Smiles vs chewing vs speech detection by similarity matching

Wenyang Cai

August 26, 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

In human-human communication, human would be able to understand each other through multiply ways. Human is able to understand others by conversation, facial expression, gesture, even intonation, mood. This phonomenon was introduced into human-computer interactions(HCI). Using computer to automatically analyze human face, voice and behaviours would be helpful in imporving HCI. This techniques not only can be used in avoid impostor attacks but also be used in automatic speech recognition. Face and voice are important and personal biometric characteristics. Compare to traditional knowledge-based and token-based person recognition method, biometric recogntion technology are more convenient and safer. With the increasing in computer computing power and development in computer vision, Automatic Speech Recognition(ASR) attracts more attention in recent years, from traditional audio-only ASR to audiovisual ASR, a huge progress was made in ASR.

2.1 Automatic Speeech Recognition

Intuition of Automatic Speech Recognition is to convert a spoken sentence into to readable text in real time by computer. It has been researched over 50 years, and the ultimate goal of ASR is to let a computer 100% accurately recognise speech of any person under any environment. However, the accurace of recognition highly depended on robust information channel, background environment, the training data base and the adaptation of speaker to database. The beneficial of ASR is quite obvious, it can help deaf people listening by convert speech into readable text and help hard-reading people on reading by convert readable text into voice. The search engine may not be limited on text searching but also speech search. With decreasing price of computing power, Speech recognition techniques are widely used on mobile devices, like siri.

As in human-human communication, visual feedback is very important, visual information plays an important role in helping people understanding each other. Visual modality was proved to have positive influence on reducing noise in ASR and the history can be quantified back to 1954 [6]. Deaf people is able to communicate with each other by reading mouth movements. [6] gave three key reasons to include vision information in human speech recognition. Firstly , it helps audio source localisation, visual information of tongue, teeth, and lips provide complementary information of articulation. It is beneficial for distinguish confusable acoustics such as unvoiced consonants $/p/$ by providing information of facial muscle movements. Facial muscle movements are robust information for ASR. This technique of using visual information to recognise speech is known as automatic lipreading or speechreading in ASR [6].

Audiovisual ASR uses both visual modality and audio modality in recognizing speech. Two main challenges are introduced by AV-ASR, how to extract visual features, how to combine it with audio features. Visual speech information mainly from speaker's face. Extract visual feature requires face detection, face alignment, tracking, feature extraction and other techniques to extract useful visual information from image with a face. Combination of two modality are also a challenge for AV-ASR.

2.2 Visual Front End

A major problem in audiovisual automatic speech recognition is extracting visual feature from images, the inputs are usuablly videos and the output should be visual speech features. Generally, visual speech feature can be classified as three types: appearance feature, shape fewature, combination of both [6]. The image of Region of Interest used to be directly used for trainning and classification. However, image data contain many noises and influenced by the lighting condition. Then some techniques in computer vision are used to extract image features from Region of Interest. Image feature contains points, edges, texture, colors and so on. Shape feature usually means coutours of speaker's face, specifically speaker's lips or including jaw and cheek. Shape feature usually means geometric-type features, such as statistical shape model or image moment descriptor of mouth, these model would be able to contain the information of the height, width and other information of a mouth. Combined feature usually is the joint of both shape and appearance feature vectors or a model that include both features like active appearance model.

In order to get appearance feature and shape feature or combined feature, there are some preprocessing before extracting feature. Face detect, tracking, alignment, and ROI extracting technique are needed. In order to align a face with feature points, some statistical models are used to fit the face. In order to remove head-pose, some technique are used to decompsite the head points. If the head is not facing the front, apperance of face would not be correct, warping techniqure are also needed.

2.2.1 Face Detection

Several main aspects that influence face detection is background, head pose and lighting. [6] reports taht many system uses traditional image processing techniques, such as color segmentation, edge detection, image thresholding, template matching or motion information and some using statistical modeling and neural networks. Once a face is detected, use face alignment technique to locate several facial feature around the face.

