

Texture-Constrained Shape Prediction for Mouth Contour Extraction and its State Estimation

Zhaorong LI, Haizhou AI

Computer Science and Technology Department, Tsinghua University, Beijing 100084, China

E-mail: ahz@mail.tsinghua.edu.cn

Abstract

In this paper, we present an automatic mouth contour and state estimation system. An efficient mouth contour extraction algorithm is proposed under the framework of Active Shape Model (ASM). Considering large mouth shape variations, we propose a texture-constrained shape prediction method for initialization. To improve accuracy and robustness of classical ASM, we use classifiers trained by Real AdaBoost to characterize the local texture model. This model is proved to have much stronger discriminative power than Gaussian model of classical ASM. After extracting the mouth contour, the mouth is classified into one of 4 typical states by Support Vector Machine (SVM) based on the shape parameter. Experiments over a large set show that extracted mouth contours have achieved good accuracy, with an average 89.5% acceptable rate, and the mouth state estimation reaches an average 93% correct rate. This automatic system reaches a speed of about 10 frames per second on a Pentium-IV 1.7GHz PC, which may have potential applications in visual speech recognition etc.

1. Introduction

Within the past decade or two, face information processing has become a hotspot in computer vision and pattern recognition area. As main components of face, face organs (eye, mouth etc) and their contours and states (close/open etc), containing rich information of facial expression and human emotion, have an important research value in applications such as driver's fatigue detection and visual speech recognition. In this paper, we only focus on mouth contour extraction and its related state estimation (close /naturally open/"o" shape / widely open) corresponding to 4 vowels' pronunciation. It can be extended to eye contour and state estimation similarly.

Active Shape Model [1] is a powerful tool in regular object contour extraction. It uses a point

distribution model to parameterize a face shape with PCA method as $S = (\bar{S} + U \cdot s)$ and interpret a facial image in a 2-stage iterative algorithm: 1) a local search is performed on each model point to find a candidate neighbor which best fits the local texture model of the landmark; 2) an optimization of shape parameters is conducted in the global shape subspace to best interpret these newly found positions. This method has strong flexibility and has been proved to be very effective in face alignment area.

Being an iterative method, ASM relies on the initialization of shape and suffers the common local minima problem, especially when the object to alignment has a vast variation of shape. Take mouth for example, with the mean shape of mouth (approximately half open) as an initialization it is hard to converge for close mouth and widely open mouth. There are two approaches to this problem: try to make a better initialization or make improvements on local texture model [2] [3] [4] to enhance local search ability. However, such strong models may result in high computation complexity, thus unsuitable for mouth contour extraction with target applications in embedded systems such as mobile devices.

To get across this obstacle, we employ two techniques. First, to replace the mean shape initialization in classical ASM, we use a texture-constrained shape prediction method to perform initialization. Second, we characterize the local texture model with classifiers learnt from training sets by Real AdaBoost [5] using rectangle features [6], which has been proved to have two merits in face detection area, very strong discriminative power and computation efficiency. The two merits are just what we want in our system.

The rest of this paper is organized as follows: Section 2 presents an overview of our automatic mouth contour extraction and state estimation system; Section 3 introduces the texture-constrained shape prediction method for shape initialization. Section 4 describes our local texture model characterized with Adaboosted

classifiers. Section 5 presents our mouth state estimation using SVM. Section 6 reports the experiment results. And finally a conclusion is given in section 7.

2. System overview

The proposed system consists of four modules, namely face detection, facial feature points (four eye corners and two mouth corners) localization, mouth contour extraction, and mouth state estimation, as illustrated in Fig. 1. Mouth states are divided into four categories that include close, naturally open, “o” shape, and widely open, as shown in Fig. 2. In this paper, only mouth contour extraction and mouth state estimation module is discussed in detail. For face detection, a variation of the cascade detector proposed by Viola and Jones [6] is used. For facial feature points localization, we use a SDAM method proposed in [7].

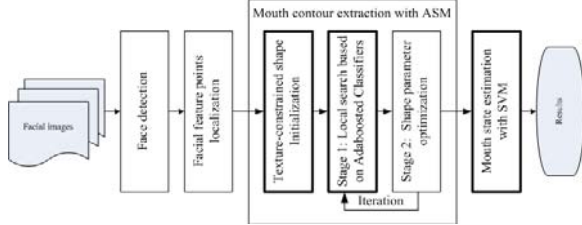


Figure 1. An automatic mouth contour extraction system. The emphasized blocks are discussed in detail.



Figure 2. Four categories of mouth state

3. Texture constrained shape prediction for initialization

As mentioned above, a problem for mouth contour extraction is that the classical ASM will run into local minima when the shape of mouth to be extracted is far from mean mouth shape. An obvious clue is that the texture around mouth region contains much information of the shape of mouth, so we can utilize this. A simple linear regression from texture to shape is used. Then, we utilize this texture constrained shape to make a better initialization than the mean shape.

