Color Features Based Speaking Detection with Hidden Markov Model in Video Sequences

Peilin Jiang[1,2], Ran Li[2], Fuji Ren[1], Shingo Kuroiwa[1], and Nanning Zheng[2]

[1] The University of Tokushima, Tokushima, Japan,
jiang@is.tokushima-u.ac.jp,
[2] Xi'an Jiaotong University, Xi'an, 710049, China

Abstract. The Human Computer Interface Technology has faced challenges of understanding user's mind actively. In the first, the speak detection is a primary technique in applications of human computer interface (HCI) and other applications like surveillance system, video conference and multimedia data base management in computer vision and speech recognition. This paper describes a novel method to detect speaker with a probabilistic model of behavior of speaking. After human face recognition, the especial components under the nonlinear transformation in color space of lip represent the specific mouth region and then combine the groups of coherent motions . Next the simple movements in the mouth region are modeled by hidden Markov models. The experimental results demonstrate that the model representing speaking is efficiency and successful in applying to driver video surveillance system.

1 Introduction

The speak detection is one of the demanding tasks in the field of computer vision and speech recognition. It has played an important role in real-time and affective computing in HCI gradually before the recognition and synthesis procedures [2, 12]. For an instance, in the video conference, several participants discussed the problems together and the video camera needs to switch among the different speakers automatically. Thus, the recognition of speaker is a primary work before video data processing and transforming start. The same problem appears before affective computing in the human computer interface in which the speech and the expression is the important ground to determine the situation of emotion states transitions. The work has been firstly focused on the work of modeling the detailed mouth movement for lip-reading [13]. The common approach to speak detection is calculating movement energy function and then judging with an empirical threshold. Also, the information from audio is used too [6, 7]. However, the first problem, is that the threshold is empirical and dependent on individuality of different person and the second one is some video sequences which we need to process is asynchronous with audio sequences. To solve these problems, the previous works include template-based match, video-audio information fusion and supervised statistical learning models methods have been applied respectively. In this paper, an unsupervised human motion learning method, which do not

Advances in Artificial Intelligence and Applications
Research in Computer Science 32, 2007, pp. 374–381

Received 16/06/07
Accepted 31/08/07
Final version 10/09/07

need mark training data ahead of time, is introduced to describe the behavior of speaking. The method estimates the motion as representation of a sequence of observations instead.

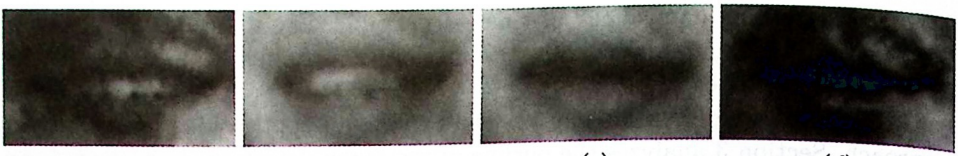
In Section 2 we discuss previous relative work and applications related to our approach. Section 3 analyzes the outline of the speck detection. In section 4, we describe the model of speaking and speak recognition, which includes prepare works before detection, color nonlinear map, and hidden Markov model learning algorithm. In section 5, we introduce the experiments of training the model and recognizing the speak behavior.

2 Related Works

Previously, the unsupervised learning approach for analysis of human motion was proposed by Tianshu Wang [3] to classify the uniform action from motion video sequences. Most earlier before that the research [14] had proved that the HMM model can be use to effectively recognize the human motions. On the other side, lip localization techniques also conclude a lot of results. Among them, an algorithm constructed mouth, eyes mapping based on light compensation and a nonlinear color transformation has been identified effective [4]. Additionally, several efforts are presented in speaker detection in many applications [1, 5, 8, 9].