2.2.2 Region of Interest

The choice of Region of Interest(ROI) is according to the purpose of project. In AV-ASR it usually include large part of the lower face, such as the jaw, and cheeks or even the entire face [6], as when people speak, the lower face would show some movements. In my project, the ROI only contains the grayscale values of mouth region, which is scaled to $8 * 32$ size square region. [6] report that experiments shows that including jaw and cheeks was beneficial. As the tracker I use for tracking face does not include the jaw and cheek of the face, I have to just include the mouth.

2.2.3 Visual Feature and Postprocessing

There are many different types of visual features you can extract from ROI. The choice of visual feature is depends on the requirements of the project. Dimensionality of feature vector depends on the feature extraction methodology. If the dimension of feature vector is too high, the common dimension reduction methods are traditional linear transforms. [6] gave some most commonly applied methods, Principle Component Analysis(PCA), discrete cosine transform(DCT), discrete wavelet transform, Hadamard and Haar transforms and a linear discriminant analysis based data projection. The visual feature are extracted from variety of lighting condition and different face. These effects can be remedied by normalization, i.e. for each visual feature vector subtract it's mean and divide by the standard deviation.

2.3 Facial Expression

Generally, face would show four kinds of signals: static facial signal, slow facial signals, artificial signals and rapid facial signals [5]. Rapid facial signals underlie facial expressions. [5] indicate that

rapid facial signals generally show five types of messages: affective attitudinal states and moods which means emotions, emblems, manipulators, illustrators, regulators. For example, smile belongs to regulators and chewing belongs to manipulators. Automatic analysis of facial signals such as rapid facial signals have potential applications in many areas. Layers, security, police could be use automatic analysis of facial signal system to monitoring and interpreting human facial signal and gain important information. For example, monitoring human reaction during inquisition, inquisitor would be able to tell whether a person is lie or not. Machine analysis of facial expression forms an important part of affective human-computer interface designs [5]. Research on machine analysis of facial expression mainly focuses on facial affect and facial muscle action detection [5]. Technologies used in this area are face detection, facial feature extraction, facial muscle action detection and emotion recognition [5].

2.4 Nonlinguistic Vocalization

Chapter 3

Processing and Methodologies

In this chapter, I will introduce the procedure of extracting facial features and the tools and methodologies I use.

