Realistic and Expressive Talking Head: Implementation and Evaluation

Der Fakultät für Elektrotechnik und Informatik

der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades

Doktor-Ingenieur

genehmigte

Dissertation

von

Kang Liu, M.Sc.

geboren am 10. Oktober 1977 in Sichuan, V.R. China

2011

Referent:	Prof. Dr.-Ing. Jörn Ostermann
Korreferent:	Prof. Dr. rer. nat. Volker Blanz
Vorsitz:	Prof. Dr.-Ing. Hans-Georg Musmann
Tag der Promotion:	08.06.2011

Acknowledgments

This work is funded by German Research Society (DFG Sachebeihilfe OS295/3-1). This work has been partially supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

I am greatly indebted to Prof. Dr.-Ing Jörn Ostermann, my supervisor and my mentor, for not only providing invaluable guidance, advice, criticism and encouragement, but also giving me the latitude I have needed to develop as a researcher. His vision and his leadership have largely contributed to making this project a success.

I would like to thank Prof. Dr. rer. nat. Volker Blanz for agreeing to serve as my second reviewer. I would also like to acknowledge Prof. Dr.-Ing. Hans-Georg Musmann, the chair of my PhD defense committee, who has introduced me to this research Institute.

My stay at the Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover, has been priceless and unforgettable. During the doctoral studies, I have had the pleasure to work with a large number of excellent researchers and skilled engineers. I want to thank Prof. Dr.-Ing. Bodo Rosenhahn for his constructive suggestions and the sharing of his knowledge. I must also remember Axel Weissenfeld, with whom I have worked on the facial animation project together for four years. Also, I wish to thank Andrew Manglos and Oktavia Ostermann for revising the English of my manuscript.

I am very thankful to my colleagues and friends: Bernd Edler, Martin Pahl, Ingolf Wassermann, Hanno Ackermann, Thorsten Thormählen, Sergej Piekh, Sven Klomp, Christian Becker, Torsten Büschenfeld, Tobias Elbrandt, Marco Munderloh, Song Hak Ri, Nikolce Stefanoski, Yuri Vatis, Minh Phuong Nguyen, Kai Cordes, Laura Leal-Taixé, Michele Fenzi, Zhijie Zhao, Mianshu Chen... Special thanks are given to Matthias Schuh and the secretaries, Mrs. Scholl, Mrs. Kemner and Mrs. Göring for their invaluable technical supports and work assistance. I would also like to thank all the people involved in the subjective tests for my thesis.

Last but not least, I would like to recognize my family members: my wife Xiao-Lu and my parents. Without their understanding and support, none of this would have been possible. I cannot begin to thank them enough, and they will always have my respect and love. I dedicate this thesis to my three little monsters: Rong-Rong (8 months), Qi-Qi (3 years) and Le-Le (6 years).

Kurzfassung

Die Kombination von Gesichtsanimation und Sprachsynthese ermöglicht innovative Mensch-Maschine-Schnittstellen, z.B. für Internetversandhäuser oder web-basierten Kundenservice. Mit bildbasierten Methoden lassen sich realistischere Animationen erstellen, als es mit 3D-Modellierung möglich ist. Ein bildbasiertes Gesichtsanimationssystem besteht aus zwei Teilen: Analyse und Synthese. In der Analyse wird eine Datenbank erstellt, die normalisierte Mundbilder und die dazugehörigen Phoneme enthält. Die Synthese generiert aus dieser Datenbank mit Hilfe von phonetischen Beschreibungen des zu sprechenden Textes natürlich wirkende Gesichtsanimationen.

Eine wichtige Aufgabe der Synthese ist die richtige Auswahl geeigneter Mundbilder aus der Datenbank für den auszugebenden Satz. Der Algorithmus zur Auswahl muss sowohl Lippensynchronisation als auch einen natürlichen Übergang zwischen aufeinander folgenden Bildern gewährleisten. Das globale Optimum des Algorithmus wird mit Hilfe des Pareto-Verfahrens gefunden. Die auf diese Weise erzeugte Gesichtsanimation erhielt den Golden Lips Award for Audiovisual Consistency des ersten "Visual Speech Synthesis" Wettbewerbs auf der "LIPS 2008".

Aufbauend auf dem zuvor beschriebenen Algorithmus wurde die Animation um einen realistischeren Gesichtsausdruck erweitert. Der Gesichtsausdruck wird durch den eingegebenen Text und zusätzliche Kontrollmarken gesteuert. Dafür werden drei Videos aufgenommen: im ersten zeigt der Sprecher keine Emotionen, in den anderen beiden lächelt er während bzw. nach dem Sprechen. Aus den unterschiedlichen Gesichtsausdrücken wird eine Datenbank erstellt, welche die Mundbilder mit und ohne Emotion und die entsprechenden Merkmale enthält. Um einen realistischen Gesichtsausdruck zu synthetisieren, werden die Viseme-Übergänge während des Wechsels des Gesichtsausdrucks analysiert. Experimentelle Ergebnisse zeigen, dass ein Betrachter nicht zwischen echter und synthetischer Sequenz unterscheiden kann.

Zusätzlich wurde ein neuer Ansatz für flexible Kopfbewegung entwickelt. Verschiedene Kopfbewegungen wie Nicken oder Kopfschütteln werden in der Datenbank gespeichert. Die Bewegung wird dann durch Auswahl der in der Datenbank gespeicherten Kopfbewegungen passend zu Text und Kontrollmarken zusammengesetzt. Um einen fließenden Übergang zwischen verschiedenen Kopfbewegungen zu erhalten, wird an den Übergangsstellen ein Morphing-Algorithmus basierend auf optischem Fluss verwendet. Experimentelle Ergebnisse zeigen, dass Animationen mit flexiblen Kopfbewegungen bessere subjektive Bewertungen erhalten.

Stichworte: Gesichtsanimation, bildbasierte Animation, Auswahl des Mundes, Pareto-Optimierung, Gesichtsausdruck, Kopfbewegung, optischer Fluss, Morphing.

Abstract

Facial animation has been combined with text-to-speech synthesis to create innovative multi-modal interfaces, such as online stores and web-based customer services. Using image-based rendering, facial animations look more realistic than facial animations generated by using 3D models. An image-based facial animation system consists of two parts: analysis and synthesis. The audio-visual analysis part creates a database, which contains a large number of normalized mouth images and related information. The synthesis part generates natural looking facial animations from phonetic transcripts of text.

An essential issue of the synthesis is the unit selection, which selects and concatenates appropriate mouth images from the database such that they match the spoken words of the talking head. Selection is based on lip synchronization and the similarity of consecutive images. The unit selection is optimized by the Pareto optimization algorithm which globally finds optimal weights. Our talking head received the Golden Lips Award for audiovisual consistency in the first visual speech synthesis challenge "LIPS2008".

Based on the optimized unit selection, the talking head is extended with realistic facial expressions, which is driven by arbitrary text input and control tags of facial expression. As an example of facial expression primitives, smile is used. First, three types of videos are recorded: a performer speaking without any expressions, smiling while speaking, and smiling after speaking. By analyzing the recorded audio-visual data, an expressive database is built. It contains normalized neutral mouth images and smiling mouth images, as well as their associated features and expressive labels. In order to synthesize realistic and smooth facial expressions, natural expression change and viseme transitions are analyzed while changing expressions. Experimental results show that the synthesized smiles are as realistic as the real ones, and the viewers cannot distinguish real smiles from synthesized ones.

In addition, a novel approach to add flexible head motions to talking heads is developed. First, head motion patterns are collected from original recordings. These head motion patterns are recorded video segments with different head motions, like nod and shake. The head motion is synthesized by selecting and concatenating appropriate head motion patterns according to the input text with head motion tags. In order to join these patterns, optical flow based morphing is used to smooth transitions without creating noticeable discontinuities. Experimental results show that animations with flexible head motions are rated with a higher average mean opinion score than the ones with repeated head motions.

Keywords: talking head, image-based animation, unit selection, Pareto optimization, facial expression, head motion, optical flow, morphing.

Contents

Formula Symbols and Abbreviations

Formula Symbols:

(R_x, R_y, R_z)	rotation vector
(T_x, T_y, T_z)	translation vector
α	interpolation parameter of morphing
$\bar{u}_k$	mean score
$\beta_1, \beta_2, \beta_3$	coefficients for unit selection
Δt	time interval of the sampling rate
δ_k	confidence interval
γ	exponential factor for phoneme distance matrix calculation
$\overline{g}$	average face texture
$\overline{PCA}$	average PCA weight
$\overline{x}$	average face shape
θ	rotation
a_R	acceleration of rotation
a_T	acceleration of translation
b_g	texture parameters
b_s	shape parameters
CC_{motion}	concatenation cost of two patterns
D	dimension of feature vector space
d	index of feature vector space U
D_{ij}	distance of two frames in the pose space
dx, dy	motion displacement
dx', dy'	motion vector of backward warping
f	visual cost
f_{PCA}	Euclidean distance of mouth images in PCA space
F_{TP}	mismatch rate of turning points
g	face texture
g_s	skip cost
i	frame index
I_α	interpolated image
K	reduced dimension of the PCA space of mouth images
k_d	weight for feature vector U
M	phoneme distance matrix
m_{y_a}, m_{y_b}	mean of series, y_a and y_b
n	index of frame number

N_{v_i}	number of mouth images of viseme v_i
p	frame difference
$p(x)$	Hermite polynomial
$p = [t_x \quad t_y \quad s \quad \theta]$	pose parameters
P_g	texture eigenvectors matrix
p_i	candidate image of state i
P_s	shape eigenvectors matrix
P_u	feature vector of candidate phoneme u
p_{v_i,v_j}	switching ability from viseme v_i to viseme v_j
ph_i	phoneme i
r	cross-correlation
$R_{closures}$	match rate of closures
R_{TP}	match rate of turning points
s	shape size
S_k	standard deviation
S_{i,p_i}	candidate image of frame i
$SL_{avg.}$	average segment length
T_i	feature vector of target phoneme i
t_x	translation in x
t_y	translation in y
$TC_{avg.}$	average target cost
U	feature vector
u	candidate index
$u_{i,k}$	score of observer i for video k
v_i	weight of frame i
v_R	speed of rotation
v_T	speed of translation
$VC_{avg.}$	average visual difference
w_a	discrimination factor for acceleration
w_i	skip cost
w_s	diagonal matrix of weights for each shape parameter
w_v	discrimination factor for speed
w_{rt}	scale factor
wcc	weight for concatenation cost
$wccf$	weight for visual cost
$wccg$	weight for skip cost
wtc	weight for target cost
x	face shape
x_0, y_0	point in image I_0
x_1, y_1	point in image I_1
y_a	the first component of PCA or mouth height of real sequence
y_b	the first component of PCA or mouth height of animated sequence

Abbreviations:

3D	3 Dimension
AAM	Active Appearance Models
ATSC	Advanced Television Systems Committee
CC	Concatenation Cost
CIF	Common Intermediate Format
CIR	Correct Identifying Rate
FACS	Facial Action Coding System
FIDM	Face Image Distance Measure
GA	Genetic Algorithm
GPA	Generalized Procrustes Analysis
HMM	Hidden Markov Model
ICA	Independent Component Analysis
LBG	Linde, Buzo, and Gray (Vector Quantization Algorithm)
LCE	Local Consistency Error
LIPS2008	the First Visual Speech Synthesis Challenge
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MSE	Mean Square Error
NCI	Number of Correctly Identified utterances
NTU	Number of Testing Utterances
PC	Path Cost
PCA	Principal Component Analysis
PSNR	Peak Signal to Noise Ratio
SAPI	Microsoft Speech Application Programming Interface
TC	Target Cost
TP	Turning Point
TTS	Text-To-Speech

1 Introduction

In recent years, the development of modern human-computer interfaces and their applications such as E-Learning, web-based information services and video games has been the focus of the computer graphics community [1] [2] [3] [4]. Talking heads become more and more common as parts of modern computer-user interfaces. A talking head can add entertainment value to a program and make it more engaging, like E-Learning. These applications will use facial animation techniques combined with dialog systems extensively in the future [5] [6].

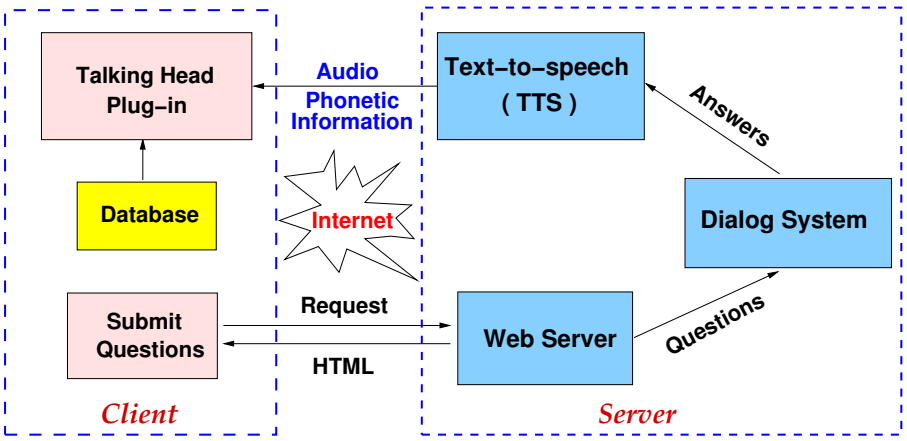


Figure 1.1: Schematic diagram of web-based information service with a talking head [7].

