A HIERARCHY PROBABILITY-BASED VISUAL FEATURES EXTRACTION METHOD FOR SPEECHREADING[1]

Yanjun Xu, Limin Du, Guoqiang Li, and Ziqiang Hou

Institute of Acoustics, CAS, Beijing 100080
Email: yanjunxu@Eudoramail.com

ABSTRACT

Visual feature extraction method now becomes the key technique in automatic speechreading systems. However it still remains a difficult problem due to large inter-person and intra-person appearance variabilities. In this paper, we extend the normal active shape model to a hierarchy probability-based framework, which can model a complex shape, such as human face. It decomposes the complex shape into two layers: the global shape including the position, scale and rotation of local shapes (such as eyes, nose, mouth and chin); the local simple shape in normal form. The two layers describe the global variation and local variation respectively, and are combined into a probability framework. It can perform fully automatic facial features locating in speechreading, or face recognition.

1. INTRODUCTION

Automatic speechreading systems, through their use of visual information to supplement acoustic information, have been shown to yield better recognition performance than purely acoustic systems, especially when environment noise is present [1][2]. As the purely acoustic speech recognition techniques have been deeply studied, visual feature extraction methods with high accuracy and robustness now became the key technique in automatic speechreading systems. However visual speech feature extraction still remains a difficult problem due to large inter-person and intra-person appearance variabilities, most current systems simplified or even avoided this issue. Generally, visual speech feature extraction methods include two categories: the image based and the model based. The image based method is simple, easy to perform and robust to small audiovisual speech database under well-controlled illumination. But it has the disadvantage of high feature dimensionality and low robustness under loose-controlled illumination. While the model based method only extracts structural features and is more invariant and robust to environment variability, so it has the advantage of low feature dimensionality and high performance, also the image based method can benefit from it through its locating accurate area of interest. So the model based method is attracting more and more attentions and may be the final solution.

By now, many visual feature extraction algorithms are put forward and utilized in audiovisual speech processing and face recognition. Some only obtain simple facial features, such as eyes or lips, with deformable templates and models; some obtain the loose positions of all facial features with low-level image processing techniques; a few combine above two methods in specific framework to form full and accurate facial description. In fact, even the third method has the disadvantage of decomposing the whole problem into independent and manageable components. This may simplify the complex task of the facial features detection and localization, however, it may result in unidirectional and irreversible flow of error from one component to the next.

Deformable templates or models have played a great role in locating simple contour or shape in complex scene. T. F. Cootes et al. [3] developed an active shape model (ASM), which modeled shape deformation by a point distribution model (PDM). J. Luettin et al. [4][5] gave an extension of the method, the appearance based model (ABM), and used it in lip localization and tracking successfully. When these methods are

[1] This research is supported by the President Foundation of the Institute of Acoustics, Chinese Academy of Sciences (No.98-02) and "863" High Tech R&D Project of China (No. 863-306-ZD-11-1).

used in fully facial features labeling, great difficulty is met. The whole face is so complex a shape, which includes eyes, nose, mouth, and chin, that a large group of parameters are needed to control the shape deformation. Also these parameters may have large covariance. This makes the energy minimization problem very hard to solve.

In this paper, we extend the normal active shape model into a hierarchy probability-based framework, which can model a complex shape model, such as human face. It decomposes the complex shape into two layers: the global shape which takes the position, scale and rotation of local shapes (such as eyes, nose, mouth and chin) as its descriptors; the local shape which takes the salient feature points as its descriptors. In this way, we can model the global variation and local variation respectively and obtain a more reasonable model. It can perform fully automatic facial features locating in speechreading, or face recognition.

2. THE HIERARCHY PROBABILITY-BASED MODEL OF HUMAN FACE

In general Active Shape Models or Active Contour Models (snakes), a contour (or a shape) is assumed to be an ordered set of points with salient image features, assuming number of points is N_s, the shape vector is defined by

$$\mathbf{V} = [v_1 \quad v_2 \quad \cdots \quad v_{N_s}]^T \tag{1}$$

where v_i is the position vector of each shape point. We can use it to model eyes, nose, mouth, and chin. It is called a simple shape.

