MULTIOBJECTIVES GENETIC SNAKES: APPLICATION ON AUDIO-VISUAL SPEECH RECOGNITION

Renaud Séguier, Nicolas Cladel

Supélec, Avenue de la Boulaie - BP28 35511 Césson-Sévigné, France.
Renaud.Seguier@supelec.fr, Nicolas.Cladel@supelec.fr

Abstract: *We propose in this article a new optimization of Genetic Snakes (GS): Multiobjectives Genetics Snakes (MGS) faster and simpler to implement. They enable us to make converge two snakes in parallel while minimizing energies of different nature. We apply them to the modeling of mouth contours within the framework of the Audio-Visual Speech Recognition (AVSR). The proposed AVSR system implements a classifier based on the STM (Sparse Template Matching) which simplicity makes it possible to consider a real time implementation. We evaluate the classifier performances on European database M2VTS, and compare the performances of the GS and MGS.*

Keywords: *Audio Visual Speech Recognition, Multiobjective Optimization, Genetic Snakes, lipreading, Sparse Template Matching.*

1. INTRODUCTION

Genetic algorithms optimization qualities have been used for several years in image processing [1]. In particular their aptitude to avoid local minima [2] makes them very attractive. The Genetics Snakes [3][4][5] give thus the possibility to overcome the problem of snakes initilization. The snakes identify a contour through the minimization of a weighted sum of several different energies (internals related to the described shape, externals related to the image). The weighting coefficients are difficult to find especially when contours to be identified vary from one image to another.
The multiobjective optimization gives a solution to this problem by considering in parallel various energies. In particular the multiobjective optimization techniques based one genetic algorithms [6][7] propose various approaches since about fifteen years. We presente in this article a multiobjective implementation of the snakes optimization by genetic algorithms.
In the field of Audio-Visual Speech Recognition (AVSR), several techniques were already proposed from neural networks [8] till HMM [9].
The algorithms which we develop are dedicated to embedded systems having very few resources. For this reason, we introduce here the use of the Sparse Template matching (STM) [10][11] in the field of the AVSR. This extremely simple classifier allows us to consider a real time implementation on light systems like PDA.

2. MULTIOBJECTIVES GENETIC SNAKES

First of all we will present the preprocessing which generate the contours points used by the snakes to model the mouth. Then we will detail the chromosomes coding used in the algorithm and at last we will describe various energies taken into account. Their parallel optimizations within the multiobjective framework will be finally introduced.

2.1. Image preprocessing

We locate [11] the face and the line Lmouth (see Fig 1a). In the V values (from YUV color coordinate system), the lips has a strong level of intensity while the teeth and the dark interior of the mouth are confused and rather dark (Fig.1b).

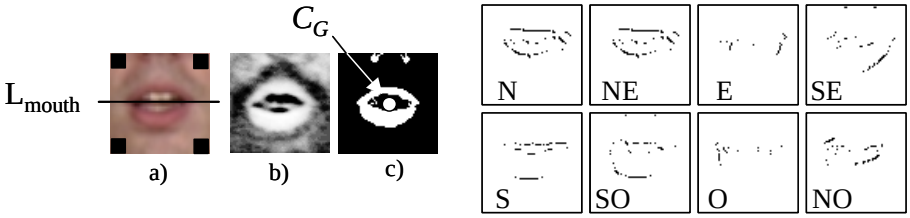

Fig. 1. Mouth Preprocessing.

On the first images, we evaluate the RVB signature of the pixels belonging to the lips (those which have a height V value around Lmouth). We also memorize the signature of those which belong to the skin (black areas in Fig. 1a.). We then classify the pixels by evaluating the Euclidean distance between the value RVB of the pixel and each of the two signatures. (Fig 1c). The gravity center C_G of the interior of the mouth is then evaluated starting from the white pixels of image c), Fig. 1. The edges are finally extracted from this image.

2.2. Chromosome coding

Our aim is to find a first snake on the external lips contour and a second one on the interior contour. Each one of these snakes is defined on eight nodes. To accelerate optimization, we make evolve the snakes only on contours points. Thus the node C (Fig 2a) will be defined only on North, North-East and North-West (Fig. 1) contours points in the area RC (Fig. 2d). This area is defined starting from the gravity centre C_G of the mouth previously given.

