

Recognition of Multi-Pose Head Gestures in Human Conversations

Ligeng Dong, Yuxin Jin, Linmi Tao, Guangyou Xu

Department of Computer Science and Technology, Tsinghua University, Beijing, P.R.China
{dongligeng99, jyx05}@mails.thu.edu.cn, {linmi, xgy-dcs}@tsinghua.edu.cn

Abstract

We address the problem of recognizing multi-pose head nodding and shaking gestures in human conversations. Existing methods mainly recognize head gestures in restricted environments like human robot interaction, where face poses are near frontal and head motions are not natural. However, in human conversations, faces of subjects might be in arbitrary poses while head gestures are often subtle. Since the face pose change and head gesture movement are of different scale, we propose to track the face of varied poses with a mixed-state particle filter and detect the subtle head movement by a Kanade-Lucas-Tomasi tracker. The motion patterns in both horizontal and vertical directions are detected and then head gestures are analyzed by a Finite State Machine. Experiments on natural human conversations demonstrated the effectiveness of our method.

1. Introduction

Head gestures, as part of bodily communication, play an important role in human communications. Head nodding and shaking are two commonly used head gestures which might mean yes or no, understanding or disbelief, agreement or disagreement respectively. Automatic recognition of head gestures can help understand people's attitude to the speaker, which is useful for human conversation analysis.

To automatically recognize head gestures in human conversations remains an open problem. The main difficulty lies in the fact that natural head movements of nod or shake might be very subtle and hard to detect, especially for low resolution and low quality videos. Another problem comes from the dynamically changing head poses, since people often turn heads to focus on different speakers, which make it difficult to track a fixed facial feature robustly. So the algorithm should be able to detect two kinds of head movements in conversations, large scale movements of head pose change and subtle movements of head gestures.

Existing methods for head gesture recognition are mainly applied in restricted environments such as human computer/robot conversation and virtual reality environment. In these cases, the faces captured are usually near frontal and head motions are not as subtle as in human conversations. Head gesture recognition is usually comprised of two parts, motion estimation and temporal sequence analysis. For motion estimation, Davis [2] used a special hardware to detect eye area and then track facial features in the area. McGlaun et al [5] employed template matching method to track the nose. Tan et al [8] applied Adaboost method to detect face and track two eyes. However, these features are only acceptable in the case of frontal faces and not suitable for non-frontal head gestures. Also, as reported in [8], those features are not suitable to detect subtle head motions. Morency et al [6] used a stereo camera to estimate the 3D head pose parameters. Tang et al [10] tracked eight features by KLT tracker. However, they both only recognized gestures in near frontal faces. Lu et al [11] used five near frontal view face models to estimate the face pose but the face pose estimation is not sensitive to small head pose changes due to subtle head motions. For temporal sequence analysis, Finite State Machine (FSM) was used in [2], Hidden Markov Model (HMM) was adopted in [6][8][11] and Neural Network (NN) was employed in [10]. G. McGlaun et al [5] compared the performance of both the FSM and HMM approaches and found that both methods provided comparable results. The advantage of the FSM method is quickness and simplicity. HMM and NN need training and are complicated. And how to segment the sequence for HMM is not a trivial problem for natural head gesture analysis, since head gestures may be followed closely by other head motions in natural conversations. In [8], a fixed length of 16 frames is used for recognizing gestures, which may cause the miss detection of continuous gestures longer than 16 frames and the false detection of gestures shorter than 16 frames.

Different from previous methods, this paper tries to recognize natural head nodding and shaking gestures of various poses in human conversations. To the best

of our knowledge, this paper is the first attempt to recognize such head gestures. We employ a mixed-state particle filter with multi-pose face detector to do the head pose tracking and a KLT tracker to detect subtle head motions in arbitrary poses. It is found that head features have different motion patterns in profile head gestures from frontal ones, and we propose to recognize the motion patterns of both vertical and horizontal directions. Finally, a Finite State Machine with frequency constraint is used to segment and recognize head gestures at the same time.

