On Model-Based Analysis of Ear Biometrics

Banafshe Arbab-Zavar, Mark S. Nixon, and David J. Hurley

***Abstract*— Ears are a new biometric with major advantage in that they appear to maintain their structure with increasing age. Most current approaches are holistic and describe the ear by its general properties. We propose a new model-based approach, capitalizing on explicit structure and with the advantages of being robust in noise and occlusion. Our model is a constellation of generalized ear parts, which is learned off-line using an unsupervised learning algorithm over an enrolled training set of 63 ear images. The Scale Invariant Feature Transform (SIFT), is used to detect the features within the ear images. In recognition, given a profile image of the human head, the ear is enrolled and recognised from the parts selected via the model. We achieve an encouraging recognition rate, on an image database selected from the XM2VTS database. A head-to-head comparison with PCA is also presented to show the advantage derived by the use of the model in successful occlusion handling.**

I. INTRODUCTION

Ears have long been considered as a potential means of personal identification, yet it is only in the last 10 years or so that machine vision experts started to tackle the idea of using ears as a biometric. French criminologist Alphonse Bertillon was the first to recognize the biometric potential of human ears [1]. Empirical evidence supporting the ear's uniqueness was later provided in studies by Iannarelli [12]. Ears have appealing properties for personal identification; they have a rich structure that appears to be consistent with age from a few months after birth. Clearly, ears are not affected by change in facial expressions. Images of ears can be acquired without the subject's participation and ears are big enough to be captured from a distance. However there exists a big obstacle — the potential occlusion by hair and earrings, which is almost certain to happen in uncontrolled environments.

One of the first ear biometric systems was introduced by Burge and Burger [2]. They modeled each individual ear with an adjacency graph. Hurley et al. [11] used force field feature extraction to map the ear to an energy field which highlights "potential wells" and "potential channels" as features. Chen et al. [4] and Yan et al. [21] exploited the 3D structure of the human ear. Yuizono et al. [22] treated the problem as an optimization task, and proposed a specially-developed genetic local search. Moreno et al. [16] used different combinations of several neural classifiers. Principal Components Analysis (PCA) approaches have also been applied in a number of studies [3], [20], [13]. However PCA has no invariance properties, thus it relies on the acquisition and pre-processing stages to window and align the data. An up-to-date survey of ear biometrics has recently been provided by Hurley et al. [10].

Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

Despite all the success in ear biometrics, no model based approach has yet been introduced. A model is explicit in its approach to identification. Being an abstract form of the object, a model capitalizes on the specific structures and thus prunes out all unnecessary detail. Furthermore, it has the advantage of being robust in noise and occlusion and has potential advantage in viewpoint invariance. Therefore we propose a new model-based approach. Using an unsupervised learning algorithm the ear model is learned from a dataset of ear images. We contend that for practical deployment a planar image of an ear is a more likely application scenario than deployment of 3D imaging and have thus concentrated on analyzing 2D images of the side view of the human head. We shall describe our new ear model in section 2, explaining our feature extraction and learning techniques. In section 3 we apply the model to a recognition task, and make a head-to-head comparison with PCA in occlusion scenarios, followed by conclusions and further work.

II. APPROACH

Our approach to modeling ears is based on distinguishing individual ear parts. The model is comprised of a number of parts, each having its specific appearance. It is worth noting that although ears might seem like random shapes, they do in fact have a definite structure just like the face. Moreover medical studies suggest that the shape of the auricle is determined by the individual growth of 6 small nodules during embryonic development [17], which supports describing ears by individual parts.

Our ear model is determined via an unsupervised learning process using a dataset of 63 ear images. Each ear image is represented by a set of features. The clusters of these features across the dataset denote the common ear features. The model is then learned by detecting these clusters and expressing them by their statistical properties. Thus our model can be considered as a constellation of generalized ear-features, the structure of which resembles the constellation models [8] used to recognize object categories.