We derive a shape from texture around mouth region given two mouth corners in the following way: First, we cut an intensity patch around the foursquare region containing mouth; Then, we represent it with Haar-wavelet transformation and get a feature t ; At last, we get a texture-constrained shape parameter

namely s_t from a linear regression $s_t = R \cdot t$. This regression matrix R can be easily learnt from training data. This process is illustrated in Fig. 3.

After getting this texture-constrained shape, we employ it to make a better initialization rather than the mean mouth shape. The only problem left is to estimate pose parameters $(tx, ty, \theta, scale)$ via two given mouth corners $p_1(x_1, y_1), p_2(x_2, y_2)$. We can simply estimate them as follows:

$$tx = \frac{x_1 + x_2}{2}, ty = \frac{y_1 + y_2}{2},$$

$$\theta = \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right),$$

$$scale = \frac{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}{width(S_t)}$$

where S_t is the texture-constrained shape data:

$$S_t = (\bar{S} + U \cdot s_t)$$

The accuracy of initialization via the texture-constrained shape is demonstrated in Fig. 4.

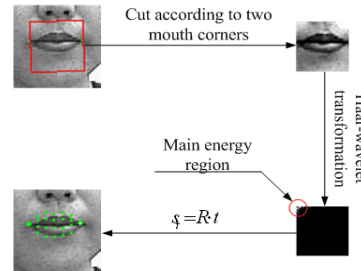
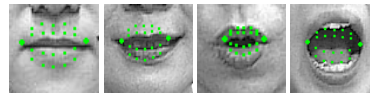
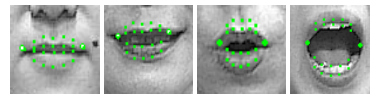


Figure 3. The process of deriving a shape from texture



a) Initialization with mean shape



b) Initialization with the texture-constrained shape

Figure 4. A comparison between initialization with mean shape and initialization with the texture-constrained shape.

4. Characterize local texture model with AdaBoosted classifiers

Another main reason for classical ASM running into local minima is that the classical ASM

characterizes its 1-D intensity profiles perpendicular to each landmark contour with a Gaussian distribution model. Because of illumination variations and other factors, this simple model can not distinguish feature points from its neighbors very well, which will guide the local search to find false candidate positions. This inadequacy activates us to find a strong classification method to characterize local textures, which not only consider positive samples (feature points), but also negative samples (neighbors of feature points). It is well known that Boosting and Support Vector Machine (SVM) are two popular classification framework, and Boosting outperform SVM in computation efficiency. Here, we chose Boosting method for efficiency. We train a classifier by AdaBoost based on 2-D Haar-like rectangle features to model the local texture of each landmark by the following method [8]:

- (1) Collect several 24x24 intensity patches centered inside a 3x3 region with its center at the ground-truth landmark position from every training image as positive training samples.
- (2) Collect several 24x24 intensity patches centered inside a 12x12 region but outside the above mentioned 3x3 region with its center at the ground-truth landmark position from every training image as negative training samples.
- (3) Train on these samples to get a classifier with Real AdaBoost using Haar-like rectangle features

The Real AdaBoost selects a small number of weak classifiers $h_i(x)$ from a large weak classifier pool to

construct a strong classifier $H(x) = \sum_{i=1}^T h_i(x) - b$.

This output indicates the confidence which class a sample belongs to. The larger it is, the more probable it is a positive sample. Based on the classifier, the position with the largest confidence is chosen to be the candidate position for the next ASM iteration.

An experiment is conducted to illustrate the classification power of our method. Take the upper feature point of upper mouth lip for instance, a comparison between classical Gaussian model, SVM-based classifier, and Real AdaBoost-based classifier is show in Table 1. It can be seen that Real AdaBoost has the strongest classification power and its processing time is comparable to that of Gaussian Model.

Table 1. A comparison of discriminative power between three local texture modeling methods: Gaussian model, SVM-based classifier, and Adaboost-based classifier.

	average classification correct rate	average processing time per intensity patch
Gaussian Model	0.783	0.01ms
SVM	0.965	4.2ms
Real AdaBoost	0.978	0.03ms

5. Mouth state estimation

The mouth state estimation is performed by applying Support Vector Machine (SVM) on the shape parameter. In the training stage, we first calculate the shape parameters of four categories of mouth in training set as training data of SVM; Then we label four categories of mouth: close, naturally open, “o” shape, and widely open as 0, 1, 2 and 3 respectively; At last, we employ SVM to train a four class classifier with RBF-kernel using one-against-one multi-class training method. In the testing stage, after extracting the mouth contour and obtaining the shape parameter, the mouth state is classified into one of the four categories with this classifier.