The rest of this paper is organized as follows. Section 2 gives an overview of our method. Face pose tracking, head motion estimation and head gesture recognition modules are described in subsequent sections. Experiment results are reported in Section 6 and conclusions are drawn in Section 7.

2. Algorithm overview

There are mainly three modules in our method, face pose tracking, head motion estimation and head gesture recognition. Face pose tracking and head motion estimation are two parallel modules and they can influence each other during tracking. The output of head motion estimation is fed into head gesture recognition module. The tracking is initialized by a multi-pose face detector and features in the face area are selected for tracking. For each frame, head motion estimation is firstly performed by the KLT tracker. If no or very subtle motion between frames is detected, face pose tracking module can keep the track of the previous frame. If big head motion is detected, pose tracking module then start tracking again. During pose tracking, face poses are estimated by our multi-pose face detectors. When there is continuous monotonous pose change, for example, the face changes from frontal pose to half-profile pose, features in the face area are reinitialized to select good features to track. The flow chart of the method is illustrated in Figure 1.

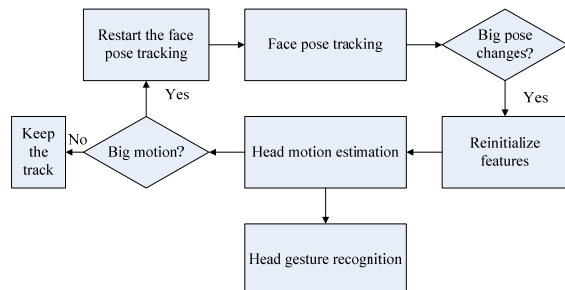


Fig. 1. The flowchart of our method

3. Face pose tracking

Conventionally, face tracking and pose estimation are organized sequentially. The pose estimation module is dependent on the result of the face tracker. However, the pose estimation result is very sensitive to the quality of the extracted face box. So a bad aligned face box might result in wrong pose estimation. Actually these two modules are correlated to each other. Good tracking results can help pose estimation and correct pose based models can help to obtain good tracking results. So we employed a mixed-state particle filter[1][3] to handle the face tracking and pose estimation jointly. The face state is mixed with continuous spatial parameter x and discrete pose parameter k . We use a boosted multi-pose face detector [9] to initialize face tracking and measure face samples. A likelihood function [12] producing probabilistic output is employed to measure the probability of a sample being a face of a certain pose. Then the face and pose are weighted from all the samples.

4. Head motion estimation

4.1. KLT feature tracking

Our head motion estimation module is based on the Kanade-Lucas-Tomasi(KLT) feature tracker[4][7]. The tracker first selects features with great intensity gradient along at least two directions as the good features to be tracked in following frames [7]. Then the most likely position of the feature in the next frame is found and tracked. This is done by matching the feature's neighboring area to the most similar area in the next frame within a certain search window in a Newton-Raphson optimization way. In order to improve efficiency, the method can be implemented in pyramids [7], where the feature is first matched in the higher low resolution pyramid and is refined in the lower high resolution pyramids iteratively. The Intel OpenCV[13] provides the implementation of KLT feature tracker and finds the coordinates of the feature point with sub-pixel accuracy.

Tang et al [10] also detected and tracked features by KLT tracker, but they only tracked eight features in near frontal faces. They used these features to estimate the 3 rotation parameters of head motions. However, in human conversations, it is difficult to estimate these parameters robustly in arbitrary poses. We detect good features for track by the KLT tracker which are far more than 8, and used them to estimate the vertical and horizontal motion directions in different poses. This is discussed in the next section.

4.2. Motion direction estimation

A set of movement directions is defined {NULL, RIGHT, UP, LEFT, DOWN}. We first estimate the motion direction of each feature and then that of the head. The motion vector (dx, dy) of each feature between frames is computed. Then the motion direction of the feature is set as the direction with the larger motion. If $|dx|$ and $|dy|$ are both smaller than a certain threshold, we treat the feature as static. The main motion direction of the head can be determined by the direction with the most number of features. For example, in frontal faces, in the up or down movements of the nodding gesture, usually $|dx|$ of each feature is small and $|dy|$ is big. So the main motion is vertical. Similarly, the main motion of shaking gesture is horizontal.