A. Feature Extractor

The Scale Invariant Feature Transform (SIFT) [14] automatically extracts potential interest points in images in a consistent manner. These features which are marked at the location of the scale-space extrema, are called the keypoints, and describe neighborhoods of pixels. The size of each neighborhood is proportional to the scale in which the feature is detected. Therefore these keypoints describe the object

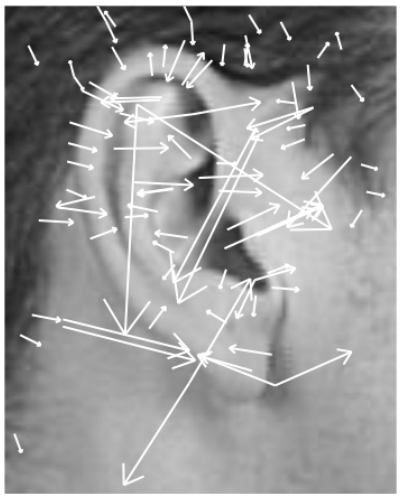


Fig. 1. The SIFT keypoints for an ear image in the XM2VTS database.

parts which are depicted in these specific neighbourhoods. SIFT's special design enables it to detect the most stable features across all possible scales.

Considering the properties of SIFT, it appears an attractive choice for feature extraction, and it is frequently used to detect features for object modelling purposes, either by itself [9] or in conjunction with other feature detectors [7], [18]. We apply SIFT following enrolment based on ear shape.

The keypoints determined by SIFT have assigned locations, scales and orientations. A distinctive descriptor is also assigned to each keypoint. These descriptors, which are 4×4 arrays of orientation histograms, are normalized with respect to scale and orientation, therefore SIFT is scale and rotation invariant. Furthermore, they are partially invariant to changes in illumination and viewpoint. The latter is particularly beneficial in 2D ear biometrics, where the ear data might be acquired from slightly different viewpoints. Lowe has shown that these assigned descriptors are highly distinctive, which allows a single feature to find its correct match with good probability in a large database of features [14]. Figure 1 shows an ear image with the detected keypoints superimposed. In this figure each keypoint is depicted by a vector which is drawn using the location, scale, and orientation of the respective keypoint.

B. The Ear Model

Our ear model is constructed using a stochastic method. In this method the SIFT keypoints of each ear image are repeatedly presented to the construction algorithm. This is known as recycling, and each cycle is called an epoch [5]. Each epoch comprises two steps: *(i)Updating clusters.* The matching keypoints between the ear and the model are determined, and the model is updated by modifying the matched keypoints using a cumulative average and also by adding the unmatched keypoints of the ear image to the model. *(ii)Revision.* A hierarchical clustering algorithm detects the clusters of keypoints in the model. These clusters are then merged to eliminate duplicate keypoints, which would otherwise cause a division of focus.

Fig. 2. A sample of training set images.

A derivative based measure is used to quantify the model alteration in each epoch. The recycling terminates when the derivative remains below a specified threshold for three consecutive epochs, indicating that the keypoints of the model have stabilised. While the description of the database we have used will come later in the results section, it is important to note here that accurate enrolment has been used in training by cropping to the average ear size to ensure the use of ear features only. Examples of the images in the training set are shown in Figure 2. Let the dataset D with ND feature sets P_m (extracted from ND images) be,

$$D = \{P_m\} \ , \ m = 1...ND \ . \tag{1}$$

The extracted feature set from each ear image P_m consists of the descriptors $\mathbf{d}$ and the locations $\mathbf{r} = \{x, y\}$ of the keypoints K which are determined using SIFT,

$$P_m = \{K_{m_i}\} = \{(\mathbf{d}_{m_i}, \mathbf{r}_{m_i})\} \ , \ i = 1...NP_m \tag{2}$$

where NP_m is the number of keypoints that are detected in the m^{th} image.

