Smiles vs chewing vs speech detection by similarity matching

Wenyang Cai

August 20, 2014

Abstract

This is abstract.

Acknowledgements

This is acknowledgement

Contents

Chapter 1

Introduction

This is Introduction.

1.1 Motivation

1.2 Thesis

1.3 Contribution

Chapter 2

Background

This is Background [2]. This is Background [5]. This is Background [3].

2.1 Automatic Speeech Recognition

2.2 Visual Front End

2.2.1 Face and Facial Part Detection

2.2.2 Region of Interest

2.3 Facial Expression

2.4 Nonlinguistic Vocalization

Chapter 3

Processing and Methodologies

3.1 Processing Flow

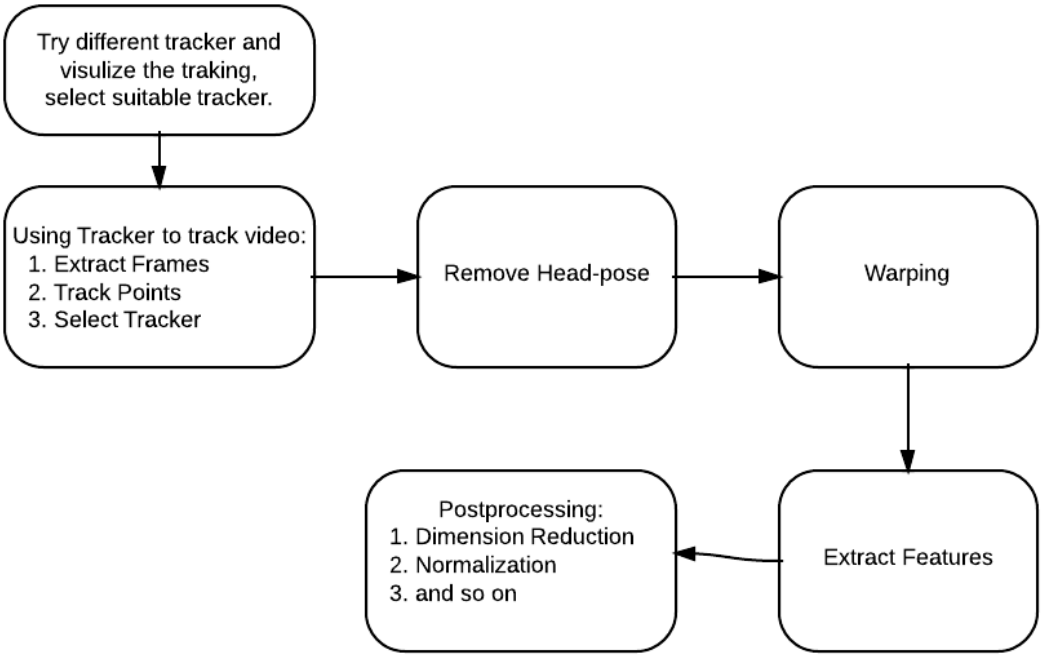

Figure 3.1: Main procedures

Processing Chart The figure above shows main procedure of the whole project. At beginning, I tried several trackers. Intraface and DRMF are two trackers I tried and compared most. Two trackers are using different methods and also implemented in different languages. Intraface are programmed in c and matlab and has great interface for matlab. I tried two version of DRMF, DRMF programmed using CUDA which uses parallel processing is quite fast. As the programme of DRMF doesn't integrate extract frames from videos. The images are extracted using external function, then tracked using DRMF. I choose Intraface as the final choose, the reason and comparison will be given in later section. Remove head-pose seems to be a very important part for this project, as subject's head moves frequently in many videos. After having tracking points without head-pose, each face in each frame is warped and scaled to same size grey image. Extracting features is to extract appearance feature of each face in the image. Post-processing is preprocessing before using the data for classification.