However, for a profile or half-profile face, in the nodding gesture, the number of features moving vertically might be comparable with the number of features moving horizontally. This is because nodding gesture in profile faces is kind of rotation-in-plane motion, where there are both motion in horizontal and vertical directions. Figure 2 illustrates the motion patterns of head nodding and shaking gestures in different poses.

To handle this case, the main motion directions of both horizontal and vertical directions are estimated separately. If there is a periodical pattern in vertical direction, it is likely that it is a nodding gesture. If there is a periodical pattern in horizontal direction, it is likely that it is a shaking gesture. If there are periodical patterns both vertically and horizontally, it is likely that it is a nodding gesture in profile faces.

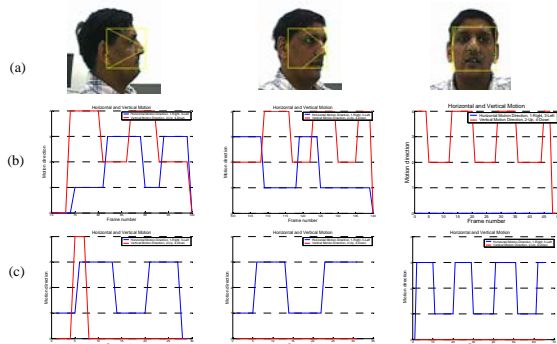


Fig. 2. Motion direction patterns of nodding and shaking in multi-pose faces. Row (a) are tracked faces where green points are features to estimate motion. Row (b) and row (c) are motion patterns for nodding and shaking gestures, respectively. The number 0-4 in y axis means the motion direction of {NULL, RIGHT, UP, LEFT, DOWN}. Red and blue lines are motions in

vertical and horizontal directions, respectively. Nodding gesture in profile and half profile faces might have both periodicities in horizontal and vertical directions. However, for all poses, motion patterns of shaking gestures are similar, although there might be noisy motions in vertical direction like figure in row (c) column (i)

5. Head gesture recognition

We model the head gestures with a Finite State Machine (FSM). The model consists of 5 states: {Start, Up, Down, Right, Left}. The initial state is the Start state. When there is a motion, the state will transfer to the corresponding state. Then the states keep transferring according to the transition rules in the FSM. A head nodding gesture will be announced when the new motion is no motion or invalid motion like right or left. And we also need to judge if the motion sequence has gone through both the up and down states. If the motion sequence only goes through one state, the motion sequence is not a nod gesture. The shaking gesture is declared in a similar way. When there are continuous nods or shakes, each nod or shake can be declared when a cycle of the states is finished. So FSM is very suitable for continuous head gesture recognition and is able to segment gestures followed other head motions. Figure 3 shows the FSM we used.

As discussed in section 4.2, there are two motion direction sequences for the horizontal and vertical direction respectively. They will be both fed into the FSM. If a certain sequence is declared as both nodding and shaking, then we classify the gesture as nodding because only nodding gestures in profile faces have such motion patterns.

In nodding or shaking, the motion direction will be periodically changing to the opposite direction. According to observation and statistics, before the motion changes to the opposite direction, there are usually one or two frames with no motion. So only if there are more than 2 null motions, the state will transfer to Start state.

Previous studies [2][5] also used FSM to analyze head gestures. However, they only analyzed the frontal head gestures. And they first segmented the head motions between null motions and then analyzed the sequence. Here, we used the FSM differently from them. Firstly, we analyzed the motion patterns in different poses. Secondly, both the vertical and horizontal patterns are analyzed and then considered together. Thirdly, in natural multi-party conversations, it is very common that people do some head gestures and then move their heads to other directions. So the

head gesture motions are closely followed by other motions. FSM is here used to segment and recognize those head motions at the same time.