A composite distance measure d is used to determine the matching keypoints. This measure combines the normalized match score for the locations and the normalized match score for the descriptors. The normalized scores are obtained by subtracting the mean and dividing by the standard deviation of distance distributions of a set of manually matched keypoints between two ear images in the training set. These matched keypoints provide us with a rough estimate for (μ_d, σ_d) and (μ_r, σ_r) which are the mean and standard deviation of the Euclidean distances of descriptors and locations of the matching keypoints in the training set respectively. The training images are well registered and thus the locations of matching keypoints are correlated. Let $\|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$ be the Euclidean distance. The distance d between two keypoints is,

$$\begin{aligned} d(K_i, K_j) &= d\{(\mathbf{d}_i, \mathbf{r}_i), (\mathbf{d}_j, \mathbf{r}_j)\} \\ &= \frac{\|\mathbf{d}_i - \mathbf{d}_j\| - \mu_d}{\sigma_d} + \frac{\|\mathbf{r}_i - \mathbf{r}_j\| - \mu_r}{\sigma_r} \end{aligned} \tag{3}$$

when both normalized scores are less than three standard deviations. Otherwise the two keypoints will be reported as mismatched keypoints.

Let $\mathbf{cd}$ and $\mathbf{cr}$ denote the cumulative average of descriptors and locations respectively, and let $n_j(k)$ be the number of keypoints that have contributed to the respective cumulative average, j, up to the k^{th} epoch. For all the image keypoints, those which are sufficiently close to a labeled model keypoint

contribute to it as,

$$\begin{cases} (\mathbf{cd}_j(k+1), \mathbf{cr}_j(k+1)) = \frac{\sum_{i=1}^{k} (\mathbf{d}_x(i), \mathbf{r}_x(i))}{n_j(k)} \\ \\ \text{if} \quad d\{(\mathbf{d}_x(i), \mathbf{r}_x(i)), (\mathbf{cd}_j(i), \mathbf{cr}_j(i))\} < \text{threshold1} \end{cases} \tag{4}$$

where $(\mathbf{cd}_j(k+1), \mathbf{cr}_j(k+1))$ and $(\mathbf{cd}_j(i), \mathbf{cr}_j(i))$ are two model keypoints in the $k+1^{th}$ and i^{th} epochs respectively. The index j indicates that $(\mathbf{cd}_j(k+1), \mathbf{cr}_j(k+1))$, which evolved through k epochs , is the descendant of $(\mathbf{cd}_j(i), \mathbf{cr}_j(i))$. Each of the image keypoints which are not matched to a model keypoint initiates a new keypoint in the model with their own descriptor and location.

Choosing a simple structure for the model, we are able to allow the number of clusters to grow as necessary without facing intractable computational problems, and thereby accurately accommodate the variations in the input keypoints.

The model at the k^{th} epoch $Mod(k)$ is a cumulative average with $N(k)$ keypoints,

$$Mod(k) = \{(\mathbf{cd}_i(k), \mathbf{cr}_i(k), n_i(k))\} \ , \ i = 1...N(k) \ . \tag{5}$$

At each epoch the model keypoints are updated, therefore the distance d between them alters. As a result in some cases this distance might fall beneath a distinction threshold,

$$d(Mod_i(k), Mod_j(k)) < \text{threshold2} \ . \tag{6}$$

This will cause the same entities in different ear images to have different corresponding keypoints in the model which is obviously not desirable. Therefore the hierarchical clustering algorithm is applied at the end of each epoch to detect these duplicated keypoints, which are then merged. Let $Mod_l(k)$ denote the new keypoint which replaces $Mod_i(k)$ and $Mod_j(k)$,

$$Mod_l(k) = \frac{n_i(k) \times Mod_i(k) + n_j(k) \times Mod_j(k)}{n_i(k) + n_j(k)} \ . \tag{7}$$

The ear model is a constellation of keypoints each describing a part of the ear that is constantly visible and distinguishable in ear images. However this model is obscured by the mass of isolated keypoints which were added to the structure so that their potential as a model keypoint would be assessed, but they have failed to construct well populated clusters. The model is revealed when we apply a threshold on the clusters' cardinalities to prune these isolated keypoints.