3.1 Processing Flow

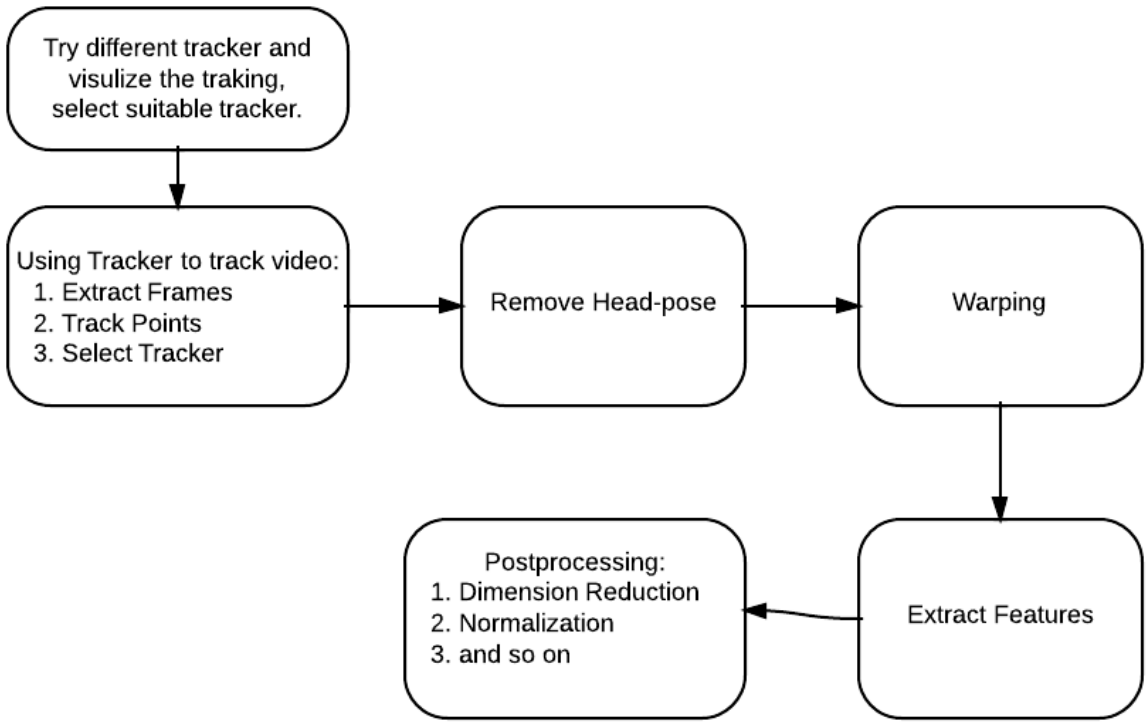

Figure 3.1: Main Procedure

Processing Chart Figure 3.1 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.

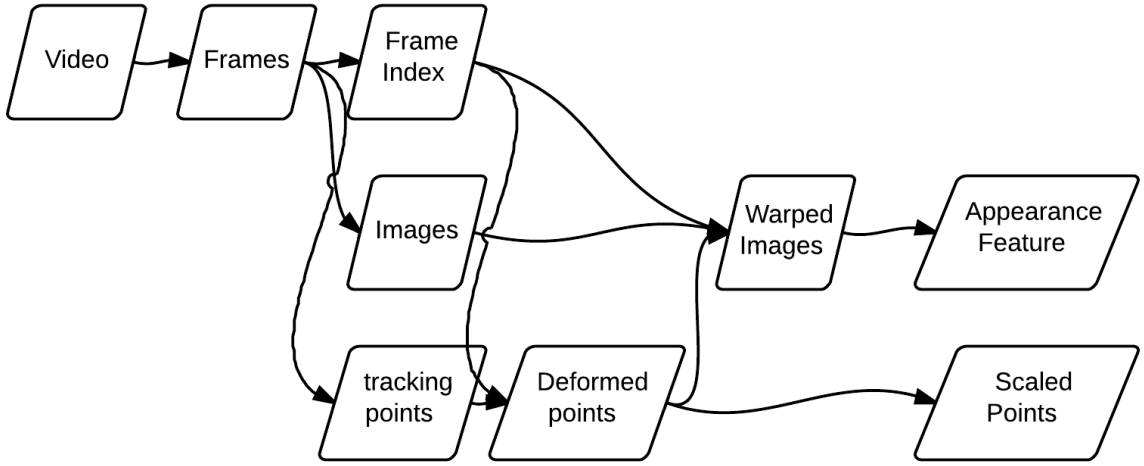

Figure 3.2: Data Flow

Data flow Chart Figure 3.2 show data I need for processing. 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 [9]. All these models are using Principle Component Analysis(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 [9]. 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 [9].

3.2.1 Active Appearance Model

Active Appearance Model (AAMs) is defined as a generative model of a certain visual phenomenon in [4]. 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 understanding purpose. According to [4], 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 [4]. Normally AAMs appears along with a fitting algorithm. However, in the following context, it only refers to a model. [4] gave a well explain about what is an AAM, most of following theory are from [4].

Shape Shape of a face s is defined by coordinates (x, y) of v vertices 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 choose 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:

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

An example of AAM instantiation is clearly shown in figure 3.3.