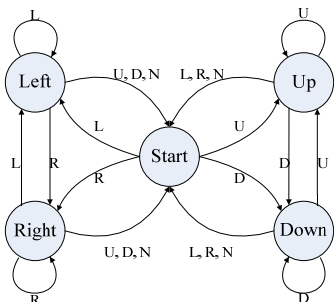


Fig. 3. Finite State Machine used in our method

6. Experiments

We captured videos of subjects in two-subject face-to-face conversations to evaluate our algorithm for head gesture recognition in multiple poses. The camera only focused on one subject A and the other subject B sat opposite to A. They talked freely and A was asked to nod or shake his head naturally showing his agreement or disagreement with B respectively. In order to capture multi-pose faces of A, B was asked to sit at different positions so that the frontal, half-profile and full profile faces of A can be captured by the fixed camera. We captured video data of 8 individuals, where each video contains head gestures in various poses. Each subject was asked to answer 6 questions by nodding or shaking in each pose. The image size of the video is 360*288, and the face size is about 50*50.

Head gestures are recognized using FSM introduced in section 5. According to observations, it is found that the motion interval in one direction of a head gesture cycle may be from 4 to 10 frames. During recognition, if a single sequence is longer than 10 frames or shorter than 4 frames, we treat them as invalid motions. And gestures should have similar timings in two motion directions. Otherwise, they are invalid gestures.

The ground truth of head gestures in our 8 test videos is labeled manually. And we compared the results of our method with the ground truth. In natural conversations, the subject may conduct continuous nods or shakes. If there are continuous head gestures, we add each gesture to the total number. Our method is evaluated by three parameters: the recognition rate (RR), the false accept rate (FAR) and the false reject rate (FRR). The result is listed in Table 1. Ground truth (GT) means the number of gestures.

The results show that frontal head gestures are easier to recognize than non-frontal ones. This is because motions in frontal head gestures are easier to

detect than in non-frontal ones. Some unrecognized gestures result from very slow head motions with more than 2 static frames at the reflection point so that a true gesture is separated as two sequences and not recognized. A large part of false accepted gestures come from arbitrary head motions which have the same motion pattern as nodding and shaking. This kind of head motions happen occasionally and are very difficult to discriminate from true head gestures.

Figure 4 shows some representative frames of a video clip where the head of the subject is oriented at different directions. Figure 5 shows the head gesture recognition result. The correctness of the head tracking is visually confirmed and all the head gestures are successfully detected and recognized. There are two false alarms in this video clip. The first is caused by arbitrary head motions around frame 30 and the second is caused by the head turning motions around frame 300, where the motion patterns are similar as those of head shakings. This is because when the head is turning, it might happen that features along the head contour move in the opposite direction of features around the face center and the former dominate the motion direction. So the head pose trajectory will be used to eliminate this kind of false alarm.

Compared with existing methods, our method is able to detect very subtle head movements. Usually head nods may be more subtle than head shakes. Table 2 lists some frames and the corresponding average motion values from a head nodding gesture. Figure 6 shows the corresponding frames and the features. The average motion value of moving features between 2 successive frames is less than 2 pixels in this gesture. It is difficult for existing methods such as nose template matching or eye tracking to detect such subtle movements in various poses. In our current implementation, we set the threshold of motion as 0.4 pixels. Any motion of features less than 0.4 pixels will be regarded as noises.

Table 1. Results of multi-pose head gesture recognition

| Nod | GT | RR | FAR | FRR |
|--------------|-----|-------|------|------|
| Frontal | 50 | 98.0% | 4.0% | 2.0% |
| Half-profile | 72 | 94.4% | 6.9% | 5.6% |
| Full-profile | 78 | 93.6% | 9.0% | 6.4% |
| Total | 200 | 95.0% | 6.5% | 5.0% |

| Shake | GT | RR | FAR | FRR |
|--------------|-----|-------|------|------|
| Frontal | 54 | 98.1% | 9.3% | 1.9% |
| Half-profile | 72 | 95.8% | 7.0% | 4.2% |
| Full-profile | 71 | 95.8% | 1.4% | 4.2% |
| Total | 197 | 96.4% | 5.6% | 3.6% |