The recycling is terminated when the model alteration is not significant in three consecutive epochs. To detect this stable state an estimate of the model evolution rate $m(k)$ is obtained by measuring the distance between the model keypoints in adjacent epochs,

$$\begin{cases} m(k) = \sum_{i \in C(k)} \|Mod_i(k) - Mod_j(k-1)\|, \\ \\ if \quad j \in C(k-1) \ , \ \forall \ j_0 \in C(k-1), \\ \|Mod_i(k) - Mod_j(k-1)\| \leq \\ \qquad \|Mod_i(k) - Mod_{j_0}(k-1)\| \end{cases} \tag{8}$$

where $C(k)$ and $C(k-1)$ are the sets of sufficiently populated clusters in k^{th} and $(k-1)^{th}$ epochs respectively.

Fig. 3. 4 samples of images in the XM2VTS database and the result of the automatic enrolment which produces 150×120 sized images of the ear regions.

III. RESULTS

To validate our model we perform a recognition test on a database of 63 individuals. A head-to-head comparison with PCA recognizing occluded probes is also presented to show how well our model handles occlusion.

A. Ear Database

We have used a database of 63 individuals, selected from the XM2VTS [15] face-profile database. These 63 individuals are those whose ear is not obscured by hair. Therefore our ear database, which comprises 4 images per individual taken in 4 different sessions over a period of five months, contains $63 \times 4 = 252$ images. This database is the same as recently used by Hurley et al. [11]. One image from each of the 63 individuals is manually registered and used for training, whilst the remainder are used for performance evaluation.

B. Ear Recognition

In our first experiment we demonstrate our model's capabilities in ear recognition. As just mentioned the test set is 3 profile images out of four for each subject, however, these profile images contain irrelevant information such as hair, eyes, neck, etc. An automatic enrolment process based on finding the elliptical shape of ears locates the ear regions by using a Hough transform for ellipses to gather votes for putative ellipse centres in an accumulator; the location of the peak in this accumulator gives the coordinates of the best matching ellipse. Once these coordinates are determined, 150×120 sized images are derived in which the ears are roughly placed in the centre. Figure 3 shows 4 images from the XM2VTS and their ear regions which were detected using this automatic enrolment.

We use our ear model to redefine the feature vectors of the ear images. The initial feature vectors are the sets of keypoints that are detected using SIFT. The model acts as a mask in keypoint selection; only those keypoints in the model are used for recognition. Our ear model comprises $n = 20$ keypoints. Therefore the cardinality of all the new feature vectors is also 20. Let MF be the new feature vector of the m^{th} image in the test set,

$$MF_m = \{\mathbf{md}_{m_i}\} \ , \ i = 1...n \tag{9}$$

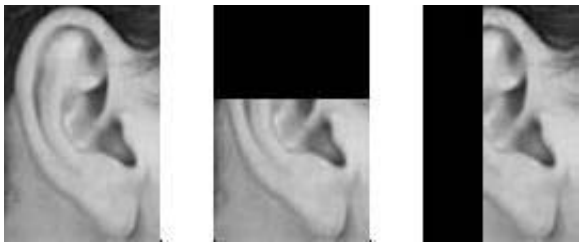


Fig. 4. An ear image (left) becomes occluded 40% from top (middle), and 40% from left (right).

where $\mathbf{md}_{m_i}$ is the descriptor of the chosen keypoint corresponding to the i^{th} model keypoint. The model selects these keypoints according to the best descriptor match criterion. Thereby the distance d' between two images is defined as the mean of Euclidean distances between their corresponding keypoints;

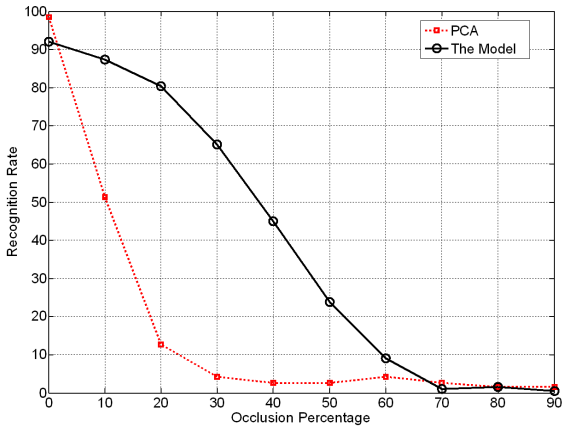
$$d'(im1, im2) = d'(MF_1, MF_2) = \sum_{i=1}^{n} \frac{\|\mathbf{md}_{1_i} - \mathbf{md}_{2_i}\|}{n} \quad . \tag{10}$$

We have used k-nearest neighbour classification with k=1 on the distance d' for recognition, and out of 189 trials, we achieved 165 correct classifications, which equates to an 87.3% recognition rate. In comparison applying PCA [19] to images with the same enrolment process achieved a recognition rate of 75.1% (See table I). K-nearest neighbour with k=1 on Manhattan distances has been used for PCA recognition.