3 Question Analysis

In the section 1, the problem is proposed that the description of speaking which we want to exploit is ambiguity because of noises, occlusion, random varieties from difference bodies. Hence, to a certain extent formerly mentioned methods are confused. The template-based method is more adaptive the localization problem than speak detection; the video-audio fusion method does not working well except the audio channel; the supervised statistical learning algorithms are effective but the data labeling is seriously heavy and hard to keep consistent. The traditional segment of mouth motion is in two phases which is showed in Fig.1. The open mouth and closed mouth respectively. Unfortunately, the borderline between these two phases is vague. Considering the speaking as the human complex dynamic such as walking and waving, we can decompose it with the probabilistic compositional framework[14]. Firstly, the work is divided into two steps: the mouth localization and speaking detection. In the localization step, the color feature in human face is extracted because of its unique property. Next, high-level complex speak motion is symbolized by Hidden Markov Model as successive phases of simple movements. Namely, the speaking is represented by a HMM model. And the HMM model is composed by the states which are observed as relatively simple dynamic models. These models are organized by temporal sequences of coherent motions that are described by low-level features. These underlying features are extracted from a sequence of input frames in which the color value of each pixel is random variable. Given a sequence of images $I_1, I_2 \ldots I_t$ includes HMM for motion, we can perform the recognition.



(a) (b) (c) (d)

Fig. 1. Examples of mouth states. The images in (a)and (b) are open mouth, (c) and (d) are close mouth.

$$P = P(HMM|I_1, I_2, \ldots, I_t) \tag{1}$$

The equation (1) presents the probability of HMMs in the image sequence. Several motions are corresponding to different HMMs.

4 Speak Model and Detection

Specifically, mouth speaking motion is represented here by a HMM model. Therefore, the detection task is converted into building the speak model and recognizing the HMM. The former is sum to calculate the optimal parameters of the hidden Markov model and then the second is estimation problem of optimal states sequences. In the section, we will introduce the processing of detection speaking with the HMM model. The first subsection describes the works before building speak model. In subsection 4.2, we describe a nonlinear transformation color feature being used as observations for coherent motion. Subsection 4.3 introduces the HMM learn model and detecting process.

4.1 Prepare Works

Before speak detection we must accurately localize the region of mouth in face area. In practical, the face detection algorithm is operated in the first step. In our work, we quote the cascade Adaboost based face detection algorithm [11] to execute the detection task. Successively the next phase is mouth localization. There are several fast and accurate algorithms, in which include color-based, texture-based, template-based and statistical learning. Compared with them we chose the luminance-chrominance space based mouth mapping for its successful stability under a wide arrange of conditions. After a light compensation, a novel nonlinear transformation is formulated to extract chrominance components which are sensitive on lip-tone region and skin-tone color. The details are introduced below.

4.2 Mouth Mapping based on Nonlinear Color Transformation

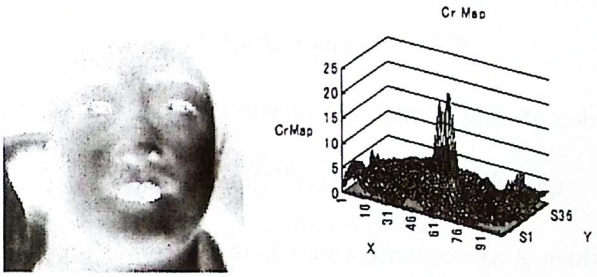
Modeling the human skin color requires firstly choosing an appropriate color space which cluster skin color together obviously. Among them, YC_bC_r space is

adopted because it is perceptually uniform and widely used in video compression standard and separable between components of luminance and chrominance [4]. The mouth region is special for its strong red color component and weak blue color component than other skin regions. These two factors correspond to the two components C_r and C_b respectively. The mouth mapping is then constructed nonlinearly:

$$Threshold_{mouth} = C_r^2 \cdot \left(C_r^2 - \eta \cdot \frac{C_r}{C_b} \right)^2 \tag{2}$$

$$\eta = 0.95 \cdot \frac{\frac{1}{n} \cdot \sum_{(x,y) \in \Im\Re} C_r(x,y)^2}{\frac{1}{n} \sum_{(x,y) \in \Im\Re} C_r(x,y)/C_b(x,y)} \tag{3}$$

The C_r/C_b feature and C_r^2 feature are found that low and high responses to the lip region in face respectively and both of them are normalized. n is the number of pixel within the face region. The h is estimated as a ratio of the average C_r^2 to average C_r/C_b The Fig.2 is the mouth map for the sample face.