Fig. 4. Representative frames with different poses

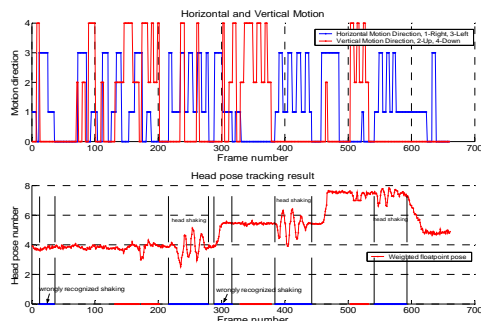


Fig. 5. Head gesture recognition result of a video. The upper figure is the motion pattern in horizontal and vertical directions. The bottom one is the pose tracking result, where red lines are weighted float point poses. The blue and red segments on the bottom mean the recognized head shaking and nodding segments

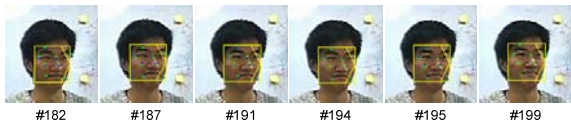


Fig. 6. The tracked face and features from a nodding gesture. Yellow rectangle is the tracked face area. The green and blue points are the features in the previous and current frame respectively

Table 2. Estimated motion direction of some frames

| Frame number | 182 | 187 | 191 | 194 | 195 | 199 |
|----------------|------|------|------|------|------|------|
| Direction | down | up | down | up | up | down |
| Average pixels | 1.95 | 1.53 | 1.54 | 0.48 | 1.21 | 0.46 |

7. Conclusions and future work

We have proposed to track face poses and recognize head gestures in natural human conversations. Our method firstly detects and tracks the multi-pose faces. Then a set of features in the face area are selected and tracked to estimate the head motions in both the vertical and horizontal directions.

After that the motion sequence is analyzed by a FSM which can segment and recognize gestures at the same time. Experiments show that our method can detect subtle head gestures of various poses in human conversations.

In the future, head poses and gestures of participants in a multi-party conversation will be combined with speaker detection to infer high level semantics, such as their attitudes to the speaker.

Acknowledgement

This work is supported by the National Science Foundation of China under grants No. 60673189.

References

- [1] S. Ba. and J.M. Odobez, "A Probabilistic Framework for Joint Head Tracking and Pose Estimation," In ICPR 2004
- [2] J. Davis and S. Vaks, "A perceptual user interface for recognizing head gesture acknowledgements," In WS on Perceptive User Interfaces (PUI 01), 2001
- [3] M. Isard and A. Blake, "A Mixed-State CONDENSATION Tracker with Automatic Model-Switching," In ICCV 1998
- [4] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," In Proc. International Joint Conference on Artificial Intelligence, pp. 674–679, 1981.
- [5] G. McGlaun, et al, "Robust Video-Based Recognition of Dynamic Head Gestures in Various Domains - Comparing a Rule-Based and a Stochastic Approach," in 5th International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction, GW 2003, Genova, Italy, April 15-17, 2003.
- [6] L.P. Morency et al, "Contextual Recognition of Head Gestures", International Conference in MultiModal Interaction (ICMI), ISBN: 1-59593-028-0, pp. 18-24, October 2005
- [7] J. Shi and C. Tomasi, "Good features to track", Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., pages 593-600, 1994.
- [8] W. Tan and G. Rong, "A real-time head nod and shake detector using HMMs," Expert Systems with Applications 25, 2003
- [9] Y. Wang et al, "Real-Time Multi-View Face Detection and Pose Estimation in Video Stream," In ICPR 2006. vol.4, no.pp. 354- 357, 20-24 Aug. 2006
- [10] J. Tang and R. Nakatsu, "A Head Gesture Recognition Algorithm," in ICMI 2000.
- [11] P. Lu et al, "Head Nod and Shake Recognition Based on Multi-View Model and Hidden Markov Model," In CGIV 2005.
- [12] P. Li and H. Wang, "Probabilistic Face Tracking Using Boosted Multi-view Detector," In PCM 2004.
- [13] Intel Open Source Computer Vision Library, www.intel.com/technology/computing/opencv/.