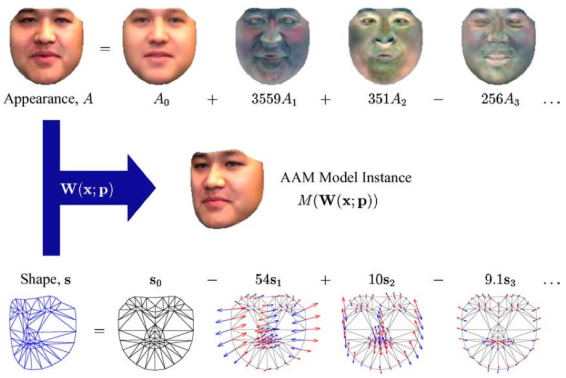

Figure 3.3: An AAM instantiation

3.2.2 Trackers

There are many different trackers for tracking facial feature points. Different tracker may use different approaches, so they may be applied into different situations. I tried two main trackers for tracking characteristic facial points, one is Intraface [9] which use suppervised decent method, the other is DRMF [1] which use discriminative response map fitting. Those two trackers not only using different approaches, the number of landmark points are also different.

Intraface

[9] 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. The running time of Intraface shows that the method is very effective and efficient.

Tracking Points The following tracker show the tracking points of Intraface. This tracker tracks 49 facial feature points. As you can see the eyes, nose, mouth, unfortunately the jaw and cheek may contain visual information that would help classification. Without the bound of face, I am unable to extract the region as I can not do warping of that area.

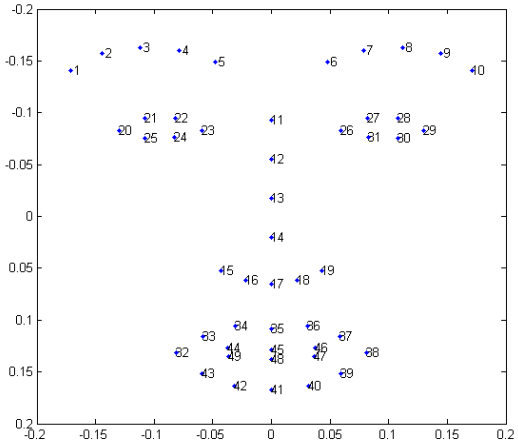

Figure 3.4: Intraface landmark points

Eating and Talking Sequence Figure 3.5 show a sequence of image of eating tracked by Intraface. The point are aligned very precisely along the face. Figure 3.6 shows a talking sequences of image tracked by Intraface. The landmark points of mouth is very accurate.

Figure 3.5: Eating sequence tracked by Intraface

Figure 3.6: Talking sequence tracked by Intraface

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. After locating the position of a face, the tracker tries to fit a trained constrained local model to fit the face. Sometimes the fitting result is not very good and the landmark points of mouth region is not very accurate.

Tracking Points DRMF tracked 66 facial feature points, the extra 17 points are the point around face bound. Other landmark points are the same with Intraface.