Fig. 1.1 shows a typical application of a talking head for a web-based information service. If the information service website is visited by a user, the talking head will start a conversation with the user. The user is warmly welcomed to experience the website. The dialog system will answer any question the user asks and send the answer to a TTS (Text-To-Speech Synthesizer). The TTS produces the spoken audio track as well as the phonetic information and their duration, which are required by the talking head plug-in embedded in the website. The talking head plug-in selects appropriate mouth images from the database to generate a video. The talking head will be shown in the website after the right download and installation of the plug-in and its associated database [8]. Fig. 1.2 shows a snapshot of a talking head embedded in a Newsreader website.

In these applications, the talking heads could either be recorded video sequences of real humans or cartoon characters that are synthesized in real time. Recorded video sequences are expensive to produce, require a lot of storage space, and limit the flexibility of the

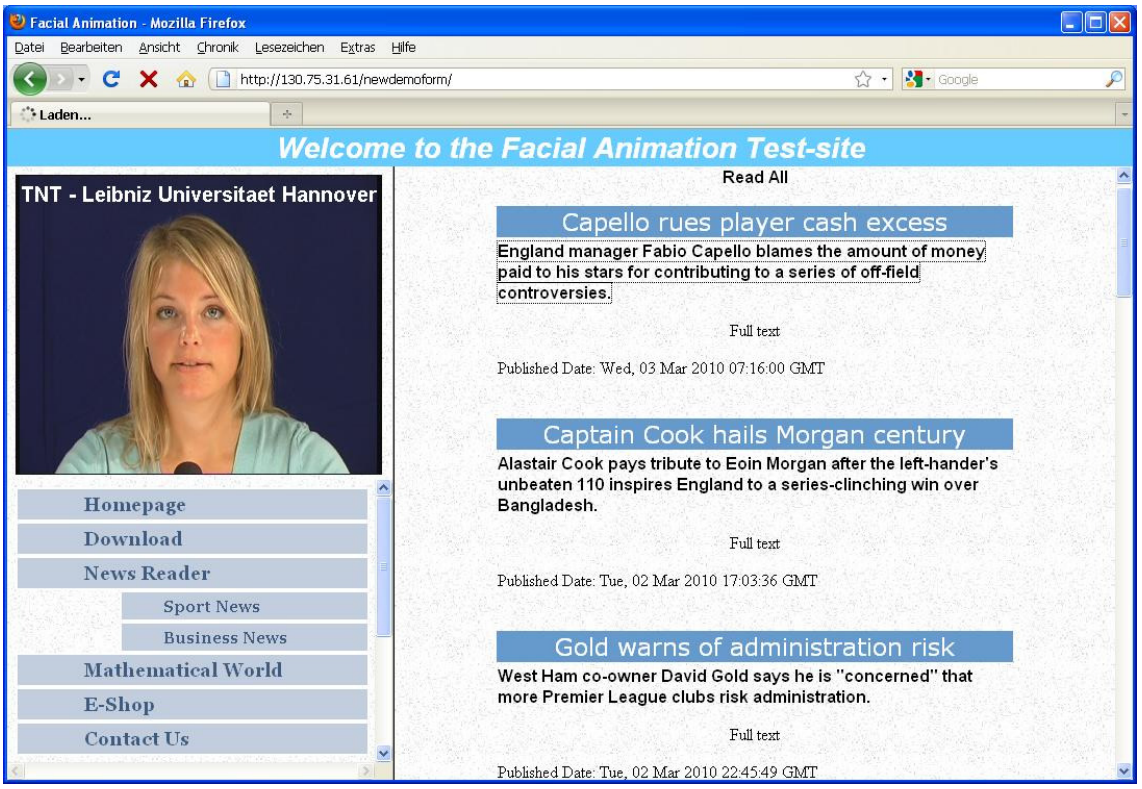


Figure 1.2: Snapshot of a Newsreader website with a talking head.

interface. Cartoon characters are suited for some applications because they can be entertaining, but often a photo-realistic talking head is more appropriate. The talking head has to be not only photo-realistic in appearance, but must also behave with realistic head movements and emotional expressions. Humans are very sensitive to the slightest facial changes, so that synthesizing a realistic talking head is still a very challenging task.

Each deformable facial part should be considered to be animated in order to achieve a natural face. These deformable parts include mouth, eyes, eyebrows, hair, forehead, cheeks, etc. The mouth, eyes, and head motion are ones of the most focused facial parts in the face-to-face communication, since the verbal information is delivered through speech and the non-verbal information is given by eyes, eyebrows, head motion, and facial expression.

It is difficult to classify facial modeling and animation techniques strictly, because recent approaches often integrate several methods to produce better results. Nevertheless, we can roughly classify facial animations into 3D model-based animation and image-based animation. Three example talking heads [5] are shown in Fig. 1.3.

3D models are based on 3D meshes, which are very flexible for generating movements and showing the head in any desired orientation. However, in order to render highly deformable facial parts, such as a mouth with a high degree of precision, complex models

Figure 1.3: Example talking heads [5]. (a) MPEG-4 standard 3D talking head, (b) a texture mapped talking head, and (c) an image-based talking head.

that are computationally expensive and generally produce faces with a synthetic look are needed. Furthermore, the movements of the 3D models are not easy to be parameterized, since the facial movements, like head motion and facial expression, are nonlinear while speaking.

An alternative approach, image-based rendering, uses recorded videos to synthesize a new face. This approach is based on recorded samples, and is able to produce a talking head which utters text the person never actually said. For that purpose, the image-based approach requires a database which contains many samples that describe the different shape and appearance of the mouth. However, this approach cannot synthesize a new face, whose facial parts are not available in the database. Despite the large number of samples, the image-based approach cannot currently handle emotional expressions. In Fig. 1.3, the hair is very roughly modeled by a 3D model, but the image-based model can directly use the recorded head to achieve realistic animations. In addition, the teeth and tongue of 3D model are not modeled as realistic as the real person due to the complex shape and appearance.

The current image-based talking head [9] [10] is so realistic that people cannot distinguish them from real videos. However, the talking head is inexpressive. The user will become bored if the talking head shows no expression for a long conversation, or the head only moves with a simple motion. Therefore, an engaging talking head is necessary for these applications, and a lively talking head makes the users trust more in human-computer communication [5] [6]. Furthermore, a context-aware talking head should be able to make the users perceive that the talking head understands them during the conversation. For this purpose, this thesis tackles the problem of synthesis of facial expression and head motion for an image-based talking head, which is driven by input text and control tags, such as smile and nod.

1.1 Previous Work

The generation of talking heads can be divided mainly into two parts: lip synchronization and talking head rendering (Fig. 1.4). The inputs of the system are speech features such as phonetic information from speech recognition or TTS (Text-To-Speech). The first part is lip synchronization, which determines visual parameters from speech features for mouth animations. Talking head rendering uses the visual parameters to generate animations via 3D model-based, physics-based or image-based approaches [6] [11] [12].

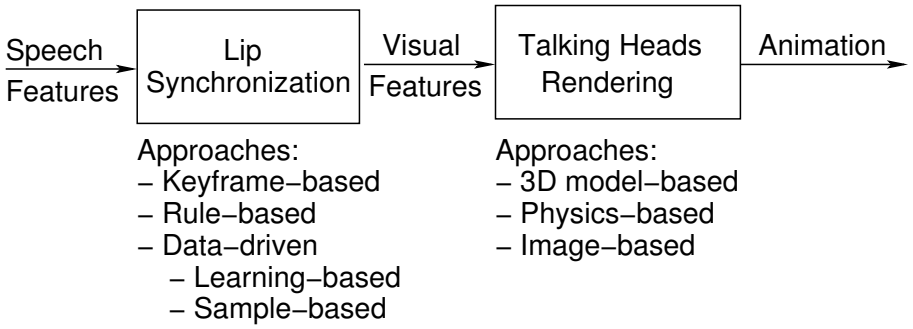


Figure 1.4: Block diagram of talking head synthesis.

3D model-based animation usually requires a mesh of 3D polygons that define the head shape, which can be parametrically deformed to perform facial actions. A texture is mapped onto the mesh to render facial parts. Such a facial animation has become a standard in ISO/IEC MPEG-4 [13]. A more complex approach is to model the face by 3D morphable models [14] [15] [16]. Additionally, eyes and teeth are also modeled in [17]. These techniques allow the creation of real surfaces, texture and motion for performance-driven animations. However, the relationship between speech and mouth movement is not well investigated.

Physics-based animation [18] [19] has an underlying anatomical structure so that the model allows a deformation of the head in anthropometrically meaningful ways [20]. The facial action coding system (FACS) is a description of the movements of the facial muscles and jaw/tongue derived from an analysis of facial anatomy [21]. The main drawback of using FACS is that FACS offers spatial motion description, but not temporal components. Furthermore, the muscular system is highly redundant and movements typically require a few dozen individual muscles whose actions need to be coordinated, sometimes in a very precise way. The FACS was designed to analyze facial expressions for the psychological study of facial behavior, however, it is not suitable for realistic talking head applications.

In addition to the above mentioned approaches, a series of image-based approaches have emerged in the past decade. We will categorize them in three families of systems: (a) systems that select appropriate segments of a large database and overlay the facial parts on

a background image [9] [10] [22]; (b) systems that generate new faces based on morphing techniques [23] [24]; (c) systems that use AAM (Active Appearance Models) [25] to model the shape and appearance and create a realistic talking face [26].

Besides mouth animation, eye animation is also important for nonverbal communications [27]. Using the image-based approach, Weissenfeld et al. [28] categorize the eye movements into two modes: listening and talking. For each mode, a Markov model is designed and the animations are as realistic as real ones.

1.1.1 Lip Synchronization

The goal of lip synchronization is to control the mouth movements which match the corresponding audio utterance as realistically as possible.

Keyframe-based rendering interpolates the frames between keyframes. The basic visemes are defined as keyframes and the transition in the animation is based on morphing visemes in [23]. A viseme is the basic mouth shape corresponding to the speech unit phoneme. For example, the phonemes "m", "b", "p" correspond to the closure viseme. However, this approach does not take into account the co-articulation models [29] [30], which indicate that a particular mouth shape depends not only on its own phoneme, but also on its preceding and succeeding phonemes.

Rule-based lip synchronization defines a mapping between the visemes and the visual control parameters [29] [31] [32]. In order to account for co-articulation effects, some visemes are left undefined and their visual control parameters are dependent on the phoneme contexts. Animations generated by these rule-based approaches largely depend on these defined rules. However, in practice, constructing accurate co-articulation functions and mappings between the visemes and visual control parameters requires a lot of painstaking manual effort. Data-driven approaches are proposed to alleviate these manual efforts.

Data-driven approaches synthesize new speech animations by concatenating prerecorded facial motion data or by sampling from statistical models learned from the data. First, facial motion data is pre-recorded. Then, a facial motion database can be built in two different ways: either statistical models for facial motion control are trained from the data (learning-based approaches), or the facial motion database contains recorded samples and related parameters (sample-based approaches). Finally, given a novel sound track (speech-driven) or text input (text-driven), corresponding visual features are generated by sampling from the trained statistical models, or by recombining recorded frames optimally selected from the facial motion database. Data-driven approaches can generate almost realistic facial animation results, but these approaches do not provide intuitive controls for animators.

Learning-based approaches model speech co-articulations as implicit functions in statistical models. HMMs (Hidden Markov Models) are used for lip motion estimation. The problem changes to estimating the missing visual features (such as mouth width and height) based on trained HMMs and given speech features (such as phonetic information). In [33] [34], an HMM-based facial control model is learned by using joint audio-visual observations. These approaches use the Viterbi algorithm through the HMMs to search for the most likely facial state sequence that matches the target speech features.

Sample-based approaches use the facial motion database for synthesizing talking faces given novel speech input. These approaches model the co-articulation effect implicitly. Bregler et al. [22] present “video rewrite" based on collected triphone video segments, which require a large database to cover the triphone samples. Cosatto et al. [35] and Liu et al. [10] extend the triphone to phoneme-based segments. Liu and Ostermann [10] improve the lip synchronization through Pareto optimization [36] [37]. Their talking head system can achieve realistic mouth animations, which are indistinguishable from real videos in the subjective test [10]. However, facial expressions are not studied in the sample-based system.

1.1.2 Facial Expression Synthesis

Facial expressions reflect one’s motivation or emotional states, which is an important aspect in communication. Methods for modeling facial expressions have been largely investigated with a few attempts to achieve realistic expressions synchronized with speech.

According to [38] [39], all facial expressions can be blended by six basic expressions: happiness, sadness, surprise, fear, anger and disgust. An intuitive method to drive facial expressions is the mesh animation in MPEG-4 [13] [40] [41]. The facial expressions are parameterized by a set of Action Units, which define the facial local motion in different regions on the face, such as eye, brow and mouth. Most of the effort has gone to the tracking and recognition of facial expressions [42] [43] [44], which utilize the static or short-term dynamics of action units for facial expressions.

Instead of deforming the 3D mesh, Pighin et al. [14] present a facial expression modeling by combining 2D and 3D techniques, which can synthesize new facial expressions by interpolating between two or more static facial expressions. The quality of these animations have significantly improved, especially through more sophisticated 3D models and new texture mapping methods in recent years. Morphing static facial expressions looks surprisingly realistic nowadays, while a realistic talking head (animation with synchronized audio) is not possible yet.