(a) (b)

Fig. 2. Extracting mouth region by YCrCb map. The region with higher CrMap value in (b) is considered the mouth map.

4.3 Hidden Markov Model Learning

The hidden Markov model is statistically finite states machine with two sets of probabilities: a transition probability and an emission probability. For example, a HMM can be represented by a topology like in Fig. 3.

There are three basic problems of interest in HMM [15]. As the estimation problem, the parameters of hidden Markov model can be evaluated by EM algorithm [10]. In a probabilistic compositional framework, our goal is to classify the speak motion composed by relatively simple movements from the video sequence. We assume that the movements is sequential and then HMM can be modeled

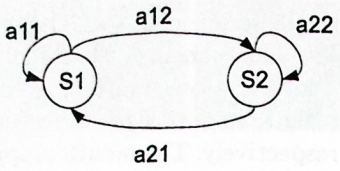


Fig. 3. Example of the two states HMM used for recognition

the motion. Each state corresponds to one simple phase. The EM algorithm provides a recursive estimation to converge on some limit of model parameters that achieve the local maximum.

In the E step, the Q function is calculated as following:

$$Q(\theta|\theta^{(t)}) = \sum_{h} P(h|v, \theta^{(t)}) logP(v, h|\theta)) \tag{4}$$

Where v is the visual part of the data and h is the hidden one.

In the M step, the maximum process is as below:

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^{t}) \tag{5}$$

In the case of hidden Markov model, the $Q(\theta|\theta^{(t)})$ is:

$$Q(\theta|\theta^{(t)}) = \sum_{s} P(s|x, \theta^{(t)}) logP(x, s|\theta)) \tag{6}$$

Where s is the hidden state sequence and x is the observation sequence.

4.4 Model Recognition

In decoding problem, we aim to recover the sequence given an observation sequence. The Viterbi algorithm provides fast way of solving the decoding problem. The followings stare the Viterbi algorithm steps:

- Initialization: $\delta_1(i) = \pi_i b_i(x_1)$ and $\psi_1(i) = 0$ for all the states;
- Recursion: for $i = 2$ to T, j represents all states, $\delta_t = \max_{(1,N)}[\delta_t(i)]$, $\psi_t(j) = \arg\max_{(1,N)}[\delta_{t-1}(i)a_{ij}]$;
- Termination: $P = \max_{(1,N)}[\delta_T(i)]$ and $S = argmax_{(1,N)}[\delta_T(i)]$;
- Recover states:$S_t = \psi_{t+1}(S_{t+1})$, where $t = T - 1$ to 1;

However, although that for Viterbi only guarantees the maximum of over all state sequences and the resultant probability is only an approximation. The previous research shows that this is mostly sufficient. In here, we utilize this method as model recognition approach.

5 Experiments

In this paper, we build driver abnormal action surveillance system in order to detect speak motion to demonstrate the feasibility of model and recognition approach.

Our training data is 5 video sequences of 3 different individuals driving during when they make cell-phone call, which is considered to be dangerous, namely abnormal action during normal driving. An independent test data of a sequence also is ready for recognition experiment. Training data sequences has not been labeled and the chrominance component based motion features are used to localize the mouth firstly and then represent the motion features. The training and test data are all recorded in a steady background in either running car or stationary lab. Fig.4 shows some examples of one experiment video sequence.

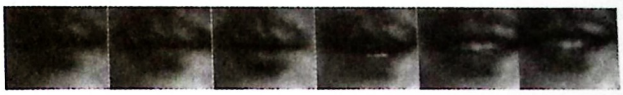


Fig. 4. Example frames of test data in training experiment.

5.1 Training Model

The training experiment is to apply with the test data that are not labeled which state the piecewise belongs to. We have tested difference numbers of states and conclude that 2 states is a reasonable opinion. We set the initial transition probability matrix and initial states randomly and train several times in order to make iterative learning procedure not converge into a local optimum.