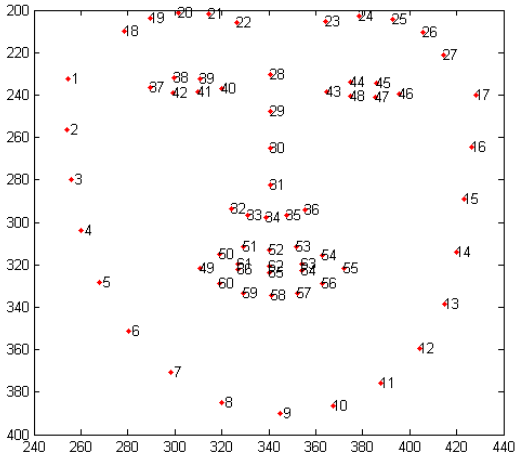

Figure 3.7: DRMF landmark points

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

Figure 3.8: Eating sequence tracked by DRMF

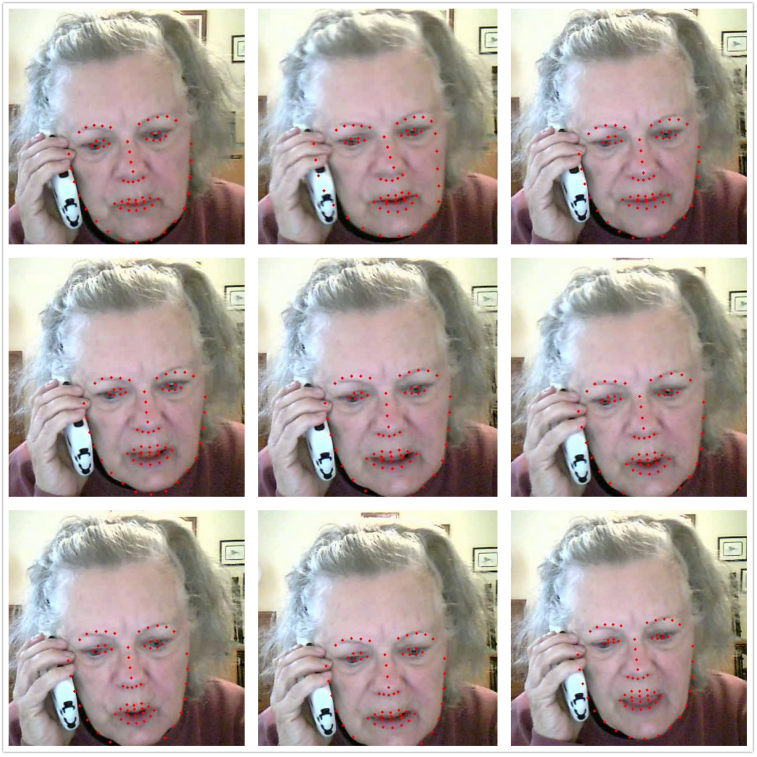

Figure 3.9: Talking sequence tracked by DRMF

3.2.3 Comparison

The following are some examples for comparing two trackers. Intraface is generally better than DRMF in accuracy and efficient. 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.

Figure 3.10: Tracking result, red points tracked by Intraface and blue points tracked by DRMF

Figure 3.11: Tracking result, red points tracked by Intraface and blue points tracked by DRMF

3.3 Remove Head-pose

The algroithm of removing head-pose from tracking points is in [7]. The following are some example of orginal track points and deformed points. Remove head pose of face and the warp the face to frontal direction is the most important part of extracting appearance feature. Here are two examples, from example one, rotation on x-y direction is mostly removed. In example two, the algorithm also show some efficient on remove head-pose of x-z direction and y-z direction.