In order to perform training, some normalization should be done. First we select two reference points on each simple shape, such as two corners of two eyes, two terminals of nose bottom, two corners of mouth, two terminals of chin contour. The the distance between the two points is defined as the scale factor of the simple shape, and its rotation θ as the orientation of the simple shape, and its center v_c as the origin of the simple shape. The other points are distributed in equal distance. Then the normalized shape vector is

$$\mathbf{x} = M\left(\frac{1}{s}, -\theta\right)[\mathbf{v} - \mathbf{v}_c] \tag{2}$$

Where $\mathbf{V}_c$ is translation vector.

$$\mathbf{V}_c = [v_c \quad v_c \quad \cdots \quad v_c]^T \tag{3}$$

M is rotation transform matrix:

$$M(s,\theta) = \begin{pmatrix} s\cdot\cos\theta & -s\cdot\sin\theta \\ s\cdot\sin\theta & s\cdot\cos\theta \end{pmatrix} \tag{4}$$

Given N shapes, the mean shape vector $\bar{\mathbf{x}}$ can be obtained as follows:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{5}$$

and the shape covariance matrix $\mathbf{S}_s$ can be calculated:

$$\mathbf{S}_s = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{6}$$

The eigenvectors and eigenvalues of the shape covariance matrix may be obtained through performing Principal Components Analysis.

$$\mathbf{S}_s\mathbf{p}_{sk} = \lambda_{sk}\mathbf{p}_{sk} \tag{7}$$

where λ_{sk}, k=1,…,$2N_s$, $\lambda_{sk} \geq \lambda_{s(k+1)}$, are eigenvalues in descending order.

The eigenvectors with the largest eigenvalues describe the most significant modes of variation, i.e. the variance of each eigenvector is equal to its corresponding eigenvalue. Most of shape variation can be represented by a small number of modes (the so-called principal modes). One way to identify the number of principal modes T_s is to choose the smallest number of modes such that the sum of their variances explains a sufficiently large proportion of λ_T, the total variance of all the variables, where

$$\lambda_T = \sum_{k=1}^{2N_s}\lambda_k \tag{8}$$

then a normalized shape can be approximated by

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{p}_s\mathbf{b}_s \tag{9}$$

where $\mathbf{P}_s = (p_{s1} \quad p_{s2} \quad \cdots \quad p_{sT_s})$ is the matrix of the first T_s eigenvectors, $\mathbf{b}_s = (b_{s1} \quad b_{s2} \quad \cdots \quad b_{sT_s})$ is a

vector of weights.

J. Luettin constructed a global profile vector for image search [5]. As the grey-level is not illumination invariant, such methods are sensitive to illumination variability. Because shiftable multiscale decompositions reflect the characterization of the image structures, which are illumination-invariant, the wavelets transform based method is more robust. Experiments also support the validity of the scheme. we use a global shiftable wavelets decomposition vector instead. E.P.Simonceli et al. [6] [7] have defind a type of shiftable wavelet transforms which are stable under translations and orientations. In our case, we use a 2-D shiftable transform. We utilize a 4-orientation filter bank and perform 3 scales decomposition on the original image. After creating a pyramid decomposition, we obtained N_p channels output. Then a N_p-dimention multiscale vector if generated at each shape point. A global multiscale vector is constructed by concatenating the muliscale vectors at all shape points as follows:

$$\mathbf{y} = \begin{pmatrix} y_1^T & y_2^T & \cdots & y_{N_s}^T \end{pmatrix}^T \tag{10}$$

Similarly, After calculating the mean global multiscale vector and its covariance matrix, PCA is then performed to get the eigenvalues and eigenvectors of the covariance matrix. Any global multiscale vector can be approximated using

$$\mathbf{y} = \bar{\mathbf{y}} + \mathbf{P}_f \mathbf{b}_f \tag{11}$$

where $\mathbf{P}_f = \begin{pmatrix} p_{f1} & p_{f2} & \cdots & p_{fT_f} \end{pmatrix}$ is the matrix of the first T_f eigenvectors , $\mathbf{b}_f = \begin{pmatrix} b_{f1} & b_{f2} & \cdots & b_{fT_f} \end{pmatrix}$ is a vector of weights.

In this way, we can model the main variation of every facial feature (eyes, nose, mouth, chin). In order to model the global shape variation, we then construct a high-level shape vector from the position, scale, orientation of all simple shapes. The global shape of mouth $\mathbf{V}_{gmouth}$ can be expressed as follows:

$$\mathbf{V}_{gmouth} = \begin{bmatrix} s_{mouth} & \theta_{mouth} & x_{cmouth} & y_{cmouth} \end{bmatrix}^T \tag{12}$$

And $\mathbf{V}_{gleye}$ for left eye, $\mathbf{V}_{greye}$ for right eye, $\mathbf{V}_{gnose}$ for nose, $\mathbf{V}_{gchin}$ for chin, in the same way. Then a global shape vector of a human face can be written as follows:

$$\mathbf{V}_g = \begin{bmatrix} \mathbf{V}_{gleye}^T & \mathbf{V}_{greye}^T & \mathbf{V}_{gnose}^T & \mathbf{V}_{gmouth}^T & \mathbf{V}_{gchin}^T \end{bmatrix}^T \tag{13}$$

Similarly, after normalizing the global shape vector and calculating the mean values and the covariance matrix, PCA is performed to get the eigenvalues and eigenvectors of the covariance matrix. Then any global face shape vector can be approximated using the sum of the mean vector and the weighted variations represented by the principal modes.