6. Experiments

Experiments have been conducted on a large data set consisting of 1,600 front facial images including four categories of mouth of nearly equal quantity. The image size is 240x320. We randomly select 1,200 images for training, and the rest 400 images for testing. The algorithm is also tested for real-time application on a desktop video from a web camera.

The accuracy is measured with relative pt-pt error, which is the point-to-point distance between the extraction result and the ground-truth data divided by the distance of two eyes.

The distributions of the overall average error of four categories of mouth are shown in Fig. 5. It shows that the close mouth is best extracted while the widely open mouth is worst. For each image, supposing if the maximum relative error of all landmarks is below 6%, the mouth contour can be considered to be acceptable. We make a comparison on the mouth contour acceptable rate in Table. 2. For those images with acceptable contour extraction, SVM-based classifier is used for mouth state estimation. The estimation correct rate is also listed in Table. 2.

The average execution time of one image with our system is about 90 ms. All of tests are carried out on an Intel Pentium-IV 1.7G PC. When tested with a web camera, our system can reach a speed of about 10 frames per second.

Some results are shown in Fig .6 and Fig.7.

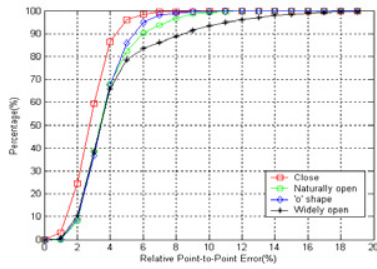


Figure 5. Distribution of relative average pt-pt error of four categories of mouth

Table 2. Mouth contour acceptable rate and state estimation correct rate

Mouth State	Close	Naturally open	“o” shape	Widely open
Classical ASM	0.84	0.72	0.74	0.74
Our Method	0.99	0.81	0.94	0.84
State Estimation	0.98	0.90	0.86	0.98

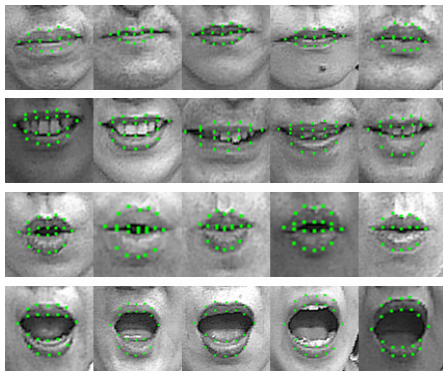


Figure 6. Some experimental results of mouth contour extraction in test set



Figure 7. Some results with a web camera (results of eye contour extraction are also shown)

7. Conclusion

In this paper, we have proposed a mouth contour extraction and mouth state estimation method. Under the framework of ASM, in order to deal with the local minima problem, we propose a texture-constrained shape prediction method for initialization, and employ AdaBoosted classifiers to characterize the local texture models. After extracting mouth contour, the mouth state estimation is performed by applying SVM on the shape parameter. Experimental results on a database of 1600 frontal face images show that our method has achieved good robustness and accuracy. The average extraction rate is about 89.5% and the average state estimation rate is around 93%. Our automatic mouth contour extraction and state estimation system reaches a speed of 10 frames per second with a webcam, which may have potential applications in driver’s fatigue detection and visual speech recognition.

8. Acknowledgement

This work is supported by National Science Foundation of China under grant No.60332010.

9. References

- [1] T Coots, D Cooper, et al, “Active shape models-their training and applications”, *Computer Vision and Image Understanding*, 1995, 61(1): pp. 1-10.
- [2] Feng Jiao, Stan Li, et. al, “Face alignment Using Statistical Models and Wavelet Feature”, *Proceedings of IEEE Conference on CVPR*, 2003, pp.321-327.
- [3] Shuicheng Yan, Mingjing Li, et al, “Ranking prior likelihood distributions for Bayesian shape localization framework”, *Proceedings of IEEE Conference on ICCV*, 2003, pp.51-58.
- [4] Ce Liu, Heung-Yeung Shum, and Changshui Zhang, “Hierarchical Shape Modeling for Automatic Face localization”, *Proceedings of ECCV*, 2002, pp.687-703.
- [5] R.E. Schapire and Y. Singer, “Improved Boosting Algorithms Using Confidence-rated Predictions”, *Machine Learning*, 1999, pp.297-336
- [6] P. Viola and M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features”, *Proceedings of IEEE Conference on CVPR*, 2001, pp.511-518
- [7] Tong WANG, Haizhou AI and Gaofeng HUANG, “A Two-Stage Approach to Automatic Face Alignment”, *Proceedings of SPIE*, Vol.5286, 2003, 558-563
- [8] Li ZHANG, Haizhou AI, et al, “Robust Face Alignment Based on Local Texture Classifiers”, *The IEEE International Conference on Image Processing*, 2005.