However, the long-term dynamics of an expressive talking face is important to model realistic facial expressions. Cao et al. [45] [46] apply ICA (Independent Component Analysis) to decompose facial motion into emotion style and speech content. The mapping between emotion spaces is trained. Using a 3D model, the system can first generate an animation without expressions according to the input audio. Then, an expressive anima-

tion is created using the trained mappings between emotion spaces. This approach does not consider transitions when the emotion changes in one sentence, which sometimes results in unrealistic animations.

The work in [47] tries to model the facial expressions using AAM, which separates the speech and expression subspaces. However, a lot of manual work is needed to build the model. Moreover, the blend shape results in blurred mouth animations.

Unit selection synthesis can generate realistic image-based animations by concatenating recorded mouth images in an appropriate order. Using a large database with a large number of units available with different appearance, shape and expression, it is possible to synthesize more natural looking facial animations than 3D model-based approaches, because the dynamics of the lips and tongue are difficult to be modeled parametrically. Experiments on the expressive synthesis indicate that the animations depend on the audio-visual database. Facial expressions [9] [48] are represented as template behavior (facial expression patterns), for example, a welcoming smile. These patterns are appended to an expressionless facial animation, which results in an unrealistic talking head, because expressions cannot be modeled by only using the patterns. Furthermore, the transitions between facial expressions in these animations are not optimized, so that the talking head looks strange and unnatural, even though the mouth movements are smooth. Moreover, expressions accompanying speech are not investigated.

1.1.3 Head Motion Synthesis

In [49], it is reported that the head motion correlates strongly with the pitch (fundamental frequency) and amplitude of the speech. The results suggest that non-verbal gestures such as head movements play a more direct role in the perception of speech than previously known. Appropriate head motion can significantly enhance human-computer interfaces. Head motion is extensively synthesized based on 3D models. However, it is not yet synthesized by using sample-based approaches. Therefore, to synthesize engaging talking heads, special attention needs to be given to appropriate speech synchronized head motions of the virtual characters.

Like mouth animations, head motion synthesis can be explained in two steps. Fig. 1.5 shows the synthesis of head movements.

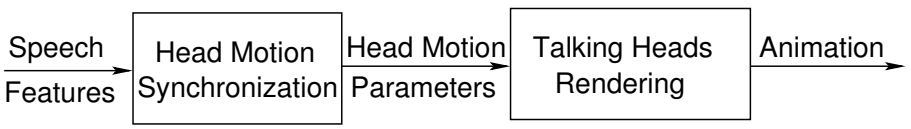


Figure 1.5: Block diagram of head motion synthesis.

Various techniques are presented to compute head motion parameters. These techniques can be categorized into two classes.

The first is parameter-driven head motion synthesis, also known as rule-based head motion animation, such as the system presented in [50] [51] [52]. If the emotion of the animation is specified, the velocity and the global pose of the head motion are generated using predefined rules, which map the labeled text to head motion. For example, the global pose was set with downward direction for the expression of sadness.

The second is data-driven head movement synthesis. In [53], head motion is classified into three basic patterns. These are nod, nod with an overshot at the return and abrupt swing of the head in one direction. Patterns of head and facial movements are strongly correlated with the prosodic structure of the text. Head motion is synthesized by concatenating the three basic motion patterns in an appropriate order according to the conditional probability distribution given the occurrence of pitch accents.

Chuang et al. [47] present an approach that synthesizes novel head motion from input speech. They first build a head motion database indexed by pitch value. Synchronized motion and audio streams are captured and then segmented and stored in the database. A new audio stream can be matched against segments in the database. A smooth path is found through the matching segments and stitched together to create synthetic head motion. The head motion parameters are used to generate 3D model-based animations.

Busso et al. [54] propose an HMM based head motion system. This approach models the specific temporal relation between emotional head motion sequences and prosodic features by building HMMs for each emotion, instead of generic models [55]. The head poses are quantized by LBG. In the synthesis, HMMs are used to generate quantized head motion sequences, which are smoothed using first order Markov models and spherical cubic interpolation.

Since a 3D model-based facial animation system can use the head motion parameters explicitly and directly, most current research has synthesized head motion based on a 3D model. However, it is difficult to generate realistic animations based on a 3D model, especially with hairs. In the literature, no head motion is synthesized by using image-based approach. In order to overcome this shortcoming, we present an image-based approach to produce animations with flexible head movements accompanying speech.

In [56], head motions are classified in both functions and forms. The selection of head motion is based on the relationship between the functions and the forms. However, the relationship does not consider the influence of other modalities such as speech, facial expression, gestures and posture. The head motions are related to the speech and the semantic meaning of the sentences. Current natural language understanding technology is not yet able to extract meaning. Therefore, in this thesis, the head motion form is determined manually by the input control tags, such as “nod” or “shake”.

1.2 Open Problems

For an expressive talking head, the animation of mouth, eye, facial expressions and visual prosody is required. Since solutions for mouth and eye animation exist, we focus our attention on the following problems in the context of an image-based talking head.

- Facial expression: The realism of facial expression of image-based talking heads is currently unsatisfying, since facial expression patterns are appended to the end of speech. Hence, facial expressions cannot synchronize with speech. Moreover, the influence of facial expressions on the movement of a talking mouth and its timing is currently not well understood.

- Head motion: Due to the use of a background sequence, the head motion of image-based talking head cannot be controlled. Furthermore, the transition between head motion sequences results in a noticeable jerky head motion. Image-based head motion synthesis is still not well investigated in the literature.

1.3 Solution and Contributions

Three main contributions for the realistic and expressive talking head are presented in this thesis:

- The reference system [35] is optimized, which has been presented in [7] [10]. In the analysis, the facial features are detected by an AAM-based approach [25] [57] (Section 2.3.1), which is robust and insensitive to illumination changes on the face resulting from head and mouth motions, compared to the template matching based approach in [58]. Another improvement is the Pareto optimization of the unit selection [36] [37] (Section 2.3.2), which can find a global optimal solution for a multi-objective problem.

- A smiling talking head is synthesized based on the unit selection from an expressive database, which consists of a large number of recorded neutral and smiling mouth images (Chapter 3). The expressive unit selection is used to select and concatenate mouth image segments from the expressive database in an optimal way. By using expression control tags, desired facial expressions are generated by appropriately switching between different expressions. This research work is published in [59].

- We introduce an approach to add flexible head motion to an image-based talking head (Chapter 4), which is presented in [60]. Head motion patterns are identified by analysis of the motion parameters of the recorded sequences. Head motion patterns are sequences with typical head motion for different states, such as speaking, listening, and idle states. In order to join these patterns, optical flow based

morphing [61] is used to smooth transitions without introducing noticeable discontinuities. The selection of head motion patterns is based on the input head motion control tags.

1.4 Thesis Structure

The remainder of this thesis is organized as follows.

In Chapter 2, a reference system of image-based talking head [35] is introduced. It consists of two parts: analysis and synthesis. Secondly, improvements of the reference system are presented, one is the reliable feature detection in the analysis part, and the other is the optimization of the unit selection in the synthesis part.

Chapter 3 introduces the technique to synthesize an expressive talking head. In order to generate expressive talking heads, an expressive database is required. The natural expression change of humans and the viseme transition of the expressive database are analyzed. Based on the analytical results, the unit selection is modified to select mouth images from the expressive database, by switching between mouth images with different expressions.

Chapter 4 describes the head motion synthesis. Image-based head motion synthesis requires a head motion database, which contains a lot of head motion patterns. These head motion patterns are recorded video segments with different head motions, like nod and shake. In the synthesis, head motion patterns are selected and concatenated according to the input head motion control tags. In order to smooth the transitions, an optical flow based morphing technique is used. In addition, the dynamics of the head motion between these patterns are analyzed.

Chapter 5 evaluates the realistic and expressive talking head. Thereby, facial expression and head motion are assessed objectively and subjectively. The evaluation results are published in [62].

The thesis is concluded in Chapter 6.

2 An Image-based Talking Head

In this chapter, the general approach to generating an image-based talking head is described in the first section. In Section 2.2, the reference system of an image-based talking head is introduced, which is divided into two parts: analysis and synthesis. In Section 2.3, the image-based talking head, based on the reference system, is improved. In the analysis, facial features are robustly and precisely detected using AAM. In the synthesis, the unit selection is optimized through Pareto optimization. Finally, the results of the optimized talking head are given in the last section.

2.1 General Image-based Talking Head System

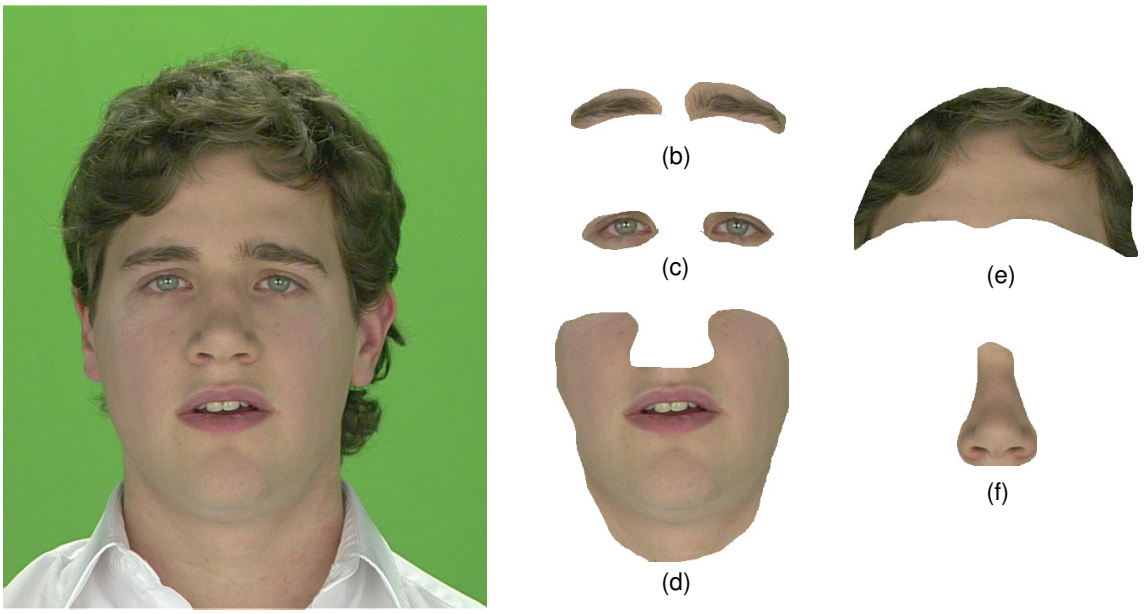


Figure 2.1: Image-based face model. (a) base face (recorded face image), basic layer: (b) eyebrows, (c) eyes, (d) mouth, cheeks and chin; optional layer: (e) forehead and (f) nose.

An image-based talking head is composed of independent image-based animations of several facial parts. The key aspect of this model is that it allows the image-based rendering of facial parts with multiple texture samples to be seamlessly merged into a whole

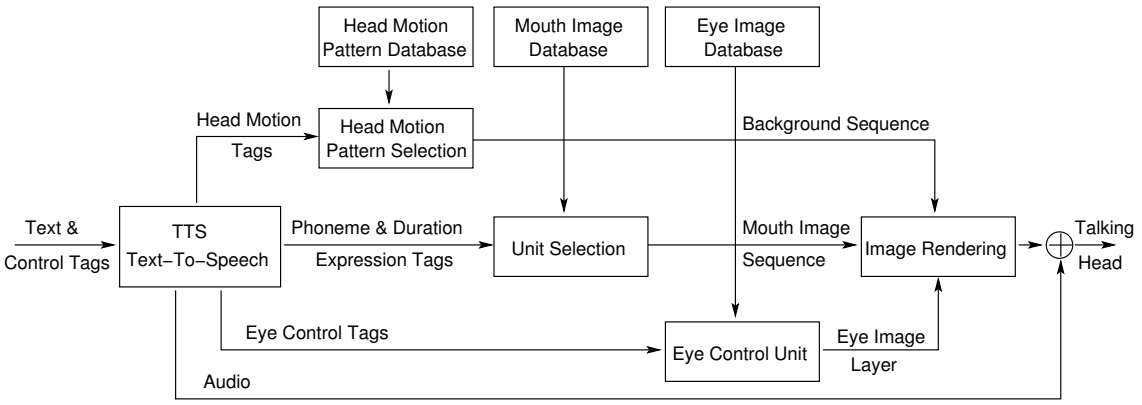


Figure 2.2: General system architecture of image-based talking head.

head. The head is separated into a basic layer and an optional layer. The basic layer includes several facial parts, such as eyes, eyebrows, and mouth, cheeks and chin. Since the movements of the mouth, cheeks and chin are tightly dependent, these three parts are segmented as one complete facial part. The optional layer includes the forehead and nose. Ears and hair are not modeled separately, but are part of a base face. The base face is a recorded face image, and also called "background image" in the synthesis. Fig. 2.1 shows the facial parts of the image-based face model.