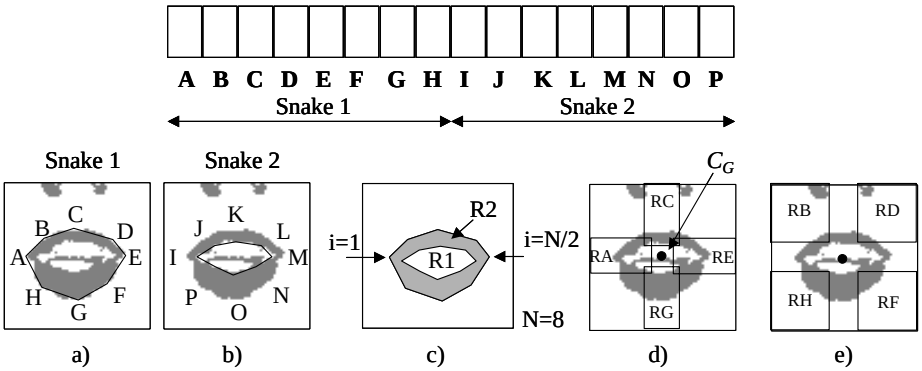

Fig. 2. Chromosome coding.

All the other nodes of the first snake are defined in the same manner by taking into account the area in which they must evolve and the contours orientation which characterize them. With regard to interior contour, insofar as the mouth is sometimes closed, it is difficult to define in a robust way the areas in which the nodes of the second snake must evolve. This is why we take into account the whole set of contours of the image c), Fig. 1 knowing that the nodes of the second snake can belong to any area of the image inside the first snake.
The position of each node of both snakes is coded on a chromosome gene as the Fig. 2 indicates it. Thus, the tenth gene codes the value $x+yL$ if L is the width of the image and (x,y) the coordinates of the node J belonging to the second snake.

2.3. Energie evaluation

Four different energies must be minimized to model the lips contours.
The first energy E_{Bend} makes it possible to control the rigidity of the curve. This constraint is evaluated on the whole nodes set of both snakes except for those corresponding to mouth corners (A, E, I, and M, see Fig.2ab)

$$E_{Bend} = \sum_{i=2, i\neq 1+N/2}^{N} \left\| V1_{i-1} - 2V1_i + V1_{i+1} \right\|^2 + \sum_{i=2, i\neq 1+N/2}^{N} \left\| V2_{i-1} - 2V2_i + V2_{i+1} \right\|^2 \quad (1)$$

where i is the number of the node on the curve knowing that a snake contains N nodes. The first node is on the left corner of the mouth (A or I in the Fig. 2ab), the node 1+N/2 is on the right corner of the mouth (E or M). Vj_i is the i-th node of the snake j.
The second energy E_{DarkS2} takes into account the quantity of pixels $NbLipsPix_{S2}$ characterizing the lips (white pixels of the Fig.1c) and belonging to the area R1 of Fig.2c.

$$E_{DarkS2} = NbLipsPix_{S2} \quad (2)$$

The third energy $E_{DarkS1S2}$ takes into account the quantity $NbDarkPix_{S1S2}$ of pixels characterizing the mouth interior or the skin (black pixels of Fig.1c) and belonging to the lips region described by the snakes (area R2 of Fig. 2c).

$$E_{DarkS1S2} = NbDarkPix_{S1S2} \quad (3)$$

The last energy $E_{LipsS1S2}$ takes into account the quantity $NbLipsPix_{S1S2}$ of pixels characterizing the lips (white pixels of Fig. 1c) and belonging to the lips region described by the snakes (area R2 of Fig. 2c).

$$E_{LipsS1S2} = -NbLipsPix_{S1S2} \quad (4)$$

2.4. Multiobjectives

According to the propositions of [12], after having evaluated energies of each chromosome, we rank the population on the basis of Pareto nondomination. A linear ranking with selective pressure 2 and a proportional fitness are then applied. A sharing function is used (σ_{share}=0.15) to maintain diversity. Because it is necessary to keep the good solutions all along the iterations [7], we create a second population in the following way.
We use a population P1 (*N1* chromosomes) and a population P2 of which the chromosomes number is augmented of one at each iteration. The algorithm is reiterated *Niter* time on a population P3 of which the chromosomes number varies from *N1* (at the first iteration) to *N1+N2* (starting from the iteration *N2* and until the iteration *Niter*).
At the beginning of each iteration, P3 is made up of P1 and the *N2* better chromosomes of P2.

At the end of each iteration (after having applied crossover and mutation operators on P3):

- P1 is made up of the N1 better chromosomes of P3,
- P2 is increased with the addition of a chromosome: the best of P3.

In this manner a chromosome representing a good solution during the convergence will not be lost, as that is often the case when is applied a multiobjective strategy not using a second population.

To define the best chromosome of a population, we proceed in the following way. We rank each chromosome according to the others by taking into account each energy separately: a chromosome will thus be characterized by a vector R of n ranks (r_1 to r_n). The value r_j is then the rank of a chromosome compared to the others by considering energy number j. The best chromosome will be that which will have the weakest rank for each energy. We calculate an Euclidian norm of this rank vector R for each chromosome. Since a weak rank characterizes an optimized energy, the chromosome having the smallest norm will be regarded as the best.

