 2, that no matter the type of noise or amount of noises added to the patterns (additive, subtractive, mixed and Gaussian noise), the behavior of the proposal for each typo of noise was almost the same.

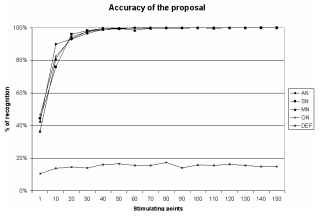

Fig. 2. Accuracy of the proposal using different number of stimulating points

For the first 5 experiments, we realized that a human was unable to perceive the voice signal if the voice signal was contaminated with noise in more than the 20%; therefore unable to recall the associated image. Using this DAM, we obtained a 100% of accuracy even if the voice signal was contaminated with noise until 90%. For the last experiments, we realized that a human was able to perceive the voice signal even if the voice signal was reproduced in different tempo and tone. Using this DAM, we obtained a 19% of accuracy.

It is worthy mentioning, that to our knowledge nobody in this field had before reported results of this type. Authors only report results when image patterns are distorted by additive, subtractive or both noises or when images under image orientations but not when the associative memory is trained with other type of patterns such as voice signals.

The results reported in this paper support the robustness of the associative model and show the versatility of the model in different environments. Furthermore, this hetero-associative model is capable to associate patterns from different domains, suggesting its applicability in a large domain of complex problems such as image retrieval using voice signal queries, translators using associative memories, control of robots using voice commands and associative memories, speech recognition using associative memories, etc.

No comparison with other AM was performed because constrains of these models limits their applicability in this problem. We must remark that capacity storage of this model is higher compared with classical models. For this experiment the DAM was trained with 40 associations while in classical models the authors used no more than 10 associations.

4 Conclusions

In this paper we have described a new concept of hetero-associative memory. From a sensory system point of view a hetero-associative memory is not only a memory that associates different patterns but it is a memory that associates patterns obtained from different sensory systems, for example, recalling of an image using as a key a voice signal.

We have shown the robustness of the dynamic associative model is a hetero-associative memory. The results obtained using a benchmark composed by 14440 voice signal samples, through the different experiments, support the applicability of this model in different complex problems that not involve only computer vision but also voice processing.

It is worth mentioning that, even without applying preprocessing methods for adequate the voice signals, the accuracy was highly acceptable, 100% for most of the experiments; in the last experiment the model presented an accuracy of 19%.

Nowadays, we are working in different applications that combine computer vision and speech recognition. We are integrating voice feature extraction techniques to increase the accuracy of the proposal when the voice signals are pronounced by different people.

Acknowledgment. This work was economically supported by SIP-IPN under grant 20082948 and CONACYT under grant 46805.

References

[1] Steinbuch, K.: Die Lernmatrix. Kybernetik 1, 26–45 (1961)
[2] Anderson, J.A.: A simple neural network generating an interactive memory. Math. Biosci. 14, 197–220 (1972)
[3] Kohonen, T.: Correlation matrix memories. IEEE Trans. on Comp. 21, 353–359 (1972)
[4] Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. 79, 2554–2558 (1982)
[5] Sussner, P.: Generalizing operations of binary auto-associative morphological memories using fuzzy set theory. J. Math. Imaging Vis. 19, 81–93 (2003)
[6] Ritter, G.X., et al.: Reconstruction of patterns from noisy inputs using morphological associative memories. J. Math. Imaging Vis. 19, 95–111 (2003)
[7] Sossa, H., Barron, R., Vazquez, R.A.: Transforming Fundamental set of Patterns to a Canonical Form to Improve Pattern Recall. In: Lemaître, C., Reyes, C.A., González, J.A. (eds.) IBERAMIA 2004. LNCS (LNAI), vol. 3315, pp. 687–696. Springer, Heidelberg (2004)
[8] Ritter, G.X., Sussner, P., Diaz de Leon, J.L.: Morphological associative memories. IEEE Trans Neural Networks 9, 281–293 (1998)
[9] Sussner, P., Valle, M.: Gray-Scale Morphological Associative Memories. IEEE Trans. on Neural Netw. 17, 559–570 (2006)
[10] James, W.: Principles of Psychology. Holt, New York (1890)
[11] Vazquez, R.A., Sossa, H., Garro, B.A.: 3D Object recognition based on low frequencies response and random feature selections. In: Gelbukh, A., Kuri Morales, Á.F. (eds.) MICAI 2007. LNCS (LNAI), vol. 4827, pp. 694–704. Springer, Heidelberg (2007)
[12] Vazquez, R.A., Sossa, H., Garro, B.A.: Low frequency responses and random feature selection applied to face recognition. In: Kamel, M., Campilho, A. (eds.) ICIAR 2007. LNCS, vol. 4633, pp. 818–830. Springer, Heidelberg (2007)
[13] Vazquez, R.A., Sossa, H.: A new associative memory with dynamical synapses. Neural Processing Letters (submitted, 2007)