Based on the above layered face model, an architecture for the synthesis of an image-based talking head is shown in Fig. 2.2. The inputs are text and control tags. The control tags include head motion tags for head motion synthesis, expression tags for facial expression synthesis, and eye control tags for eye animation. A TTS synthesizer converts the text into its phonetic information, as well as the spoken audio data. Depending on the face model, facial parts are selected from the databases. The databases are categorized into two layers: basic layer and optional layer. The basic layer includes mouth, eyes, eyebrows, and base face. The optional layer includes nose, forehead and hair. For the basic layer, three databases are built: the mouth image database for mouths, the eye image database for eyes and eyebrows, as well as the head motion pattern database for base faces. The selection of these facial parts is the main issue for the synthesis of a realistic talking head. The head motion pattern selection is to select the background sequences synchronized with the speech according to the head motion tags, such as nod and shake. Depending on the phonetic information and the expression tags, the unit selection is to select the appropriate expressive mouth images which coarticulate with the speech. The eye control unit is to control the eye motion related to the speech, generating non-verbal animations. The optional layer could be added to the system as required. For example, the wrinkle on the nose and forehead can be layered on the base head.

Once all facial parts have been located from the databases, the image rendering integrates all parts into the base face on the background image, so that a new face is generated. The new face and the audio are synchronized and a talking head is displayed.

2.2 Overview of the Reference System

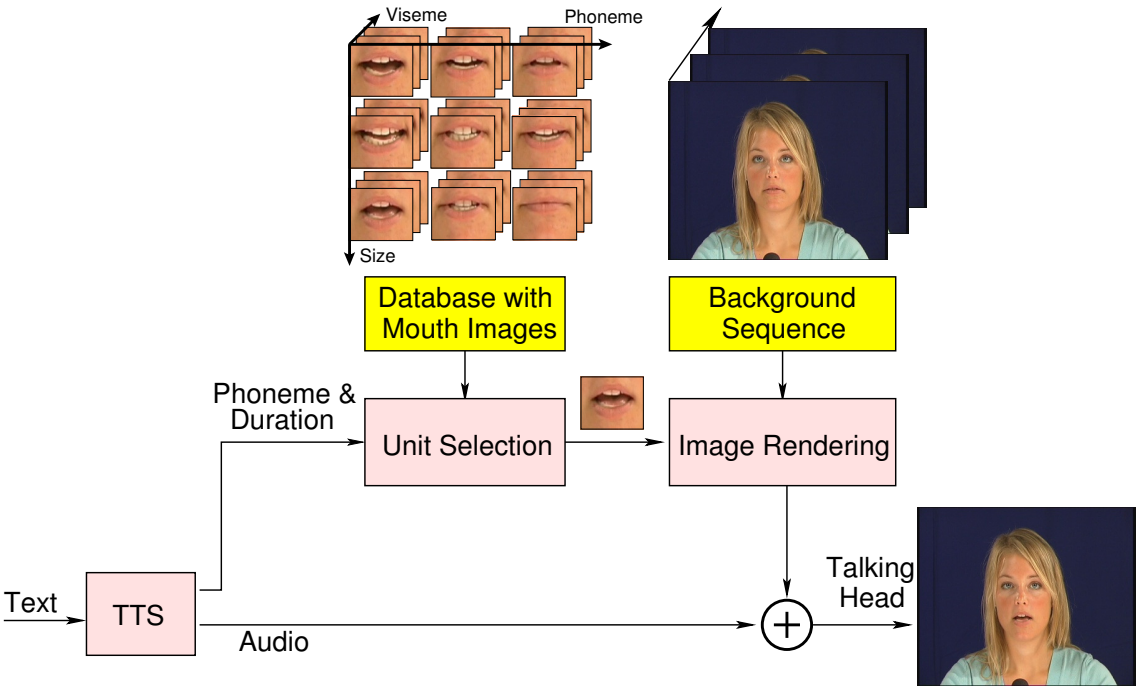


Figure 2.3: Reference system of synthesizing an image-based talking head.

The system presented in [35] is our reference system. The reference talking head system is depicted in Fig. 2.3. First, a segment of text is sent to a TTS synthesizer. The TTS provides the audio track as well as the sequence of phonemes and their durations, which are then sent to the unit selection. Depending on the phoneme information, the unit selection selects mouth images from the database and assembles them in an optimal way to produce the desired animation. The unit selection balances two competing goals: lip synchronization and smoothness of the transition between consecutive images. A cost function is defined for each goal, where both of them are functions of the mouth image parameters. The cost function for lip synchronization considers the coarticulation effects by matching the distance between the phonetic context of the synthesized sequence and the phonetic context of the mouth image in the database. The cost function for smoothness reduces the visual distance at the transition of images in the final animation, favoring transitions between consecutively recorded images. Then, an image rendering module stitches these mouth images to the background video sequence. The mouth images are first wrapped onto a personalized 3D face mask and then rotated and translated to the correct position on the background images. The wrapped 3D face mask is shown in the left image of Fig. 2.4. The right image of Fig. 2.4 shows the projection of the textured 3D mask onto a background image in a correct position and orientation. Background videos

are recorded video sequences of a human subject with typical head movements. Finally the facial animation is synchronized with the audio, and a talking head is displayed.

Figure 2.4: Image-based rendering. The left image is the 3D face mask with wrapped mouth and eye textures. The right image is a synthesized face by projecting the textured 3D mask (left image) onto a background image in a correct position and orientation. Alpha-blending is used on the edge of the face mask to combine the 3D face mask seamlessly with the background.

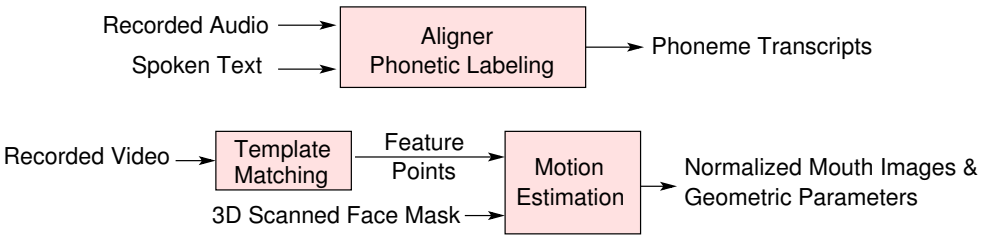
2.2.1 Analysis

The goal of the analysis is to build a database for a real time visual speech synthesizer. The analysis process is completed in two steps, as shown in Fig. 2.5. Step one is to analyze the recorded video and audio to obtain normalized mouth images and related phonetic information. Step two is to parameterize normalized mouth images. The resulting database contains normalized mouth images and their associated parameters.

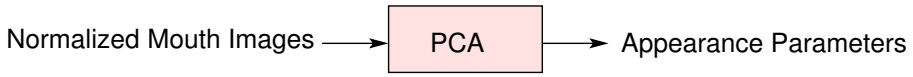
Audio-Visual Analysis

The audio-visual analysis of recorded human subjects results in a database of mouth images and their relevant features suitable for synthesis. The audio and video of a human subject reading a predefined corpus of texts are recorded. As shown in Fig. 2.5(a), the recorded audio and the spoken texts are phonetically labeled by the aligner, and the recorded video data is analyzed by the template matching and the motion estimation.

The recorded audio and the spoken text are processed by speech recognition to recognize and temporally align the phonetic transcription of the text to the recorded audio data. The process is referred to the aligner. Finally, the timed sequence of phonemes is aligned up to the sampling rate of the corresponding video. Therefore, for each frame of the recorded video, the corresponding phoneme and phoneme context are known. The phonetic context is required due to the coarticulation, since a particular mouth shape depends not only on its associated phoneme, but also on its preceding and succeeding phonemes.



(a) Audio-visual analysis of recorded human subjects



(b) Parameterization of the normalized mouth images

Figure 2.5: Database building by analyzing recorded human subject. (a) Analysis of recorded video and audio, (b) Parameterization of the normalized mouth images.

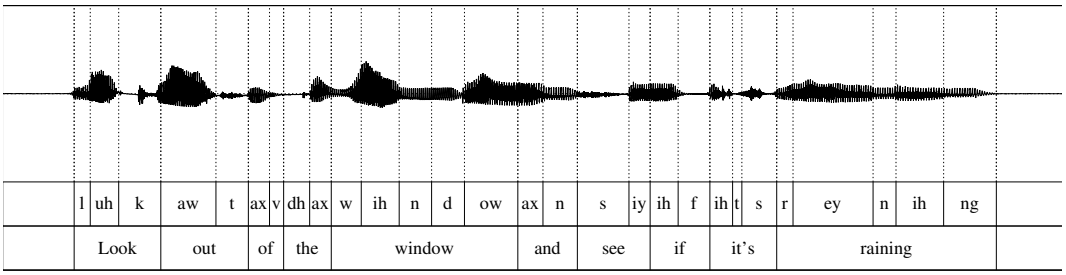


Figure 2.6: Phonetic labeling of audio data. The top row is the audio wave form of the sentence "Look out of the window and see if it's raining.". The middle row is the phonetic labeling of the audio data.

Table 2.1 shows the American English phoneme and viseme inventory that used in this thesis to phonetically transcribe the input text. The mapping from phoneme to viseme is based on the similarity of the appearance of the mouth. In our system, we define the mapping of 43 phonemes to 22 visemes using the American English Phoneme Representation of Microsoft Speech API (version SAPI 5.1). Fig. 2.6 shows an example of phonetic labeling of an audio data.

All image frames of the recorded videos are individually analyzed to extract and normalize individual facial features. The first step in obtaining normalized mouth images is to locate the facial feature points. Based on the collected templates, such as eye and mouth corners, these features are manually marked on the templates. A template matching approach tries to find the best template which is most similar to features in the input image.

To find the position and orientation of the head, the pose estimation technique [63] is used. A 3D face mask of the recorded subject is acquired using a Cyber scanner. Knowing

Table 2.1: Phoneme-viseme mapping of SAPI American English Phoneme Representation. 43 phonemes are classified into 22 visemes.

Viseme No.	Phoneme	Viseme No.	Phoneme
1	silence	12	ay
2	ae, ax, ah	13	h, hh
3	aa	14	r
4	ao	15	l
5	ey, eh, uh	16	s, z
6	er	17	sh, ch, jh, zh
7	iy, y, ih, ix	18	th, dh
8	w, uw	19	f, v
9	ow	20	d, t, n
10	aw	21	k, g, ng
11	oy	22	p, b, m

the values of both the 3D feature points and their corresponding 2D points in the image plane, the pose of the object can be inferred by the equations of perspective projection. This is known as the perspective-n-point (PnP) problem. Using the recovered rotation and translation parameters of the head, the face image is projected onto the 3D face mask. The textured 3D mask is moved to a normalized position, and is reprojected onto the image plane using bilinear interpolation. Thus, the resulting image is a frontal, normalized view of the facial part.

Parameterization of Normalized Mouth Images

Once the normalized mouth images are extracted and stored in the database, several features need to be computed and attached to the images for the unit selection in the synthesis.

PCA parameters are used to describe the appearance of mouth images as shown in Fig. 2.5(b). Reflecting the original data structure, we use PCA parameters to measure the distance of the mouth images in the objective criteria for system training, as PCA transforms the mouth image data into principal component space.

In addition to the appearance parameters, geometric parameters, such as mouth width and mouth height, are derived directly from the detected feature points in Fig. 2.5(a). Furthermore, the visibility of teeth and tongue is characterized for every mouth.

The sentence number and the frame number are used to preserve the inherent smoothness of real, recorded facial movements when selecting images for an animation.

For the background sequence, the rotation and translation parameters of the head are saved for each background image, such that, at synthesis time, the facial parts can be projected onto the background image with the correct orientation.

Database

All the parameters associated with an image are also saved in the database. Therefore, the database is built with a large number of normalized mouth images. Each image is characterized by geometric parameters (such as mouth width and height), texture parameters (PCA parameters), phonetic context, sentence number and frame number.

The size of the database is directly related to the quality required for animations. Better lip-synchronization requires more sequences and a higher image resolution makes larger files.

2.2.2 Synthesis

To synthesize a new animation, the input text is first run through a text-to-speech synthesizer (TTS) generating the audio track and its phonetic transcript [64]. The given video frame rate of animations, together with the length of the audio, determine the number of video frames to be synthesized. To achieve the synchronization of the mouth with the audio track while keeping the resulting animation smooth and visually pleasing, a technique of unit selection from concatenative speech synthesis [65] is adapted, which is detailed as follows.

Unit Selection

The unit selection chooses the mouth images corresponding to the phoneme sequence, using a target cost and a concatenation cost function to balance lip-synchronization and smoothness. As shown in Fig. 2.7, the phoneme sequence and audio data are generated by the TTS system. For each frame of the synthesized video, a mouth image is selected from the database for the final animation. The selection is executed as follows.

First, a search graph is built. Each frame is populated with a list of candidate mouth images that belong to the viseme corresponding to the phoneme of the frame. Using a viseme instead of a phoneme increases the number of valid candidates for a given target, given the relatively small database. Each candidate is fully connected to the candidates of the next frame. The connectivity of the candidates builds a search graph as depicted in Fig. 2.7. Target costs are assigned to each candidate and concatenation costs are assigned to each connection. A Viterbi search through the graph finds the optimal path with minimal total cost. Given in Fig. 2.7, the selected sequence is composed of several segments. The segments are extracted from the recorded sequence. Lip synchronization is achieved by defining target costs that are small for images recorded with the same phonetic context as the current image to be synthesized.