In our applications, we use a classical crossover (probability: *0.9*), a uniform mutation (probability: *0.03*) and *N1=N2=10* chromosomes. At the end of the *Niter=10* iterations, the snake coded by the winner chromosome describes the contours of the lips.

3. AVSR SYSTEM

The system describes in Fig. 3 consist of an audio and video preprocessing module and of a classifier based on Sparse Templates.

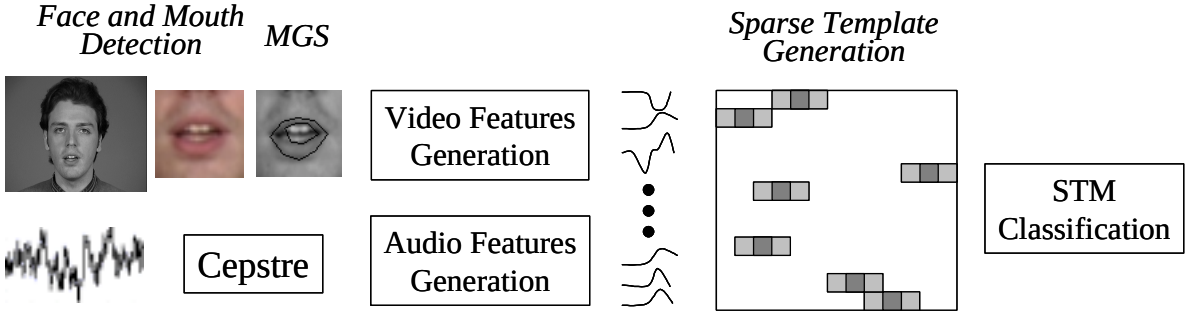

Fig. 3. Audio-Visual Speech Recognition System.

Audio Prepreocessing. On a 40ms sliding window, we calculate as in [13] the first 12 coefficients of the cepstrum, the logarithm of the signal energy in the window and the temporal derivate of those thirteen parameters.

Video Prepreocessing. We extract the height and the width mouth starting from the snakes and evaluate the percentage of dark and light pixels in the mouth [11]. Finally we calculate the temporal derivative of these parameters.

We normalize each audio and video parameters over the entiere sequence.

STM Classifier. We use the same classifior as [10]: starting from the audio and video parameters we generate a Sparse Template (ST). A matching between the ST identifying the signature of each word to be recognized and the current ST enables us to identify the word.

4. RESULTS

Within the framework of separated word recognition, we tested our system on the first person of the European Data Base M2VTS (Multi Modal Checking for Teleservices and Security applications [14]). This base is dedicated to audio-visual recognition and identification. The person pronounces four times (at one week interval) the digits from 0 to 9. We chose this base because it characterizes well the conditions of use in which the real time implementation of our system will have to function. The images were acquired at 25Hz with a weak resolution (288x360 pixels in 4:2:2).

Multiobjectives Genetic Snakes. Applied to this type of image, Multiobjective Genetic Snakes (MGS) give the same results as Genetics Snakes (GS) as one can see it on Fig. 4.

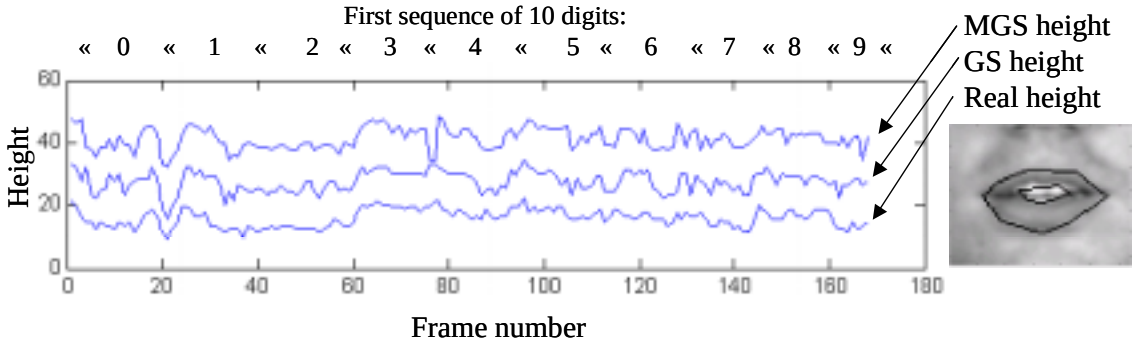

Fig. 4. MGS, GS and real mouth height along the first sequence.

All the interest of the MGS comes from their simplicity of implementation, and the speed of convergence in this application. Ten iterations are enough (against 300 within the framework of the GS), a less great number of chromosomes are necessary (from 10 to 20 at the end of the convergence against 30 in the case of the GS) and a shorter running time: 14s against 52s (Matlab implementation on a P3-900Mhz, video preprocess included). In one pass learning, four evaluations are carried out according to the number of the sequence used for the training (tests being performed on the three other sequences). In this framework, the lipreading is carried out correctly in 77% of the words (GS or MGS).

