A New Intelligent Systems Approach to 3D Animation in Television

Charles B. do Prado
Research and Development Department
Globo TV Network
Rio de Janeiro, Brazil
charles.prado@tvglobo.com.br

Felipe M. G. Franca
COPPE
Federal University at Rio de Janeiro
Rio de Janeiro, Brazil
felipe@cos.ufrj.br

Eduardo Costa
Research and Development Department
Globo TV Network
Rio de Janeiro, Brazil
eduardo.costa@tvglobo.com.br

Luiz Vasconcelos
Research and Development Department
Globo TV Network
Rio de Janeiro, Brazil
luiz.vasconcelos@tvlgobo.com.br

Categories and Subject Descriptors

J.2 [**Physical Sciences and Engineering**]: Engineering

1. INTRODUCTION

Nowadays, the automatic processing for 3D facial animation has been studied intensively and a variety of techniques has been proposed so far. For instance, some facial animation softwares are based on audio capture [5] and others are based on motion capture [7, 6]. Independently of the approach used, it is fundamental that the capture solution assure lip movements synchronization with audio. Surely, this issue represents the state-of-art and many efforts have been made to solve it. Lip synchronization is the most complex process in the production of an animation and, commonly, it takes a long time.

Since 2004, a virtual reality team of Globo TV Network has been developing a virtual reporter, named *Eva Byte* [2]. She always appears on news program presented on Sunday nights. To perform her production, it is common to animate up to 2000 frames during three days of work. In order to obtain excelent results, it is necessary to spend more time with mouth animation to avoid problems of lip synchronization. For this reason, the first step of the animation process consists on marking each keyframe, using 3D Studio Max environment [1], each keyframe corresponding to a specified phoneme. After that, for each keyframe found, all the necessary adjustment is done to obtain its 3D model. Obviously, this type of process is quite work intensive and the animator becomes very tired in the end.

In this work, a new automation method for the lip animation process, based on the WISARD weightless neural network model, is detailed [4]. In the initial step, we populated the image database that contains mouth images using various videos previously recorded of a real presenter, each one stored as bitmap representing different phonemes (see Figure 1). Initially, 300 images with dimension 88x88 are used. A few samples are separated for training a WISARD network dedicated to finding the mouth position into video frames. Once a mouth is locked in a video frame, another WISARD network is applied to retrieve the most similar mouth class, in the image database, the target mouth belongs. This is done by associating each mouth class, each representing a specific phoneme, to a WISARD neural discriminator. Thus, the neural discriminator producing the maximum output represents the mouth class, and also the 3D model, of the mouth appearing in that frame. Conversely, when the mouth found in the current frame is not visually compatible with the classification produced by the winning WISARD descriminator, and/or the winning discriminator is low, the system allows for users to include the "new mouth" type into the database at anytime, without stop the animation process. At the same time, the WISARD network learns the new mouth pattern, either by associating it to an existing class or by creating a new WISARD descriminator.

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-733-9/07/0007.

2. WISARD

WISARD is an adpative pattern recognition machine which is based on neural principles [3]. The WISARD's neuron is a RAM-type memory unit, having n address inputs and it is able to store 2^n bits (all positions are initialized with 0's). That way, each neuron is able to learn and recognize n-bit words ("tuples"). The training of a new tuple consists in writing '1' on the neurons' position addressed by it; the positive recognition of an input tuple bu a neuron is merely checking if it addresses a stored '1'.

The RAM neuron, while fast, lacks generalisation power: it only recognises previously learned tuples. To overcome this limitation, a set of neurons can be organized in a structure called discriminator, where each neuron is responsible for the learning/recognition of a subset of a (bigger) input pattern (Figure 2). The subpattern assigned to each neuron is defined in a randomly created input-neuron mapping, which is used in both learning and recognition phases. The training of a discriminator consists in writing 1's in the posi-

Figure 1: Examples of mouths stored onto database.

tions addressed in each of the discriminators' neurons. The recognition of a pattern is given by analysing the discriminator's output, which is the sum of the neurons' outputs bits for that pattern. By having a graded output, a discriminator can recognise similar but different version of a trained pattern, thus showing generalisation ability.

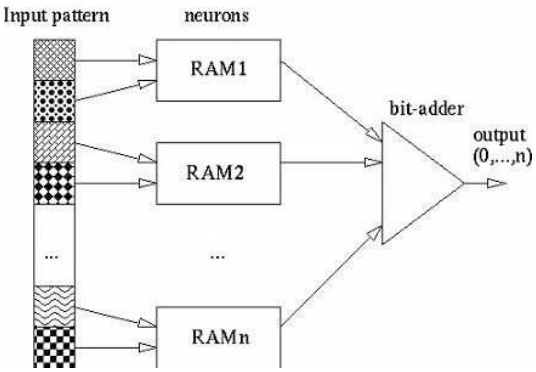

Figure 2: The discriminator.