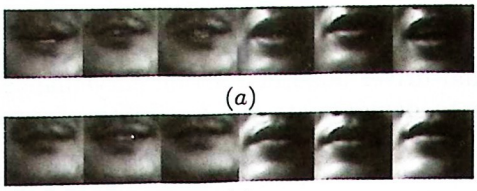
5.2 Recognizing the Speaking in Image Sequences

The final process was done to apply the learned HMM for recognition task on the test sequence. In term of the model and an unlabeled input sequence, we process the recursive decoding to obtain the optical states sequences. The sequence is split into pieces and the piecewise with high likelihood is the winner. In the result, the speak segment in test sequence is correctly recognized. The recognized segment is below in Fig.5:

In the next phase, we are going to performing more experiments on a large database.

6 Conclusion and Future Work

The human behavior like speaking, waving can be modeled by a probabilistic state transition model empirically. Like the other complex movements and gestures of human being such as walking and skating, the speaking can be clustered



(a)

(b)

Fig. 5. The example of detected speak video sequences. The upper row images (a)are segmented as speaking results, the bottom row images (b) are from no-speaking results

into certain sequent states corresponding to the observation sequences composed by low-level visual features and can be recognized separately. We conducted the training and testing experiments on video sequences of simulated driver action detection with the HMM and derived a satisfying result. Additionally, in a more general case, the observations in mouth region are a variant vector under the different poses of head and the improvement on the feature extraction and model parameter selection will be applied in the future work.

References

1. James M. Rehg, Kevin P. Murphy, Paul W. Fieguth: Vision-Based Speaker Detection Using Bayesian Networks. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99) - Volume 2 (1999)
2. Todor Ganchev, Nikos Fakotakis, George Kokkinakis: One-Speaker Detection-Limited Data: The WCL-1 System. NIST2003 Speaker Recognition Workshop, MD, USA (2003)
3. Tianshu Wang, Nanning Zheng, Yingqing Xu and Heung Yeung Shum: Unsupervised Clustering Analysis of Human Motion. Journal of Software (2003) 209-214
4. Rein-Lien Hsu, Mohamed Abdel-Mottaleb, Aril K.Jain: Face Detection in Color Images. IEEE Trans. Pattern Analysis and Machine Intelligence, vol.24, no.5, (2002) 696-706
5. Vladimir Pavlovic, Ashutosh Garg, James M. Rehg, and Thomas S.Huang: Multimodal Speaker Detection using Error Feedback Dynamic Bayesian Networks. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2000) 34 - 41
6. Ross Cutler and Larry Davis: Look who's talking: Speaker detection using video and audio correlation. IEEE International Conference on Multimedia and Expo (ICME), Manhattan, New York (2000)1589 - 1592
7. Nicolas Scheffer, Jean-Francois Bonastre: Speaker Detection using Acoustic Event Sequences. INTERSPEECH Lisboa, Portugal (2005)3065-3068
8. Mathieu Ben, Guillaume Gravier, Frederic Bimbot: A model space framework for efficient speaker detection. Proc. European Conf. on Speech Communication and Technology (2005) 3061C3064
9. Alexandre Preti, Nicolas Scheffer, Jean-Francois Bonastre: Discriminant Approaches for GMM Based speaker Detection Systems. MMUA, Toulouse, France (2006)

10. Rabiner Lawrence, Juang Biing-Hwang: Fundamentals of Speech Recognition. New York: Prentice Hall, (1993) 321-387
11. Paul Viola, Michael Jones: Rapid Object Detection Using a Boosted Cascade of Simple Features. IEEE Computer Vision and Pattern Recognition, Kauai Hawaii, USA. Volume 1.(2001) 511-518
12. Saitoh Takeshi, Kobayashi Harato and Konishi Ryosuke: Automatic speaker detection for videoconferencing. Technical report of IEICE. PRMU. Vol.106. No.72(20060518) (2006) 25-30
13. P Duchnowski, M Hunke, D Busching, U Meier: Toward movement-invariant automatic lip-reading and speech recognition. International Conf. Acoustics, Speech and Signal Processing, ICASSP-95 (1995)
14. Christoph Bregler: Learning and Recognizing Human Dynamics in Video Sequences. Proc. IEEE Conf. Computer Vision and Pattern Recognition (1997)568 - 574
15. Rabiner, L.R: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. Volume 77, Issue 2, (1989) 257 - 286

Intelligent Tutoring Systems