The Target Cost (TC) is the distance measured between the phoneme at frame i and the

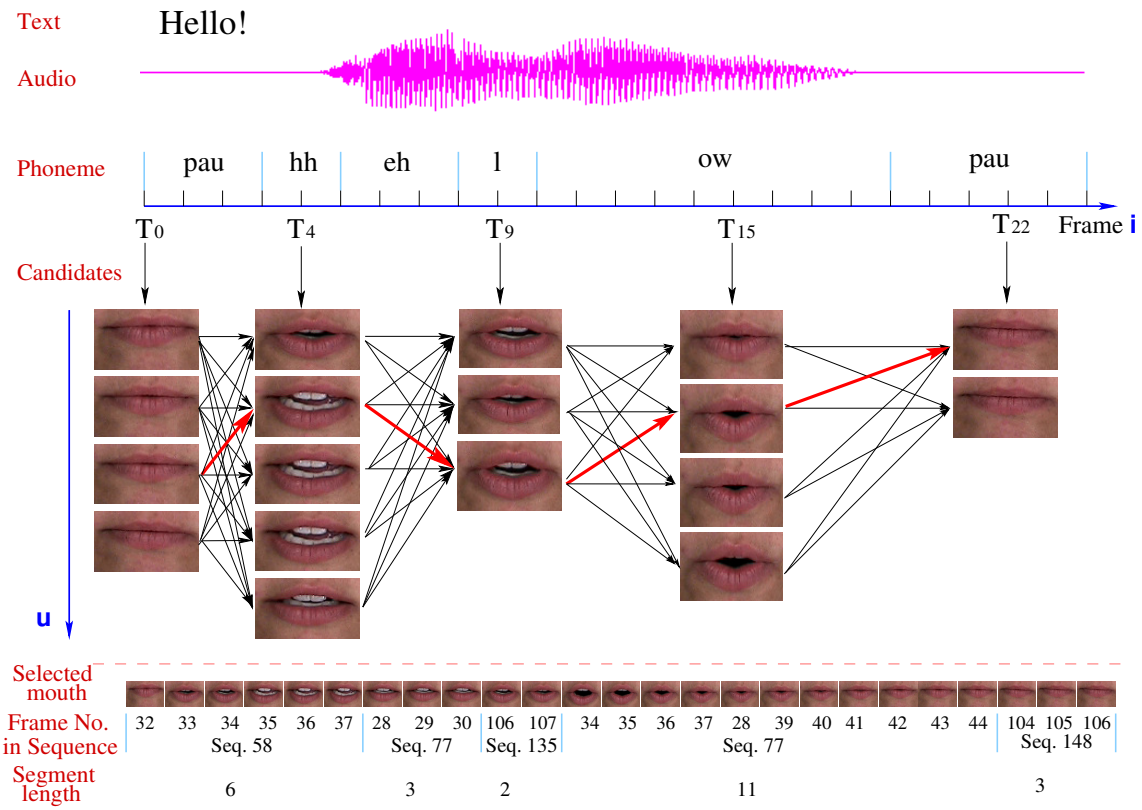


Figure 2.7: Illustration of unit selection algorithm. The text is the input of the TTS synthesizer. The audio and phoneme are the outputs of the TTS synthesizer. The candidates are from the database and the red path is the optimal animation path with a minimal total cost found by Viterbi search. The selected mouth images are taken from several original video segments.

phoneme of image u in the candidate list:

$$TC(i,u) = \frac{1}{\sum_{t=-n}^{n} v_{i+t}} \sum_{t=-n}^{n} v_{i+t} \cdot M(T_{i+t}, P_{u+t}) \tag{2.1}$$

where a target phoneme feature vector:

$$\overrightarrow{T_i} = (T_{i-n}, ..., T_i, ..., T_{i+n})$$

with T_i representing the phoneme at frame i, a candidate phoneme feature vector:

$$\overrightarrow{P_u} = (P_{u-n}, ..., P_u, ..., P_{u+n})$$

consisting of the phonemes before and after the u^{th} phoneme in the recorded sequence and a weight vector:

$$\overrightarrow{v_i} = (v_{i-n}, ..., v_i, ..., v_{i+n})$$

with $v_i = e^{\beta_1|i-t|}$, $i \in [t-n, t+n]$, n is phoneme context influence length, depending on the speaking speed and the frame rate of the recorded video, we set $n = 10$, if the frame rate is $50Hz$, $n = 5$ at $25Hz$. β_1 is set to -0.3. M is a phoneme distance matrix with size 43×43, which denotes visual similarities between phoneme pairs. M is computed by the weighted Euclidean distance in the PCA space:

$$M(Ph_i, Ph_j) = \frac{\sqrt{\sum_{k=1}^{K} \gamma_k^2 \cdot (\overline{PCA}_{Ph_i,k} - \overline{PCA}_{Ph_j,k})^2}}{\sum_{k=1}^{K} \gamma_k} \tag{2.2}$$

where $\overline{PCA}_{Ph_i}$ and $\overline{PCA}_{Ph_j}$ are the average PCA weights of phoneme i and j, respectively. K is the reduced dimension of the PCA space of mouth images. γ_k is the weight of the k^{th} PCA component, which describes the discrimination of the components. We use exponential factor $\gamma_k = e^{\beta_2|k-K|}$, $k \in [1, K]$, with $\beta_2 = 0.1$ and $K = 12$.

The Concatenation Cost (CC) is calculated using a visual cost (f) and a skip cost (g_s) as follows:

$$CC(u_1, u_2) = wccf \cdot f(U_1, U_2) + wccg \cdot g_s(u_1, u_2) \tag{2.3}$$

with the weights $wccf$ and $wccg$. u_1 and u_2 are one of the candidate images from frame i and from frame $i-1$, respectively. U_1 and U_2 correspond to the feature vector of u_1 and u_2. The feature vector considers the articulator features including teeth, tongue, lips, appearance, and geometric features.

The visual cost measures the visual difference between two mouth images. A small visual cost indicates that the transition is smooth. The visual cost f is defined as:

$$f(U_1, U_2) = \sum_{d=1}^{D} k_d \cdot \left\| U_1^d - U_2^d \right\|_{L2} \tag{2.4}$$

$\left\| U_1^d - U_2^d \right\|_{L2}$ measures the Euclidean distance in the articulator feature space with D dimensions. Each feature is given a weight k_d which is proportional to its discrimination. For example, the weight for each component of PCA parameters is proportional to its corresponding eigenvalue of PCA analysis.

The skip cost is a penalty given to the path consisting of many video segments. Smooth mouth animations favor long video segments with few skips. The skip cost g_s is calculated

as:

$$g_s(u_1,u_2) = \begin{cases} 0 & ; \ |f(u_1)-f(u_2)| = 1 & \wedge \ s(u_1)=s(u_2) \\ w_1 & ; \ |f(u_1)-f(u_2)| = 0 & \wedge \ s(u_1)=s(u_2) \\ w_2 & ; \ |f(u_1)-f(u_2)| = 2 & \wedge \ s(u_1)=s(u_2) \\ \vdots & & \\ w_{p-1} & ; \ |f(u_1)-f(u_2)| = p-1 & \wedge \ s(u_1)=s(u_2) \\ w_p & ; \ |f(u_1)-f(u_2)| \geq p & \vee \ s(u_1)\neq s(u_2) \end{cases} \tag{2.5}$$

where f and s are the frame number of the recording and the sentence number, respectively and $w_i = e^{\beta_3 i}$. We set $\beta_3 = 0.6$ and $p = 5$.

A path (p_1,p_2,...,p_i,...,p_N) through this graph generates the following Path Cost (*PC*):

$$PC = wtc \cdot \sum_{i=1}^{N} TC(i, S_{i,p_i}) + wcc \cdot \sum_{i=2}^{N} CC(S_{i,p_i}, S_{i-1,p_{i-1}}) \tag{2.6}$$

with candidate image S_{i,p_i} belonging to the frame i. *wtc* and *wcc* are the weights of the two costs.

Substituting Eq. (2.3) in Eq. (2.6) yields

$$PC = wtc \cdot C1 + wcc \cdot wccf \cdot C2 + wcc \cdot wccg \cdot C3 \tag{2.7}$$

with

$$\begin{aligned} C1 &= \sum_{i=1}^{N} TC(i, S_{i,p_i}) \\ C2 &= \sum_{i=2}^{N} (f(S_{i,p_i}, S_{i-1,p_{i-1}})) \\ C3 &= \sum_{i=2}^{N} (g_s(S_{i,p_i}, S_{i-1,p_{i-1}})) \end{aligned}$$

The best path through the graph is the path that produces the minimum cost. The weights *wtc* and *wcc* are used to fine-tune the emphasis given to concatenation cost versus target cost, or in other words, to emphasize acoustic versus visual matching. A strong weight given to the concatenation cost will generate a very smooth animation, but the synchronization with speech might be lost. A strong weight given to the target cost will generate an animation which is perfectly synchronized to the speech but might appear visually choppy or jerky. This is due to the high amount of jumping within database sequences. These weights are selected by maximizing the average segment length over a set of training sequences.

2.2.3 Drawbacks and Limitations

The feature-based motion estimation of the reference system depends on the feature points. The template-based feature detection results in an inaccurate corresponding feature points, which makes the motion estimation incorrect, especially for frames with the mouth wide open [9]. For this reason, a final manual inspection of all frames is necessary to ensure a clean database with facial features, even though the motion estimation is performed entirely automatically.

In order to eliminate these deficiencies, our system uses model-based motion estimation [66] which estimates the head motion parameters for each frame. Using a 3D head scan of the recorded speaker, the model-based estimation algorithm stores texture information of the object and tries to find the motion parameters of the rigid head in a new frame by minimizing the difference between the textured 3D model and the face in the new frame. The 3D head scan is a 3D face representation, which is a polygon mesh consisting of a collection of vertices and polygons that define the shape of a face in 3D. Using the motion parameters, all frames are normalized to a reference position and the mouth regions of all frames are cropped to build a database, so that the normalized mouth images can be parameterized by PCA correctly.

In the reference system [9], the feature detection is implemented in two steps. First, the face is located by color based segmentation. Second, feature points are detected using manually predefined feature templates, such as mouth corners cropped from mouths with different open sizes. One of the deficiencies is that it is time-consuming and very sensitive to the skin color and lighting of the subject. If the color contrast between lip and skin is small, the threshold cannot be found correctly. Second, feature templates are selected manually, and they are not always consistent in all the templates. In this thesis, the facial features are detected by AAM [25] [57], which uses not only the color information but also the shape of the subject such that the approach can locate the feature points very precisely and robustly. Moreover, a new approach to reduce the manual error when creating the training data is proposed in this thesis [7].

The second improvement in our facial animation system is to smooth the transitions between segments by morphing instead of blending [9]. A segment is a chunk of consecutively recorded mouth images. Since mouth animations are composed of several segments, the discontinuities of mouth movements occur only at the joint of these segments. Blending introduces the motion blur artifacts. Furthermore, morphing [67] requires several corresponding feature points between two images. The color based method can only detect the features points such as mouth corners, which are not enough for morphing. However, the results from AAM can satisfy this need.

Another drawback of the reference system is that the weights of the unit selection are not optimized. In [65] two approaches are proposed to train the weights of the unit selection for a speech synthesizer. In the first approach, weight space search is to search a range of weight sets in the weight space and find the best weight set which minimizes the differences between the natural waveform and the synthesized waveform. In the second

approach, regression training is used to determine the weights for the target cost and the weights for the concatenation cost separately. Exhaustive comparison of the units in the database and multiple linear regressions are involved. Both methods are time consuming and the weights are not globally optimal. An approach similar to weight space search is presented in [68], which uses only one objective measurement to train the weights of the unit selection. However, other objective measurements are not optimized. Therefore, these approaches are only sub-optimal for training the unit selection, which has to create a compromise between partially opposing objective quality measures. Considering multi-objective measurements, Pareto optimization is proposed to train the unit selection in this thesis.

2.3 Optimization of the Reference System

The main improvements of the reference system are presented in [10]. In this section, we describe two improvements. One is the feature detection in the analysis, and the other is unit selection optimization in the synthesis.

The analysis of the recorded videos are improved as shown in Fig. 2.8. Compared with the reference system, our system estimates the head motion by the model-based algorithm [66] and the geometric features are detected by AAM instead of the template-based approach. In this thesis, the reliable feature detection by AAM is presented in the section 2.3.1.

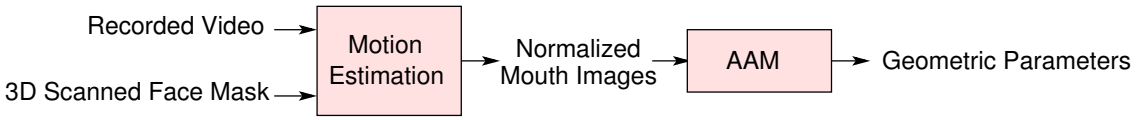


Figure 2.8: Block diagram of the analysis of recorded videos in the optimized system.

With respect to synthesis, the unit selection is optimized by Pareto optimization, which is proposed in the Section 2.3.2

2.3.1 Reliable Feature Detection using AAM

The geometric parameters, such as mouth corner points and lip position, are obtained by template matching based approach in the reference system [9]. This method is very sensitive to the illumination change resulted from mouth movement and head motion during speaking, even though the environment lighting is consistent in the studio. Furthermore, the detection error of the mouth corners may be higher when the mouth is very wide open. The same problem exists also for the detection of eye corners, which will result in an incorrect motion estimation and normalization.