According to the probability theory, we can write:

$$P_{face} = P_{global} \cdot P_{leye} \cdot P_{reye} \cdot P_{nose} \cdot P_{mouth} \cdot P_{chin} \tag{14}$$

Because the model is trained by learning a great quantity of data, we can assume it conform to Gauss distribution, then the product of probability items is converted into the sum of energy items of global shape vector and all simple shapes:

$$E_{face} = E_{global} + E_{leye} + E_{reye} + E_{nose} + E_{mouth} + E_{chin} \tag{15}$$

The energy of the global shape vector is only the internal energy (deviation from the mean shape), while the energy of the local simple shape vector is the weighted sum of the internal energy and the external energy (deviation form the mean evidence).

Then the visual feature extraction problem is turned into an energy minimization problem.

3. SEGMENTAL ENERGY MINIMIZATION METHOD

We adopt a coarse-to-fine strategy in the image searching. First we perform an energy minimization procedure on the smallest scale, then perform the larger scale after obtaining the best match, and finally get the most accurate face and facial feature localization.

Energy minimization methods (EMM) are fundamental techniques in the field of computer vision and pattern recognition. Although the global minimum of the energy function is the only target, the energy function often has many local minima. The degree of difficulty for finding the global minima depends largely on the constraint on the images, under which the algorithm should work.

The Downhill Simplex Method (DSM) is one of the most important EMMs, which has the important property of not requiring derivatives of function evaluations and that it can get out of local minima. It has been used in many visual feature

extraction algorithms. A DSM on n-Dimension energy function includes following steps：

1) Generate a simplex with n+1 vertices in the n-dimension space, and evaluate the energy function at each vertex;
2) Find the worst point, the less worst point, the best point in the simplex;
3) Calculate the reflection point of the worst vertex in the super plane defined by the remaining vertices;
4) Generate new vertices according to some specific rules;

Repeat（2）~（4） until the distance between the vertices is smaller than the specific precision.

However, We found that the minimization procedure of DSM always undergoes first the larger variance modes then the smaller variance modes. This inspired us to divide the parameters vector into several segments, perform the DSM minimization procedure sequentially on each segment. As a result, the convergence speed is greatly accelerated and the performance of localization is more robust, while the localization accuracy remains excellent.

4. DISCUSSION AND FUTURE WORK

In this paper, we extend the normal active shape model to a hierarchy probability-based framework, which can model a complex shape model, such as human face. It decomposes the complex shape into two layers: the global shape including the position, scale and rotation of local shapes (such as eyes, nose, mouth and chin); the local simple shape in normal form. The two layers describe the global variation and local variation respectively, and are combined into a probability framework. It can perform fully automatic facial features locating in speechreading, or face recognition.

We have designed and created a Chinese Audiovisual Bimodal Speech Database CAVSR1.0. It consists of 78 isolated Chinese characters spoken by 20 subjects (12 male, 8 female) in 2 repetitions. There are total 99840 image frames taken under loose-controlled illumination. We want to label a small quantity of data by hand, then train the hierarchy face model with the labeled data. Finally, we use the trained model to perform fully automatic facial features extraction. In this way, the audiovisual speech database may be enlarged (number of subjects, speech data, and repetitions) easily, and speechreading may be studied on a larger corpus. Also other vision research like face recognition may benefit from it.

5. REFERENCES

1. E.D.Petajan, *Automatic Lipreading to Enhance Speech Recognition*. Ph.D thesis, University of Illinois at Urbana-Champain, 1984.
2. P.L. Silsbee, Computer Lipreading for Improved Accuracy in Automatic Speech Recognition, *IEEE Trans. On Speech and Audio Processing*, Vol.4, No. 5, 1996.
3. T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, Active Shape Models—Their Training and Application, *CVIU: Computer Vision and Image Understanding,* Vol. 61, No. 1, January, pp. 38-59, 1995.
4. J. Luettin, N.A. Thacker, and S.W. Beet, Speechreading using shape and intensity information, In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, vol. 1, pp. 137-140, 1996.
5. J. Luettin, *Visual Speech and Speaker Recognition*. Ph.D thesis, University of Sheffield, 1997.
6. E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger, Shiftable Multiscale Transforms, *IEEE Trans. On Information Theory*, Vol.38, No.2, 1992.
7. W.T. Freeman and E.H. Adelson, The Design and Use of Steerable Filters, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol. 13, No.9, 1991.