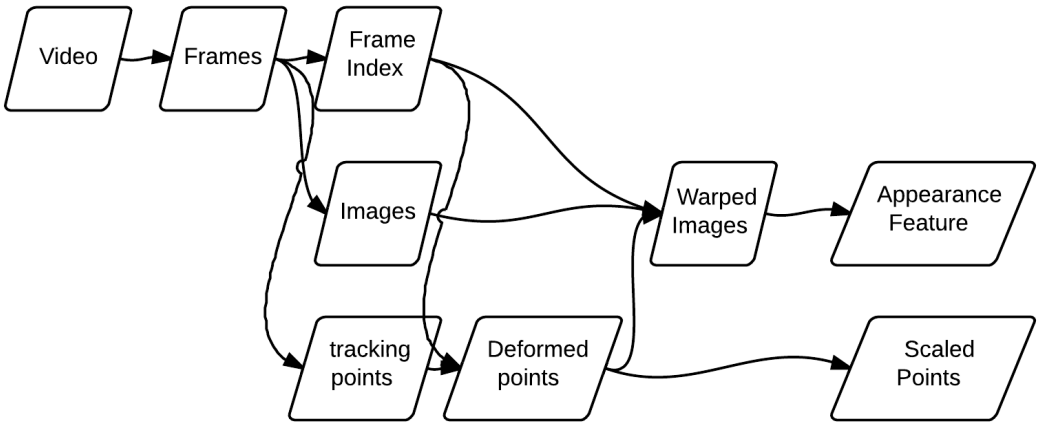

Figure 3.2: Eating sequence tracked by Intraface

Data flow Chart There are two types of encoded video, one is in format of fly and the other is avi. Extracting frames from videos is proceed with Intraface and stored in formats of jpeg and mat which used for processing of matlab. There are several situations that a track is unable to track a face in the image such as no subject in the image, the head is face to a very large angle from frontal face, face is partially not show in the frame. Frame index points to those images which the tracker is able to tracker a face in a frame. Image is the frame images stored in mat format. Different tracker may tracks different number of characteristic facial points. Intraface tracks 49 facial points and DRMF tracks 66 points. Deformed points is the tracking point after removed head-pose. Warped Images is the face after remove head-pose and background which only leaves the meshes build by tracking points. Appearance feature is face feature extracted using local binary pattern (LBP). As the image size from image to warped images are changed, the points is rescaled from deformed point to scaled points.

3.2 Face Alignment

Face alignment is to align face in one image with respect to the same face in another image. Face alignment techniques are used to track characteristic facial points in image sequences. In this project, the aim of face alignment is to localise the feature points on face images. The points are usually around eyes, nose, mouth, and outline. Face alignment techniques are essential on face recognition, modelling and synthesis. There are three main different approaches Parametrized Appearance Models(PAMs), Discriminative approaches, Part-based deformable models. Parametrized appearance models contains many models such as active appearance models (AAMs), morphrable models, eigentrackings, and template tracking [5]. All these models are using PCA method to parametrize a face. A face could approximately decomposed as linear combination of shape basis and appearance basis. The problem of face alignment could be refer as minimising the difference between the constructed PAM and the face. Common approach is use Gauss-Newton methods [5]. Discriminative approaches are to learn the linear regression between the head move and appearance change. Part-based deformable model perform face alignment by maximising the posterior likelihood of part locations given image [5].

3.2.1 Aative Appearance Model

Active Appearance Model (AAMs) is defined as a generative model of a certain visual phenomenon in [2]. AAMs are conceptually related to morphable models, constrained models and active blobs. In this project, it is refer to a model of face. As AAM is conceptually related to other parameterized appearance model, so it is introduced as an example of parameterized appearance model for under-

standing purpose. According to [2], there are two types of AAMs, one refers as independent shape and appearance models, which model shape and appearance independently, and the other refers as combined shape and appearance models, which parameterized shape and appearance model with a single set of linear parameters [2]. Normally AAMs appears along with a fitting algorithm. However, in the following context, it only refers to a model. [2] gave a well explain about what is an AAM, most of following theory are from [2].