AAM-based feature detection uses not only the texture but also the shape of the face. AAM models are built from a training set including different appearances. The shape is manually marked. Because the AAM is built in a PCA space, if there are enough training data that can construct the PCA space, the AAM is not sensitive to the illumination change on the face. Typically the training data set consists of about 40 mouth images.

Active Appearance Models (AAM)

The active appearance models [25] were introduced by Edwards et al. a few years ago, and have since been the subject of much research. An appearance model is a joint statistical model of the shape and the texture of an object, while the active appearance model is an appearance model combined with a directed search algorithm for adapting the model to an image. This section will summarize the traditional AAM algorithm.

From our mouth image database, several different mouth images are selected and landmarked. Each image is related to a landmark vector. The landmark vectors can be used to build the shape model for the object (mouth, in our context). The statistical shape is parameterized by PCA according to

$$x = \overline{x} + P_s b_s \tag{2.8}$$

$\overline{x}$ is the standard (average) shape of the model, the columns of P_s are the shape eigenvectors, and the b_s contains the shape parameters.

The shapes of all images are aligned by Generalized Procrustes Analysis (GPA). The textures under the landmarked shapes are extracted and warped to the mean shape. The warped texture data vectors are parameterized by PCA, a linear model is obtained as

$$g = \overline{g} + P_g b_g \tag{2.9}$$

$\overline{g}$ is the mean texture, the columns of P_g are the texture eigenvectors, and the b_g contains the texture parameters.

In order to recover the correlation between the shape and the texture, the parameters b_s and b_g are combined in a third PCA space. A combined parameter c is computed as

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s P_s^T (x - \overline{x}) \\ P_g^T (g - \overline{g}) \end{pmatrix} = \begin{pmatrix} P_{cs} \\ P_{cg} \end{pmatrix} c = P_c c \tag{2.10}$$

where W_s is a diagonal matrix of weights for each shape parameter, allowing for the difference in units between the shape and texture model. P_c is the matrix consisting of the combined eigenvectors.

The Appearance model is parameterized by c as

$$\begin{cases} x = \overline{x} + P_s W_s^{-1} P_{cs} c \\ g = \overline{g} + P_g P_{cg} c \end{cases} \tag{2.11}$$

For a given c, the texture g and the shape x can be calculated by the Eq. (2.11). A synthesized image is generated by warping the texture of g into the shape of x and applying the current pose parameters $p = [t_x\ t_y\ s\ \theta]$ where t_x, t_y and θ denote in-plane translation and rotation, and s denotes the shape size.

The goal of AAM search is to find the optimal adaptation of the model to the input image, that is to find the parameter c and p that minimizes the distance between the model and the image using the L_2-norm as a cost function. The iterative updating scheme is based on a fixed Jacobian estimate or a principle component regression [25].

Feature Detection by AAM

Facial features are defined as landmarks on the mouth part (Fig. 2.9) in our system. The landmarks are divided into three groups: anatomical-, mathematical- and pseudo-landmarks [57]. For example, the mouth corners are anatomical landmarks, the lowest points of inner and outer lip are mathematical landmarks, and the points on the outline or between landmarks are pseudo landmarks.

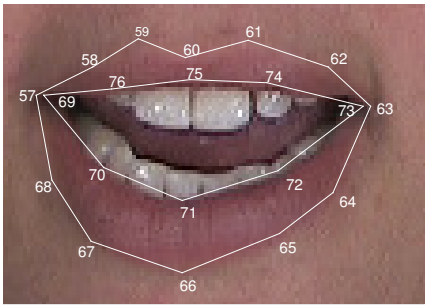


Figure 2.9: 20 Landmarks of mouth part.

An appropriate training set is required for AAM building. According to different mouth appearance such as lip open size, teeth and tongue, about 40 mouth images are selected from the database with 20000 mouth images. The training set includes the mouth images and their landmarks.

Optimization of AAM Model

Because the manual work of placing landmarks is a subjective task, in general, the landmarks are not consistent in the training set. Improving the accuracy of the landmarks in the training set should be carried out, which is also called a repeatability and reproducibility study. Stegmann [57] has proposed that by letting a set of operators annotate the data set several times, the average landmarks are estimated as the optimal landmarks for the data set. However, this method needs a lot of manual work and it is tedious. In this part,

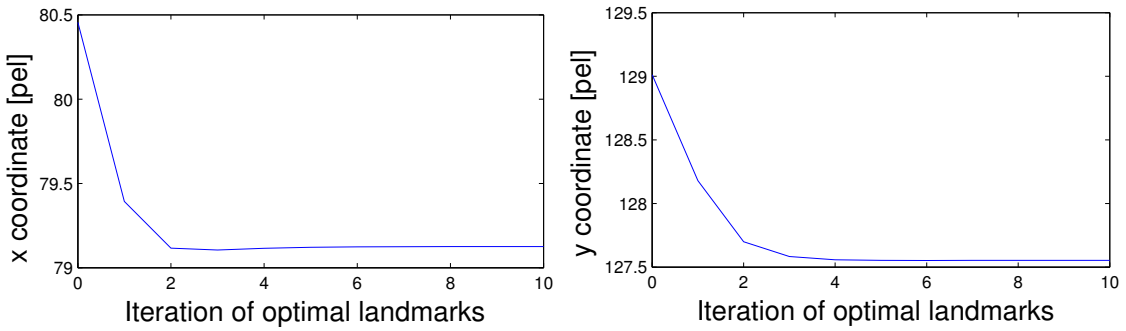
we will propose a new approach, which can reduce the manual error of landmarks in the training set.

The optimization process is described as follows:

- Step 1: The training set is used to build the AAM model.
- Step 2: Features of each image from the training set are detected by the AAM model. If the features did not change from the last iteration, the optimization process will stop. Otherwise, do next step.
- Step 3: The landmarks in the training set will be updated by the detected feature points. Go to step 1.

Generally the landmark converges to a stable position after 3 to 5 times iterations.

Fig. 2.10 shows an optimization process of a mouth corner. The x coordinate of the mouth corner is shown in Fig. 2.10 (a), the y in Fig. 2.10 (b). The optimization starts from the manual landmark. In this case, 80.5 is the x coordinate of mouth corner. After 4 iterations, the x coordinate is converged to 79.13. The position of the landmark moves about 1.37 pixels in x, 1.46 pixels in y. The average correction of the landmarks is about 0.78 pixel in x and 0.89 in y.



(a) x coordinate (b) y coordinate

Figure 2.10: Manual errors of landmarks are reduced iteratively by AAM optimization.

Sub-pel Accurate Feature Detection

The optimized AAM is used to detect feature points of the mouth images, resulting in precise positions. We have also defined objective criteria to measure the accuracy of the feature points.

The most popular method to measure the accuracy is to mark a test sequence manually, and the average difference between the detected features and the manually marked features is calculated as the error. This method is time-consuming and cumbersome. Some

call this manually marked features ground truth. However, manual marks can be inconsistent in different images. Therefore, this definition of ground truth is unsuitable, and consistent feature points are needed as ground truth.

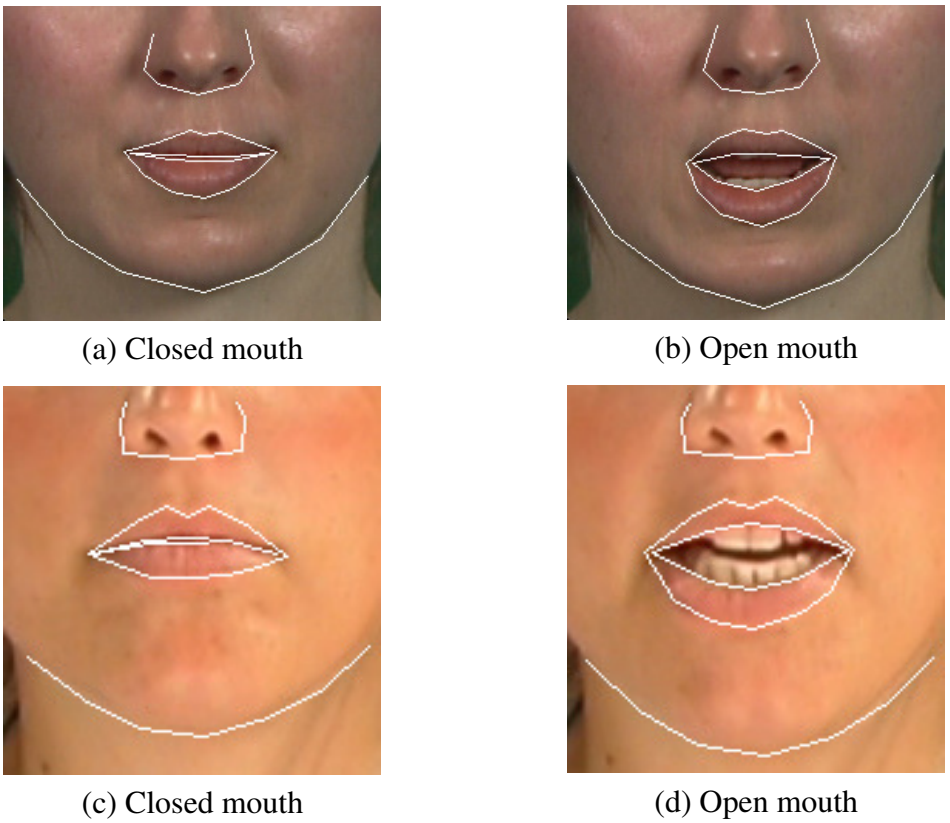
We start with an optimized AAM based on 40 training images with landmarks. We manually mark landmarks in 100 test images. Next we build an extended optimized AAM using training and test images. This extended AAM detects the feature points in the test images. We define these feature points as the ground truth for the test images. The manual error is measured by the difference between the ground truth and the manually marked features. The detection error is measured by the difference between the feature points detected using the AAM based on the training images and the ground truth.

(a) Low contrast (b) High contrast

Figure 2.11: Results of AAM-based feature detection under different lighting and color changes.

Using the 140 images as the training data, the ground truth features are detected by the optimal AAM approach. The average error between the manually marked features and ground truth is about 1.15 pixels. Using 40 images as the training set and 100 images as the test set, the average error between the ground truth and the features of the test images, detected by the non-optimized AAM approach, is 0.44 pixels. The error is reduced to 0.17 pixels by the optimized AAM. The results of feature detection under different conditions such as lighting and color changes are shown in Fig. 2.11. Fig. 2.12 shows the AAM-based feature detection used for the test data [4] (Fig. 2.12(a)(b)) and the data from our Institute (Fig. 2.12(c)(d)). The overall accuracy improvement by the optimized AAM approach is smaller than 0.5 pixels. The standard deviation of the detection error is 0.04 pixels, and the maximal error is 0.46 pixels.

The mouth corners are detected by optimized AAM as accurately as by the template-base method [9], while AAM based detection saves much manual work to train the model and it is more robust to the lighting and color changes. The lip contour can also be detected with high accuracy by optimized AAM, but the template based method cannot. The detected mouth features are further used to control the morphing for animation.



(a) Closed mouth (b) Open mouth

(c) Closed mouth (d) Open mouth

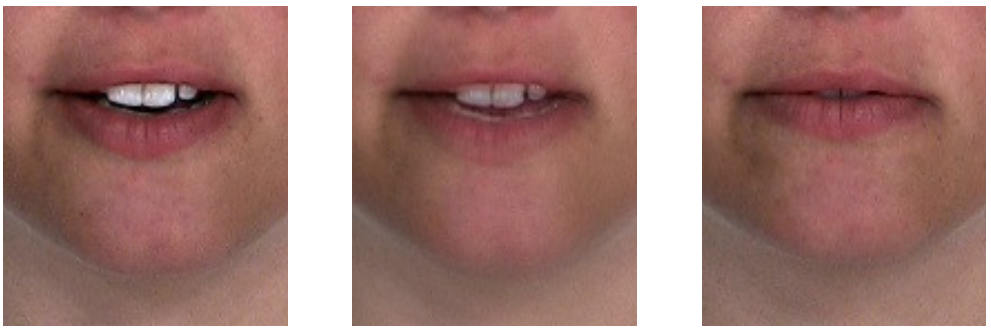
Figure 2.12: AAM-based feature detection on normalized mouths of different databases.

Morphing for Animation

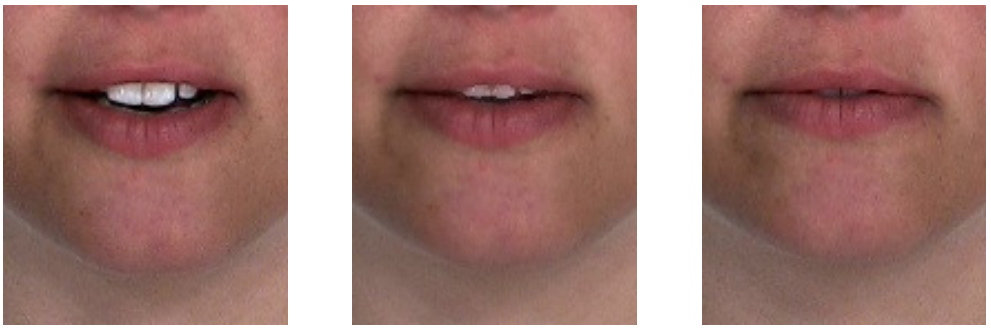
The selected mouth images are made from segments of consecutively recorded image samples. The Viterbi search has done its best to minimize visual differences at the junction of these segments. However, due to the database having a limited size, many junctions remain for which a large visual difference exists. These abrupt transitions are generally noticeable on the final animation and greatly lower its overall quality. Instead of blending, field morphing [69] is implemented when the junctions have a large visual difference. Morphing requires feature points of the mouth part. The correspondences of the feature points in the mouth images have been detected by optimized AAM approach. Therefore, morphing can be implemented in our system. The whole system runs also in real time because the mouth part is small and morphing is required only at junctions with large visual differences.