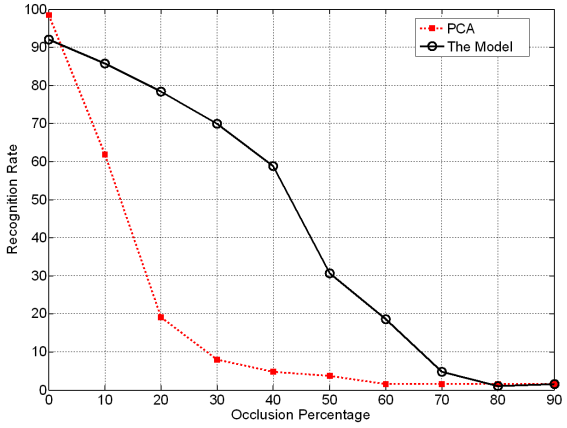
C. Occlusion Test

One of the biggest advantages of a model-based approach is its robustness in occlusion. In this experiment we assess our model capability in handling occlusion. Clearly more occlusion means less information to decide upon. In the limit the ear is totally occluded and recognition is impossible. Thus it is only natural that the recognition rate drops as the ears become more and more occluded. However model-based methods are better at dealing with missing data. We choose PCA as a good example of the holistic approaches since it is a well developed and widely used method. PCA provides us with some benchmark results against which we can gauge our results.

For this test, synthetically occluded probes are compared against a gallery of unoccluded images. This scenario is supported by the actual conditions of a working system in which we presume the gallery images are taken in advance, under controlled conditions. However less control is applied when acquiring the probes, thus they might be occluded. Due to PCA's inability to offer satisfactory results for our automatically enrolled data (See table I), we have used a manually registered test set where images are rotated to a mean angle and registered to 111×73 sized images. This registration was specially developed by Hurley et al. [11] to reinforce the PCA's recognition capabilities, for which PCA achieves a 98.4% recognition rate, and our model



(a) Occlusion from top



(b) Occlusion from left

Fig. 5. Model-based recognition compared to PCA-based recognition for occlusion from top and left.

yields a 91.5% recognition rate which equates to 173 correct classifications out of 189 probes. Furthermore Hurley et al. [11] reported a recognition rate of 99.2% using the force field transform on the same data with poor registration.

Observing the partially occluded images in the XM2VTS database, it appears that one of the most common types of occlusion by hair in ear images is occlusion from the top. Therefore we occlude the probe images from the top with solid black colored bars which grow toward the lobe to present more occlusion. We also examine the effects of occlusion inwards from the helix (left to right for our database). Synthetic occlusion has been used to give better judgement and control over the extent of occlusion and generate more samples for performance evaluation. Figure 4 shows examples of occluded probes.

K-nearest neighbour classification with k=1 on the dis-

TABLE I

COMPARISON OF EXAMPLE MODEL-BASED AND PCA RESULTS.

	Model-based	
	Recognition(%)	Decidability
Manual Registration (111×73 sized image)	91.5%	2.36
Automatic Registration (150×120 sized images)	87.3%	2.71
20% occlusion from top	80.4%	1.96
	PCA	
	Recognition(%)	Decidability
Manual Registration (111×73 sized image)	98.4%	3.56
Automatic Registration (150×120 sized images)	75.1%	1.90
20% occlusion from top	12.7%	1.26

tances d' (10) and Manhattan distances is used for model-based and PCA recognition respectively, and their recognition rates in various occlusion scenarios are depicted in figure 5. Table I shows example results of the model-based and PCA methods. The decidability is the Daugman's decidability measure [6], which evaluates the potential decisiveness of a biometric task. As can be seen in figure 5, the model-based approach achieves much better results than PCA for occluded images. Therefore we propose that our model-based approach is suitable for ears which suffer from a high likelihood of partial occlusion.