Shape Shape of a face s is defined by coordinates (x, y) of v vectices of face points and the mesh they built:

$$s = (x_1, y_1, x_2, y_2, ..., x_v, y_v)^T \tag{3.1}$$

s also can be expressed as a base shape s_0 plus linear combination of n shape vectors s_i:

$$s = s_i + \sum_{i=1}^{n} p_i s_i \tag{3.2}$$

Appearance For all pixels x in the mesh s_0, appearance $A(0)$ can be expressed by base appearance $A_0(x)$ and m appearance images $A_i(x)$.

$$A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) \qquad \forall x \in s_0 \tag{3.3}$$

AAMs are usually computed by applying Principle Component Analysis (PCA) to choosen images. The chosen images contains a variety of shapes. The base shape s_0 is the mean shape and the vector s_v is the eigenvector corresponding to the largest v eigenvalues. The base appearance A_0 and the appearance A_i is computed by applying Principle Component Analysis to a set of shape normalised images.

Model $W(x : p)$ is the warp from s_0 to s. Then the model M set the appearance of $W(x : p)$ to $A(x)$.

$$M(W(x : p)) = A(x) \tag{3.4}$$

Combined AAMs
Combined AAMs just use parameter $c = (c_1, c_2, ...)^T$ to parametrize shape:

$$s = s_0 + \sum_{i=1}^{l} c_i s_i \tag{3.5}$$

and appearance:

$$A(x) = A_0(x) + \sum_{i=1}^{l} c_i A_i(x) \tag{3.6}$$

3.2.2 Trackers

I tried two main trackers for tracking characteristic facial points, one is Intraface [5] which use suppervised decent method, the other is DRMF [1] which use discriminative response map fitting. The number of landmark points, tracking effect and programming execution time are quite different.

Intraface [5] implies image alignment can be posed as solving a nonlinear optimization problem. It uses Supervised Descent Method for minimising Non-linear Least Square(NLS) function, which avoids calculating the Hessian and the Jacobian that could be computationally expensive. Intraface tracks points very effective and efficient.

Tracking Points

Figure 3.3: Eating sequence tracked by Intraface

Eating and Talkingsequence A image sequences of eating tracked by Intraface is shown above. The point are aligned very precisely along the face. A image sequence of talking tracked by Intraface is shown above.

DRMF DRMF uses novel discriminative regression based on Constrained Local Models(CLMs) for face alignment. The basic idea of DRMF is to fit a face for each frame of a video. There are 66 points tracked by this video.

Tracking Points

Talking and Talking Sequence A image sequences of eating tracked by Intraface is shown below.

Figure 3.4: Talking sequence tracked by Intraface

3.2.3 Comparison

The following are some examples for comparing two trackers. There are two version of DRMF tracker one is implemented by CUDA language the other is by C language. Although the C version of DRMF is very slow and very easy to run out of memory, the version implementd by CUDA is very fast as CUDA is using parallel computing. However, DRMF is not as accurate as Intraface and not suitable for this project. In many situations, DRMF try to fit a face and the fitting result is awful. The images with red points are tracked by Intraface, and the images with blue points is tracked by DRMF. From figure above, it seems both tracker can only be used to trak one face at one time. DRMF detect the smaller face instead of the bigger one is possiblly because of the algorithm. The second and third frames are tracked by DRMF is not very accurate on the nose area. As the track point of mouth, intraface is better than DRMF. In some frame, partial of face is out of frame. Intraface is better dealing with this type of situation. Intraface ignore those points that out of the images. DRMF tries to fit a face forcibly. It often lead to bad influence on the traking result shown in the figure.

3.3 Remove Head-pose

The algroithm of removing head-pose from tracking points is in [4].The following are some example of orginal track points and deformed points:

Figure 3.5: Talking sequence tracked by DRMF

3.4 Warping

In order to have the appearance image of the face after removed head-pose, it is necessary to warp the face with head pose. Basic idea is to for each triangles builded by tracking points, the image points in the triagnles are projected to the corresponding triagnles built by deformed points. The following are some examples of face before and after warping:

3.5 Feature Extraction

The image after warping is not directly used for classfication. The data for classfication is the features of the image. There are many techniques to extract features from images, in this experiment, Local Binary Pattern are used for extracting image feature.

3.5.1 Local Binary Pattern

Effective facial representation of the original face iamges is an important part of successful facial expression recognition.

3.6 Postprocessing

Due to the time limits, in the experiment part, we only use support vector machine to do classfication.