An animation sequence (25 fps) is produced for the utterance “We are working on facial animation.” Only 3 of 17 transitions need morphing. On average 25% of transitions are morphed in animations. The mouth part of animations with blending and morphing the transitions are shown in Fig. 2.13. From Fig. 2.13(a), the blended image shows the

transparency of the lips and teeth. In Fig. 2.13(b), the morphed mouth has a much clearer mouth shape than the blended one. The teeth are a little darker than the left one, because the morphing technique tries to blend the forward and backward warped mouth images. The forward warped mouth has a light white color, while the backward warped mouth has a dark white color. Therefore, the brightness of the teeth becomes darker after blending the warped images.



(a) Blending the transition



(b) Morphing the transition

Figure 2.13: Snapshots of mouth part of facial animations. The middle image is blended in (a), morphed in (b). The visual quality of mouth part is improved by morphing. The blending factor is set to 0.5.

The improved training method can reduce the average position error of landmarks from 1.15 pixels to 0.17 pixels iteratively. In contrast to the reference method, the feature detection using the improved AAM model performs more accurately and robustly. The detected geometric features are used for the synthesis of facial animation. Furthermore, the feature points are also used for morphing, which smooths the transition between segments with large visual differences automatically. Since the two improvements are built in our system, the overall quality of animations is better than these by the reference method as evaluated by subjective tests.

2.3.2 Automatic Parameter Optimization of Unit Selection

As discussed in section 2.2.2, several weights, influencing TC, CC and PC, should be trained. Generally, the training set includes several recorded original sentences (as ground truth) which are not included in the database. Using the database, an animation with the audio of the training set will be generated using the given weights for unit selection. We use objective evaluator functions as Face Image Distance Measure (FIDM). The evaluator functions are average target cost, average segment length, and average visual difference between segments. The average target cost indicates the lip synchronization, the average segment length and average visual difference indicate the smoothness.

Pareto Optimization

Pareto optimization, also known as genetic algorithm (GA), is population-based search algorithm. Inspired in natural evolution ideas, GA evolves a population of candidates solution (i.e. weights) adapting them to a given environment, or fitness function (i.e. unit selection cost). This process takes advantage of mechanisms such as the survival of the fittest and genetic material recombination. The basic GA searches optimal weights at multiple locations in the search space rather than just one location. It does this by producing a population of various possible solutions distributed throughout the search space and then operates on these solutions using basic operations to create a new generation of better solutions. The operators correspond closely with the operations of nature as pertains to survival of species [70] [71].

The basic GA is summarized as follows:

- Step 1: An initial population consisting of several possible solutions in the search space is created. The issues to be addressed at this step are the size of the initial population and the criteria for choosing the initial population so as to distribute the search at the various locations in the search space. In our system, the initial population is the weights of the unit selection.

- Step 2: Each solution in the current population is evaluated by using a fitness function. The fitness function is a measurement of the suitability of the solution to survive into the next generation. A new population is selected by sampling the current one.

- Step 3: The individuals of the new population are recombined in two different phases. The first is crossover, which recombines the solutions producing two new offsprings. Moreover, the offspring replace their parents in the population. The second phase is mutation, which introduces random perturbations to the solutions with a given probability. This process is a mechanism to extend search to unexplored domains in the search space.

- Step 4: The steps 2 and 3 are repeated for several generations to allow evolution of very stable and fit population. Generally some termination criteria are used to stop the evolution after several iterations.

Use of a GA for unit selection is advantageous as we can terminate evolution at any step and pick the best solution. Also as the size of the database grows, GA based algorithm can be more and more efficient than enumerative (brute force) optimization approaches [68].

Multi-Objective Measurements

In this section, we define the multi-objective measurements as the fitness functions.

A mouth sequence (p_1,p_2,...,p_i,...,p_N) with minimal path cost is found by the Viterbi search in the unit selection. Each mouth has a target cost (TC_{p_i}) and a concatenation cost including a visual cost and a skip cost in the selected sequence.

The average target cost is computed as

$$TC_{avg.} = \frac{1}{N}\sum_{i=1}^{N} TC_{p_i} \tag{2.12}$$

As mentioned before, the animated sequence is composed of several original video segments. We assume that there are no concatenation costs in the mouth image segment, because they are consecutive frames in a recorded video. The concatenation costs occur only at the joint position of two mouth image segments. When the concatenation costs are high, indicating a large visual difference between two mouth images, this will result in a jerky animation. The average segment length is calculated as

$$SL_{avg.} = \frac{1}{L}\sum_{l=1}^{L}(SL_l) \tag{2.13}$$

where L is the number of segments in the final animation. For example, the average segment length of the animation in Fig. 2.7 is calculated as $SL_{avg.} = (6+3+2+11+3)/5 = 5$.

The Euclidean distance (f_{pca}) between mouth images in the PCA space is used to calculate the average visual difference in the following way:

$$VC_{avg.} = \frac{1}{L-1}\sum_{i=1}^{N-1} f_{pca}(i,i+1) \tag{2.14}$$

where $f_{pca}(i,i+1)$ is the visual distance between mouth images at frame i and $i+1$ in the animated sequence. If the mouth image at frame i and $i+1$ are two consecutive frames in an original video segment, the visual distance is set to zero. Otherwise, the visual distance for the joint of the mouth image segments is calculated as

$$f_{pca}(i,i+1) = \left\| \overrightarrow{PCA}_i - \overrightarrow{PCA}_{i+1} \right\|_{L2} \tag{2.15}$$

where PCA_i is the PCA parameter of the mouth image at frame i.

The optimal average visual difference $VC_{avg.}$ should consider not only the appearance features (PCA parameters) but also the geometric features as the cost function definition in the concatenation cost. However, the weight used for the appearance and geometric features is to be searched by the Pareto optimization. In the subjective test, we have observed that the appearance feature is more discriminative than the geometric one, so we choose the PCA parameters for the suboptimal measurement.

Pareto Optimization of Unit Selection

Similar to the works in [72] and [73], we adapt the Pareto optimization to our visual speech synthesis. The Pareto algorithm starts with an initial population. Each individual is a weight vector containing weights to be adjusted. Then, the population is evaluated by the multi-objective evaluator functions (i.e. FIDM). A new population is created by recombining the individuals of the current population in two steps, i.e. crossover and mutation. The first step recombines the weight values of two individuals to produce two new children. The children replace their parent in the population. The second step introduces random perturbations to the weights with a given probability. Finally, the new population is obtained to replace the current one, starting the evolutionary cycle again. This process stops when a certain finalization criteria is satisfied.

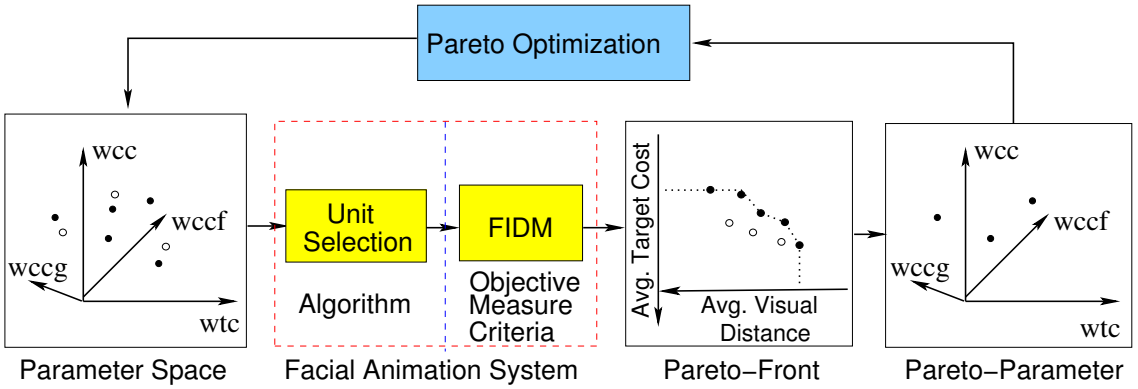


Figure 2.14: The Pareto optimization for the unit selection.

FIDM is used to evaluate the unit selection and the Pareto optimization accelerates the training process. The Pareto optimization (as shown in Fig. 2.14) begins with ten thousand combinations of weights of the unit selection in the parameter space, where ten settings were chosen for each of the four weights in our experiments. For each combination, there is a value calculated using the FIDM criteria. The boundary of the optimal FIDM values is called Pareto-front. The boundary indicates the animation with smallest possible target cost given a visual distance between segments. Using the Pareto parameters corresponding to the Pareto-front, the Pareto optimization generates new combinations of the weights

for further FIDM values. The optimization process is stopped as soon as the Pareto-front is declared stable.

Once the Pareto-front is obtained, the best weights combination is located on the Pareto-front. However, the visual qualities of these animations are different. The subjective test is the ultimate way to find the best weights combination, but there are many weight combinations performing similar results that subjects cannot distinguish. Therefore, it is necessary to define objective measurements to find the best weight combination automatically and objectively.

The measurable criteria consider the subjective impression of quality. We have performed the following objective evaluations. The similarity of the real sequence and the animated sequence is described by directly comparing the visual parameters of the animated sequence with the real parameters extracted from the original video. We use the cross-correlation of the two visual parameters as the measure of similarity. The visual parameters are the size of open mouth and the texture parameters.

Appearance similarity is defined as the correlation coefficient (r_{pca}) of the PCA weights of the animated sequence and the original sequence. If the unit selection finds a mouth sequence, which is similar to the real sequence, the PCA parameters of the corresponding images of the two sequences have a high correlation. Movement similarity is defined as the correlation coefficient (r_h) of the mouth height. If the animated mouth moves just as the recorded mouth, the coefficient approaches 1.

Since the PCA coefficients are ordered in a descending significance, the first PCA coefficient represents the most statistical variance of mouth images. Therefore, the first PCA coefficient could be used for appearance similarity instead of all the coefficients. The mouth height is visually more significant than the mouth width in our informal subjective tests.

The cross-correlation is calculated as

$$r = \frac{\sum_{i=1}^{N}[(y_{a,i} - m_{y_a}) \cdot (y_{b,i} - m_{y_b})]}{\sqrt{\sum_{i=1}^{N}(y_{a,i} - m_{y_a})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_{b,i} - m_{y_b})^2}} \tag{2.16}$$

where $y_{a,i}$ and $y_{b,i}$ are the first principal component coefficient of PCA parameter or the mouth height of the mouth image at frame i in the real and animated sequence, respectively. m_{y_a} and m_{y_b} are the means of the corresponding series, y_a and y_b. N is the total number of frames of the sequence.

In addition to the appearance similarity and movement similarity criteria, the timing of mouth closure is also used as an evaluator. Experiments indicate that human viewers are very sensitive to closures, and getting the closures at the right time may be the single most important criterion for providing the impression that lips and sound are synchronized. The timing of mouth closures is very critical for animations with plosive and bilabial phonemes such as phoneme “p”, “b” and “m”.

2.4 Experimental Results

2.4.1 Data Collection

In order to test our talking head system, two data sets are used, comprising the data from our Institute (TNT) and the data from LIPS2008 [4].

In our studio a subject is recorded while reading a corpus including about 300 sentences. A lighting system is developed for the audio-visual recording, which minimizes the shadow on the face of a subject and hence reduces the change of illumination on the moving head. The capturing is done using a HD camera (Thomson LDK 5490). The video format is originally 1280×720 at $50fps$, which is cropped to 576×720 pixels at $50fps$. The audio signal is sampled at $48kHz$. 148 utterances are selected to build a database to synthesize animations. The database contains 22762 normalized mouth images with a resolution of 288×304.

The database from LIPS2008 consists of 279 sentences, supporting the phoneme transcription of the texts. The video format is 576×720 at $50fps$. 180 sentences are selected to build a database for visual speech synthesis. The database contains 36358 normalized mouth images with a resolution of 288×288.

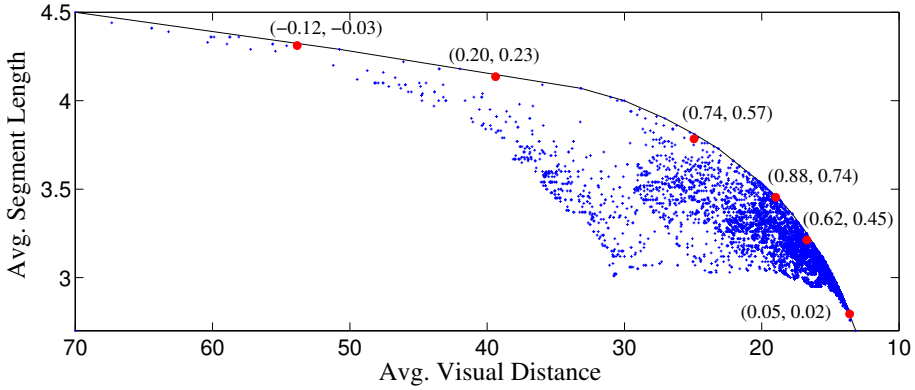
A snapshot of example images extracted from two databases is shown in Fig. 2.15.