IV. CONCLUSIONS AND FURTHER WORK

We have shown that an ear model can be built using an unsupervised learning algorithm and a dataset of ear images. We have validated our model in a recognition task and also demonstrated its advantage in handing occlusion. Its performance benefits are not without cost: PCA can outperform our approach on unoccluded ears but its performance in noise and occlusion drops rapidly compared with the new approach.

In our further work we aim to promote the model and use it for feature selection. Our model has further potential in exploiting the mutual position, scale and orientation information of the parts. Furthermore it identifies corresponding features between ear images which can be used in a feature selection algorithm to identify the most important ear parts from a recognition perspective. Applying a complementary feature extraction technique to cover the undetected features by SIFT -for example, boundary type features- may also prove beneficial, building up a fused model.

REFERENCES

[1] A. Bertillon. *La photographie judiciaire, avec un appendice sur la classification et l'identification anthropometriques*. Gauthier-Villars, Paris, 1890.

[2] M. Burge and W. Burger. Ear biometrics. In A. Jain, R. Bolle, and S. Pankanti, editors, *BIOMETRICS: Personal Identification in a Networked Society*, pages 273–286. Kluwer Academic, 1998.

[3] K. Chang, K.W. Bowyer, S. Sarkar, and B. Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9):1160–1165, 2003.

[4] H. Chen and B. Bhanu. Human ear recognition in 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):718–737, 2007.

[5] V. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods*. John Wiley & Sons, Inc., NY, USA., 1998.

[6] J. Daugman. Biometric decision landscapes. Technical Report TR482, University of Cambridge, Computer Laboratory, 2000.

[7] G. Dorko and C. Schmid. Object class recognition using discriminative local features. Technical Report RR-5497, INRIA - Rhone-Alpes, February 2005.

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of the IEEE Conf on Computer Vision and Pattern Recognition*, pages 264–271, Madison, Wisconsin, 2003.

[9] S. Helmer and D. Lowe. Object recognition with many local features. In *In Workshop on Generative Model Based Vision 2004 (GMBV)*, Washington, D.C., July 2004.

[10] D. J. Hurley, B. Arbab-Zavar, and M. S. Nixon. The ear as a biometric. In A. Jain, P. Flynn, and A. Ross, editors, *Handbook of Biometrics*. Forthcoming 2007.

[11] D. J. Hurley, M. S. Nixon, and J. N. Carter. Force field feature extraction for ear biometrics. *Computer Vision and Image Understanding*, 98:491–512, 2005.

[12] A. Iannarelli. *Ear Identification*. Paramount Publishing Company, Freemont, California, 1989.

[13] K. Iwano, T. Hirose, E. Kamibayashi, and S. Furui. Audio-visual person authentication using speech and ear images. In *Proc. of Workshop on Multimodal User Authentication*, pages 85–90, 2003.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[15] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Proc. AVBPA*, Washington D.C., 1999.

[16] B. Moreno and A. Sanchez. On the use of outer ear images for personal identification in security applications. In *Proc. IEEE 33rd Annual Intl. Conf. on Security Technology*, pages 469–476, 1999.

[17] J. L. Northern and M. P. Downs. *Hearing in Children*. Lippincott Williams & Wilkins, 5 edition, 2002.

[18] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. of the 8th European Conference on Computer Vision*, volume 2, pages 71–84, Prague, Czech Republic, 2004.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[20] P. Yan and K. W. Bowyer. 2d and 3d ear recognition. In *Biometric Consortium Conference*, 2004.

[21] P. Yan and K. W. Bowyer. Biometric recognition using 3d ear shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1297–1308, 2007.

[22] T. Yuizono, Y. Wang, K. Satoh, and S. Nakayama. Study on individual recognition for ear images by using genetic local search. In *Proc. of the 2002 Congress on Evolutionary Computation*, pages 237–242, 2002.