STM Classifier.

As Fig. 5 indicates it, the fusion of the audio and video signals makes it possible to improve the audio recognition rates when we add white noise on audio signal.

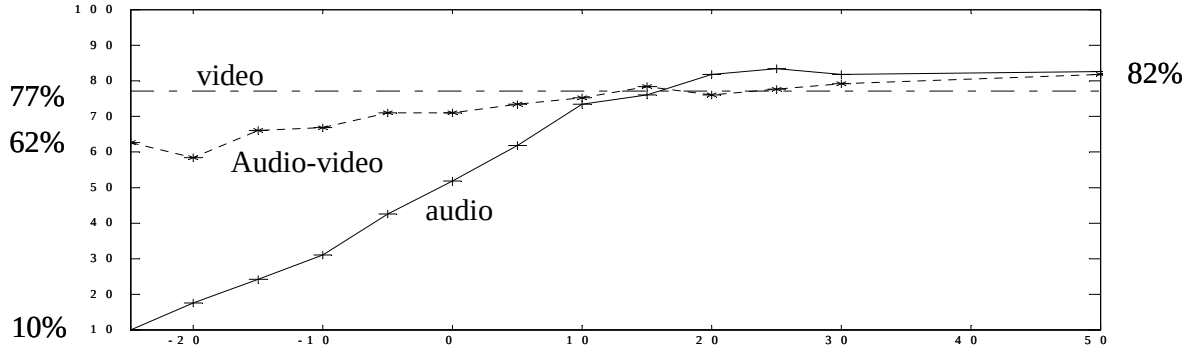

Fig. 5. Percentage of correct classification versus signal to noise ratio.

5. CONCLUSION

We proposed in this article a new optimization of the Genetic Snakes : the Multiobjectives Genetics Snakes illustrated in an AVSR system. We showed that the MGS were at the same time simpler to implement and more rapid than GS. Moreover they allow us to make converge in parallel two snakes on different energies and permit the description of the lips contours. However, it remains us to apply them not on only one person but on the whole M2VTS database in order to really illustrate their robustness.

Acknowledgment. This research was supported by Brittany Region ("Région Bretagne") in France.

REFERENCES

[1] C. Bounsaythip and J.T. Alander, Genetic Algorithms Applied to Image Processing - A Review, *Proc. of the 3rd Nordic Workshop on Genetic Algorithms (3NWGA)*,1997.
[2] K. Sakaue and A. Amano and N Yokoya, Optimization approaches in computer vision and image processing, *IEICE Trans. Inf. and Syst.*, 1999.
[3] A. Cagnoni and A. Dobrzeniecki and R. Poli and J. Yanch, Genetic algorithm-based interactive segmentation of 3D medical images, *Image and Vision Computing*, 17(12):881-895, 1999.
[4] L. Ballerini, Genetic snakes for color images segmentation, *Lecture Notes in computer sciences 2037*, 2001.
[5] N. Covavisaruch and T. Tanatipanond, Deformable Contour for Brain MR Images by Genetic Algorithm: From Rigid toTraining Approaches, *Proceedings*, *Image and Vision Computing New Zealand (IVCNZ '99)*, 1999.
[6] Carlos A. Coello, An updated survey of {GA-based} multiobjective optimization techniques, *ACM Computing Surveys*, vol. 32, num. 2, 2000.
[7] David A. van Veldhuizen and Gary B. Lamont, Multiobjective Evolutionary Algorithms: Analyzing the State-of-the-Art, *Evolutionary Computation*, vol. 8, num. 2, 2000.
[8] P. DAUBIAS and P. DELEGLISE, Evaluation of an automatically obtained shape and appearance model for automatic audio visual speech, *Eurospeech*, 2001.
[9] S. Dupont and J. Luettin, Audio-Visual Speech Modeling for Continuous Speech Recognition, *IEEE Transactions on multimedia*, 2000.
[10] G. Sullivan and all, Model-based vehicle detection and classification using orthographic approximations, *Proc. of 7th British Machine Vision Conference*,1996.
[11] R. Séguier and Nicolas Cladel, Genetic Snakes. Application on Lipreading, *International Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA)*, 2003.
[12] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Ed. Addison-Wesley 1989.
[13] R. Séguier and D. Mercier, Audio-visual speech recognition: one pass learning with spiking neurons, *International Conference on Artificial Neural Networks (ICANN)*, 2002.
[14] S. Pigeon, M2VTS, *www.tele.ucl.ac.be PROJECTS/M2VTS/m2fdb.html*, 1996.