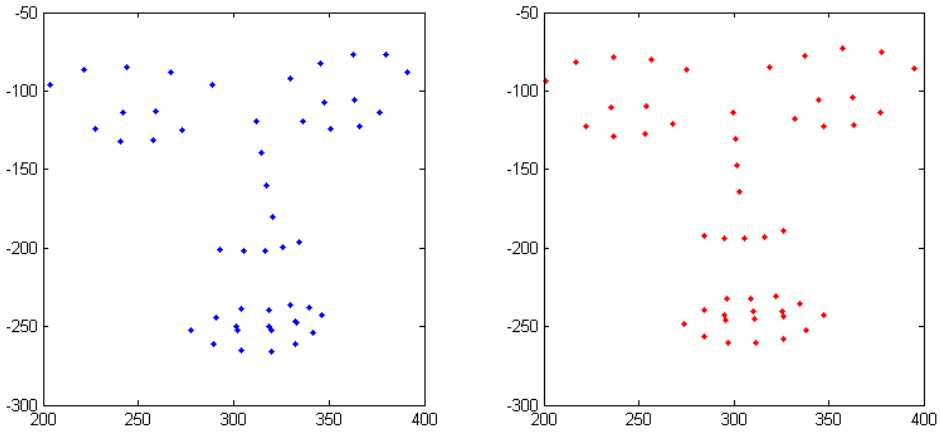

Figure 3.12: Traking points and Deformed Points, Example 1

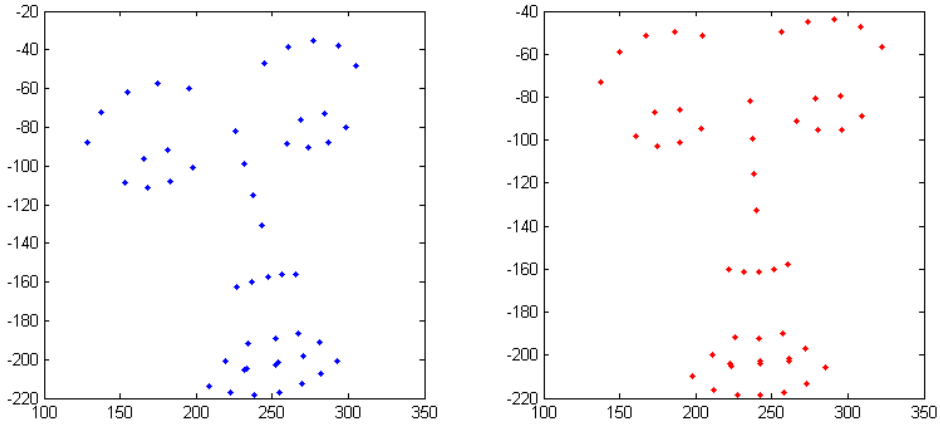

Figure 3.13: Traking points and Deformed Points, Example 2

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 shape 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:

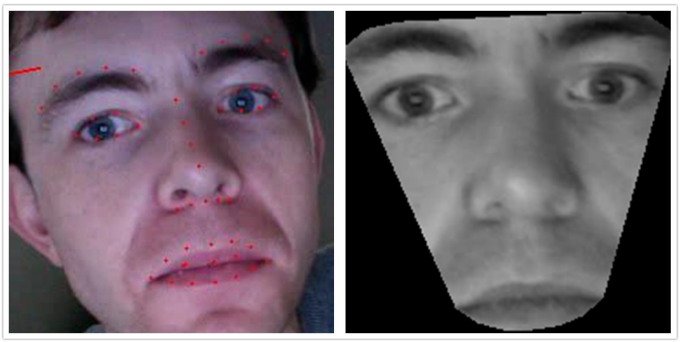

Figure 3.14: Talking sequence tracked by DRMF

3.5 Feature Extraction

The image after warping is not directly used for classfication, image feature are stelected to represent the image. Points, edges, ojects and texture are important features of an image. In this project, Local Binary Pattern are used for calssification as it's a very powerful feature for texuture classfication.

3.5.1 Local Binary Pattern

LBP is chosen to be the feature for representing region of interest. [8] obtained best recognition result by using Support Vector Machine with Boosted-LBP features. Moreover, [8] shows LBP features perform stably and robustly on low-resolution face images. In the beginning, LBP was used for texture analysis, it has natural advantage on computational simplicity and ignoring illumination changes.

3.6 Postprocessing

As it is known large margin classifiers are sensitive to the way features are scaled, it's better to normalize either the data or the kernal function [2]. Feature of a image is represented by a vector, the number in the vector would influence the weight of feature in this dimension. As I would like to treat each dimension similar, I scale the number in the range of $[0, 1]$.

Normalization The perfromance of SVM is usually better if the data is normalized. There are two ways of applying normalization, standardlizing the input features or normalizing the kernal function. As I am using the builtin function of libsvm [3], so I standardlising the input features by substracting its mean and divide by its standard deviation.

Scaling The range of appearance feature vector and shape feature vector is different. I would like to treat then as the same. So I scale all the vector into the range of $[0, 1]$, by substract the minmun and divide by the maximum number of each dimension.