Figure 3.6: Eating sequence tracked by DRMF

Normalization

Scaling

Figure 3.7: Tracking result, red points tracked by Intraface and blue points tracked by DRMF

Figure 3.8: Tracking result, red points tracked by Intraface and blue points tracked by DRMF

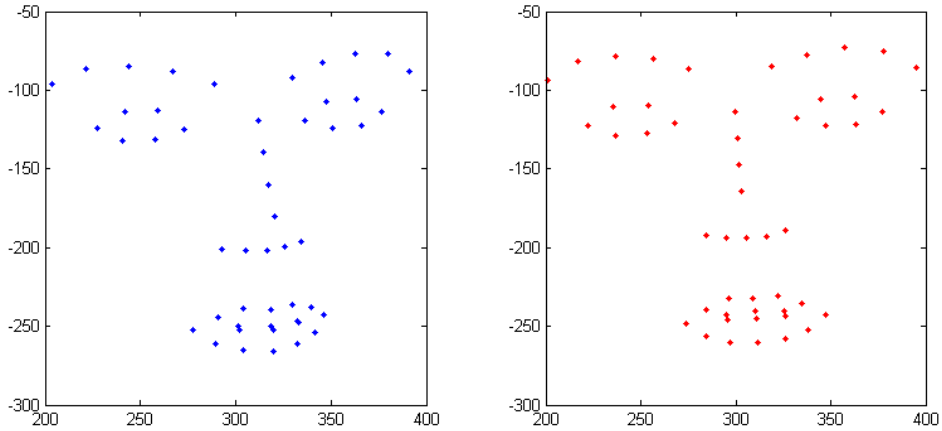

Figure 3.9: Traking points and Deformed Points, Example 1

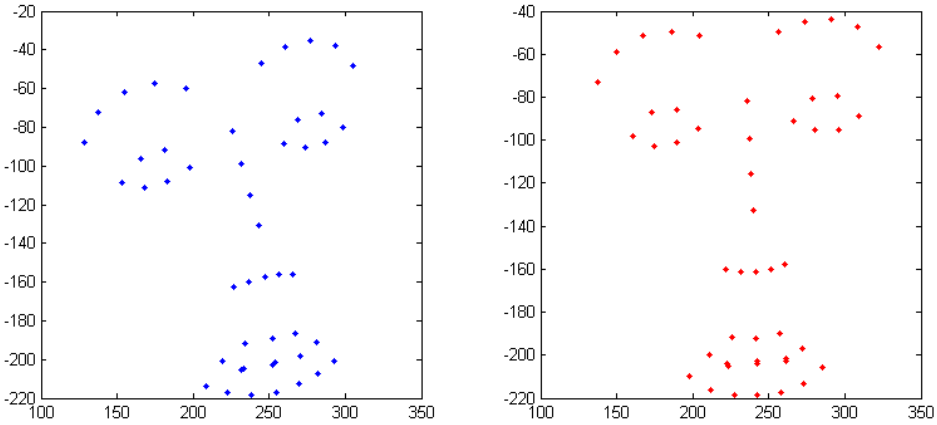

Figure 3.10: Traking points and Deformed Points, Example 2

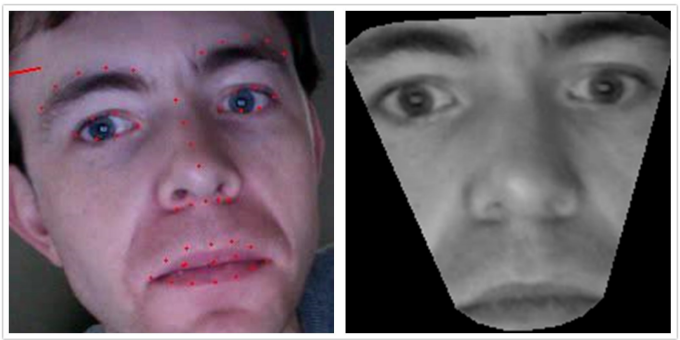

Figure 3.11: Talking sequence tracked by DRMF

Chapter 4

Experiment and Results

4.1 Database

4.1.1 Feature and Data

4.2 Methodology

RBF

4.3 Experiments

4.4 Results and Analysis

Chapter 5

Conclusion and Future Work

Bibliography

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE, 2013.

[2] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[3] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[4] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[5] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.