The WISARD network is an array of discriminators, each representing a different class of patterns (Figure 3). The WISARD is trained in a supervised fashion: one must select a discriminator and train it with selected patterns from the respective class. The determination of a pattern's class by a WISARD is made in a competitive way: the input pattern belongs to the discriminator which presented the highest recognition level (output) for that input, assigning to itself the winner discriminators' class label.

3. IMPLEMENTATION

The semi-automated lip animation system was designed using two different WISARD networks. The first one is capable of finding the correct position of the mouth inside

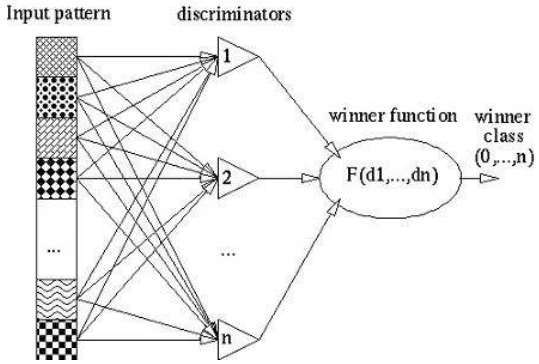

Figure 3: The WISARD network.

a frame. For this purpose, a few samples taken from the database of mouth images are used to train a WISARD discriminator to recognize mouths. Once a mouth is found, it is necessary to match it with the most similar mouth present in the image database. In order to achieve that, a WISARD network was designed having each discriminator trained to recognize one mouth image in one mouth class. Thus, a total of 300 neural discriminators were implemented.

For each specific frame, the system searches for the mouth position using the first discriminator. This operation is performed through a search window with the same dimension of the mouth image. Initially, this is done at all pixels of the input image. As a result, the position that induces the maximum output count for the mouth discrimininator, represents a mouth found. After this processing step, some patterns are presented to the second WISARD network around the position found. This technique allows for precisely determining which descriminator, among the 300 ones, gives the biggest output count. Consequently, the winner discriminator represents the most similar mouth for that pattern. As the 3D model is associated to 2D models (images), it is possible to import these 3D parameteres into the graphic animation software, like 3D StudioMax.

Figure 4 shows the semi-automated lip animation system developed. The woman shown is the journalist that provides all movements and expressions to be reproduced by the Eva Byte virtual reporter. As a result, the mouth that appear into the small square is the most likely for the current frame, considering all the 300 possible classes.

4. WALKTHROUGH

In this section, we want to describe the technical demo to be presented. Initially, a brief explanation about concepts related to 3D animation and lip sybchronization will be discussed. After that, the semi-automated lip animation system developed in order to obtain a almost perfect lip synchronization is presented (Figure 4). Moreover, we will demonstrate all advantages of this implementation describing the WISARD neural system as a powerful tool for image retrieval. Finally, we will show some Eva Byte's videos that were made using the system presented in this work. Figure 5 shows some frames of Eva Byte on news TV programs.

Figure 4: The software developed.

Figure 5: Eva Byte on news programs.

5. ACKNOWLEDGMENTS

This work was supported by Globo TV R&D Department. The authors would like to thank Mr. Luiz Amaral, Mr. Flávio Reis and Mr. Ricardo Morais, all of them creators of Eva Byte.

6. REFERENCES

[1] 3d studio max. http://www.the3Dstudio.com.

[2] Eva byte. http://www.globo.com/fantastico.

[3] I. Aleksander and H. Norton. *An Introduction to Neural Computing*. Chapman and Hall, London, 1990.

[4] I. Aleksander, W. Thomas, and P. Bowden. Wisard, a radical step forward in image recognition. *Sensor Ver.*, pages 120–124, 1984.

[5] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM Trans. on Graphics.*, 24(4), October 2005.

[6] C. Curio, M. Breidt, and M. Kleiner. Faces and gestures: Semantic 3d motion retargeting for facial animation. In *Proceedings of 3rd symposium on Applied perception in graphics and visualization APGV*. ACM, July 2006.

[7] A. Hornung, E. Dekkers, and L. Kobbelt. Character animation from 2d pictures and 3d motion data. *ACM Trans. on Graphics.*, 26(1), January 2007.