Chapter 4

Experiment and Results

4.1 Data

There are 450 videos in two formats, avi and flv, 89 of them are flv videos, 361 of them are avi videos. The time of videos are from seconds to dozens of seconds. Videos were recorded using web-cam of different PCs. Video background are vary as the video are recorded in a place chosen by subject. People are free to do anything while recording the video, I even observe a male disappear from the camera for half of the video sequence while recording. As a result, there are no faces in most of the video frames. The Intraface tracker [9] is used to track those videos. It is able to track 439 videos. 1 video is tracked, but the tracking result does not match the label. 10 videos are tracked, but unable to identify the face in the video. The face in those untracked video can be clear identify by visual. One possible may be the resolution of the image. Maximum untracked video frame is $240 * 960$ pixels. Minimum tracked video frame size is $240 * 960$, the same as the maximum size of untracked video frame. It is reasonable to say tracker [9] may not be good at tracking low resolution videos. The observation mentioned in comparing tracker [9] and tracker [1] also support this hypothesis. The maximum frame tracked is $600 * 2400$. For each video there is a label file indicate the label of each frame, The frame number of each video is quite different, from around 200 to more than 900. The frame rate is 30 fps. Table4.1 shows the tracking result of tracker [9].

Label / Title	Normal Face	Eating	Talking	Looking Away	Occluded	Other Problem
Total	35361	10409	5623	7730	21394	8422
Tracked	33776	9460	5196	5405	19014	3884
Rate	0.9552	0.9088	0.9241	0.6992	0.8888	0.4612

Table 4.1: Frame tracking result by tracker [9]

There are six labels for each frame, normal face, eating, talking, looking away, occluded, other problem. A frame labelled as eating belong to a image sequence of eating. Label normal face, eating, talking and looking away are disjoint, but one frame can be labelled as one of them and occluded or other problem. It is very hard to track a face that face to camera from a certain angle, so the tracking rate is very small for a face looking away. I only use frame with label normal face, eating, talking in the experiment. Not all three labels are included in all videos, most of videos miss one or two or even all three labels.

4.1.1 Feature

Each face is aligned with 49 facial feature points as shown in figure 3.4. As the tracker doesn't provide face bound point, so I am unable to include jaw and cheek in the ROI. I just kept the mouth as Region of Interest. I extract Local Binary Pattern feature from ROI. The size appearance feature vector are different if I divide image of ROI into different number of blocks. I tried dividing ROI into 1 block and $1 * 3$ blocks, size of appearance feature vector is 95 and 177. As there are 49 shape

feature points, so the size of shape feature points is 98.

4.2 Methodology

In the classification part, I just tried to use Support Vector Machine to do classification. I tried two different type of appearance feature vectors, their dimensionality is 95 and 177, to see whether with more detailed appearance feature vector would be better for classification. I would like to know whether apply normalisation to each video would improve the classification. I tried two different feature vectors, one is normalized by each video and then normalized together, the other is just normalized together. So there are 10 groups of experiment.

Normalisation (True if By Video)	T	T	T	T	T	F	F	F	F	F
Appearance Feature (Divide by 1 or 3 Block)	1	1	3	3	1/3	1/3	1	1	3	3
Feature A: Appearance S: Shape	A	A+S	A	A+S	S	S	A	A+S	A	A+S

Table 4.2: Experiments

I tried both linear kernel function and non-linear kernel, the Gaussian Kernel shows better result. Gaussian and polynomial kernels often leads to over-fitting in high dimensional database, while linear kernel is easier to tune because the only parameter that affects performance is the soft-margin constant [2]. The most important parameters for Gaussian Kernel is penalty parameter c and γ in equation 4.1.

$$K(x, x') = e^{-\gamma||x-x'||^2} \tag{4.1}$$

4.2.1 Find Parameter c and γ

A general way to find parameter c and γ is using cross-validation and grid-search. In n-fold cross-validation, first equally divide the data into n fold, leave out one fold of data as testing data and use other $n-1$ fold of data to train the classifier. Thus all the data is predicted once and the cross-validation accuracy is the percentage of data are correctly classified. Grid-search is try various pairs of c and γ and choose the one with best cross-validation accuracy. Grid search approach is very simply and the computational time is no more than advanced method. To shorten the time of grid search, it is better to search with a coarse grid and then proceed with a more specific search in the identified grid.

4.3 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] Asa Ben-Hur and Jason Weston. A users guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer, 2010.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

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

[5] Maja Pantic and Marian Stewart Bartlett. Machine analysis of facial expressions. 2007.

[6] 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.

[7] 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.

[8] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[9] 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.