(a) TNT (b) LIPS2008

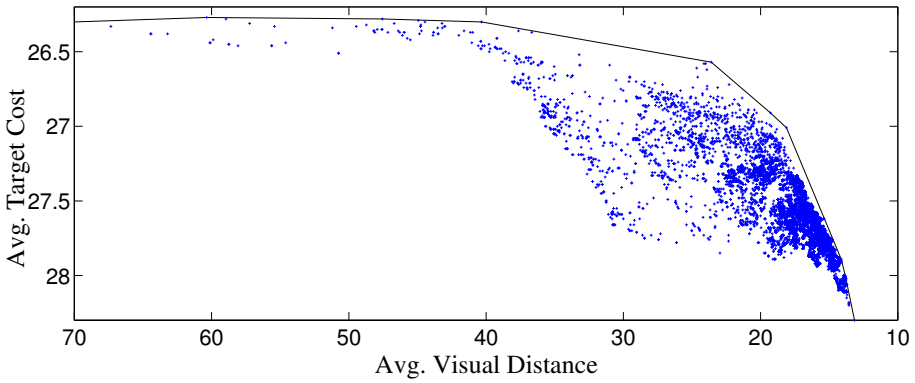
Figure 2.15: Snapshot of an example image extracted from recorded videos at TNT and LIPS2008, respectively.

2.4.2 Unit Selection Optimization

The unit selection is trained by Pareto optimization with 30 sentences. The Pareto-front is calculated and shown in Fig. 2.16. There are many weight combinations satisfying the objective measurement on the Pareto-front, but only one combination of weights is determined as the best set of weights for unit selection. We have tried to generate animations by using several weight combinations and find out that they have similar quality subjectively in terms of naturalness, because quite different paths through the graph can produce very similar animations given a quite large database.



(a) Evaluation space for $VC_{avg.}$ and $L_{avg.}$



(b) Evaluation space for $VC_{avg.}$ and $TC_{avg.}$

Figure 2.16: Pareto optimization for unit selection. The curves are the Pareto-front. Several Pareto points on the Pareto-front marked red are selected to generate animations. The cross correlation coefficients of PCA parameters and mouth height (r_{pca}, r_h) between real and animated sequences are shown for the selected Pareto points.

To evaluate the Pareto-front automatically, we use the defined objective measurements to find best animations with respect to naturalness. The cross correlation coefficients of

PCA parameter and mouth height between real and animated sequences on the Pareto-front are calculated and shown in Fig. 2.17. The red curve is the cross correlation of PCA parameter of mouth images between real and animated sequences. The blue curve is the cross correlation of mouth height. The cross correlation coefficients of several Pareto points on Pareto-front are labeled in Fig. 2.16(a), where the first coefficient is r_{pca}, the second is r_h. Given in Fig. 2.17, the appearance similarity (red curve) and the movement similarity (blue curve) run in a similar way, and reach the maximal cross correlation coefficients at the same position with the average visual distance of 18.

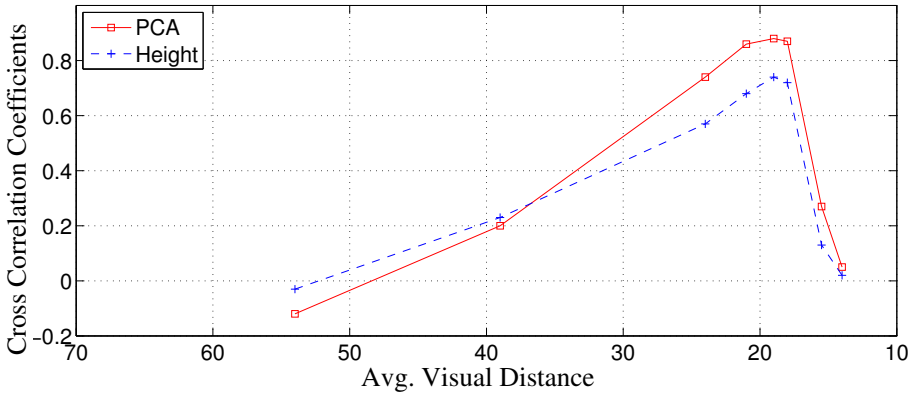
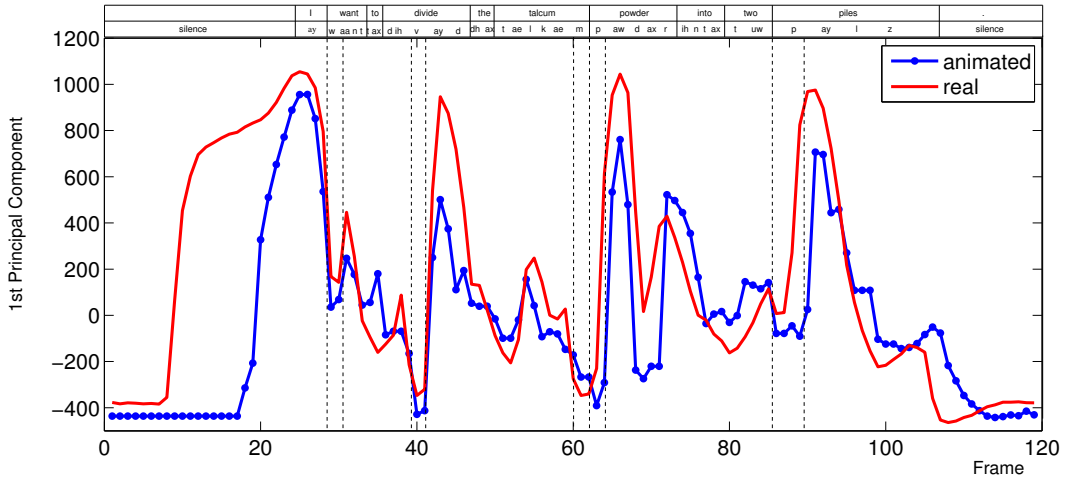


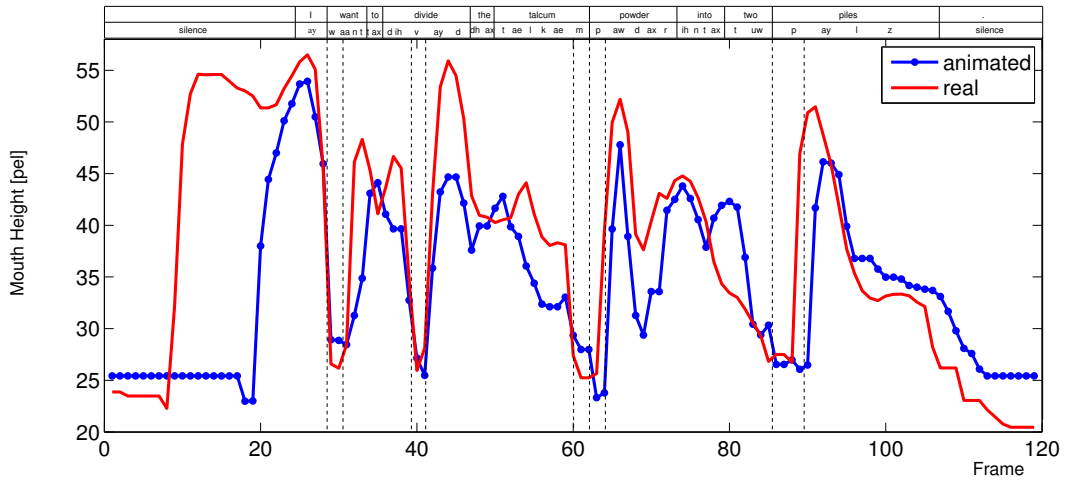
Figure 2.17: Cross correlation of PCA parameters and mouth height of mouth images between real and animated sequences on the Pareto-front. Red curve is cross correlation of PCA parameter between real and animated sequences. The blue one is the cross correlation of mouth height.

Fig. 2.18(a) shows the first component of PCA parameters of mouth images in real and animated sequences. The mouth movements of the real and synthesized sequences are shown in Fig. 2.18(b). We have found that the curves in Fig. 2.18 do not match perfectly, but they are highly correlated. The mouth height trajectory of the animated sequence is different from the one of the real sequence at several positions in the Fig. 2.18(b). For example, the mouth of silence phoneme in the real sequence is open at the beginning of the sentence, while the mouth of the animated sequence begins with closure. The resulting facial animations look realistic compared to the original videos. One of the most important criteria to evaluate the curves is to measure how well the closures match in terms of timing and amplitude. The phonemes “w”, “v”, “m”, and “p” in Fig. 2.18(b) have closed mouths. The timing of these mouth closures are matched very well in the animated and real sequences. Furthermore, objective criteria and informal subjective tests are consistent to find the best weights in the unit selection. In such a way the optimal weight set is automatically selected by the objective measurements.

The weight set corresponding to the point on the Pareto-front with maximal similarity (r_{pca}, r_h) are used in the unit selection. Animations generated by the optimal facial animation system are used for the following formal subjective test.



(a) Trajectory of the first PCA weight



(b) Trajectory of mouth height of real and animated sequences

Figure 2.18: The similarity measurement for the sentence: "I want to divide the talcum powder into two piles". (a) shows the appearance similarity, (b) shows the mouth movement similarity. The red curve is the PCA parameter trajectory and the mouth movement of the real sequence, the blue curve is the PCA parameter trajectory and mouth movement of the animated sequence. The cross correlation coefficient of PCA parameters between the real and animated sequence is 0.88, the coefficient for mouth height is 0.74. The mouth height is defined as the maximal top to bottom distance of the outer lip contour. The dashed lines indicate the interval of closures for phonemes “w”, “v”, “m”, and “p”.

Rendering Performance

The performance of visual speech synthesis depends mainly on the TTS synthesizer, the unit selection, and the OpenGL rendering of the animations. We have measured that the TTS synthesizer has about 10 ms latency in a WLAN network. The unit selection is running as a thread, which only delay the program at the first sentence. The unit selection for the second sentence is run when the first sentence is rendered. Therefore, the unit selection is done in real time. The OpenGL rendering takes the main time of the animations, which relies on the graphics card. For our system (CPU: AMD Athlon XP 1.1GHz, Graphics card: NVIDIA Geforce FX 5200), the rendering needs only 25 ms for each frame of a sequence with CIF format at 25 fps.

2.4.3 Subjective Test

A subjective test is defined and carried out to evaluate the facial animation system. The goal of the subjective test is to assess the naturalness of animations whether they can be distinguished from real videos.

Assessing the quality of a talking head system becomes even more urgent as the animations become more lifelike, since improvements may be more subtle and subjective. A subjective test where observers give feedback is the ultimate measure of quality, although objective measurements used by the Pareto optimization can greatly accelerate the development and also increase the efficiency of subjective tests by focusing them on the important issues. Since a large number of observers are required, preferably from different demographic groups, we designed a Website for the subjective test.

In order to get a fair subjective evaluation, we let the viewers focus on the lips and separate the different factors. Since head motions and facial expressions influence the speech perception, we selected a short recorded video with neutral expressions and tiny head movements as the background sequence. The mouth images, which are cropped from a recorded video, are overlaid to the background sequence in a correct position and orientation to generate a new video, named original video. The corresponding real audio is used to generate a synthesized video by the optimized unit selection. Thus a pair of videos, uttering the same sentence, is ready for the subjective test. Overall 5 pairs of original and synthesized videos are collected to build a video database available for the subjective test on our Website. The real videos corresponding to the real audios are not part of the database.

A Turing test was performed to evaluate our talking head system. 30 students and employees of Leibniz Universität Hanover were invited to take part in the formal subjective test. All video pairs from the video database were randomly selected and the video pair was itself presented to the participant randomly only once. The participant should decide whether it is an original or a synthesized video immediately after the video pair was displayed.

The results of the subjective test are summarized in Table 2.2. The Turing test can be

quantified in terms of the Correct Identifying Rate (CIR), which is defined as

$$CIR = \frac{Number\ of\ Correctly\ Identified\ utterances(NCI)}{Number\ of\ Testing\ Utterances(NTU)} \tag{2.17}$$

Table 2.2: Results of the subjective test for talking heads. 5 video pairs were shown to 30 viewers. The number of the viewers, which identified the real and synthesized video correctly (NCI), was counted. The correct identifying rate (CIR) for each video pair was calculated. NTU represents the number of testing utterances.

Video pair	1	2	3	4	5
NCI	21	16	17	11	21
NTU	30	30	30	30	30
CIR	70%	53%	57 %	37 %	70%

Table 2.2 shows the results of the subjective test. Snapshots of the talking heads extracted from 5 video pairs are shown in Fig. 2.19. CIR 50% is expected, which means the animations are as realistic as the real one. From the results of the subjective test, we can find that the original videos of video pair 1 and 5 are correctly recognized by 70% of the viewers. The video pair 2 and 3 is almost indistinguishable to the viewers, where the CIR is approaching 50%. The synthesized video of video pair 4 is decided by most viewers as original video. After the subjective test, we have asked the viewers, how they recognized the videos as synthesized. They gave a clue that the mouth animations of the video pair 1 and 5 are as good as the real ones, but the other facial parts, such as cheeks and neck, look unnatural in several frames, where some lighting changes appear on the cheeks and the neck parts due to the difference between the mouth images and the background images.

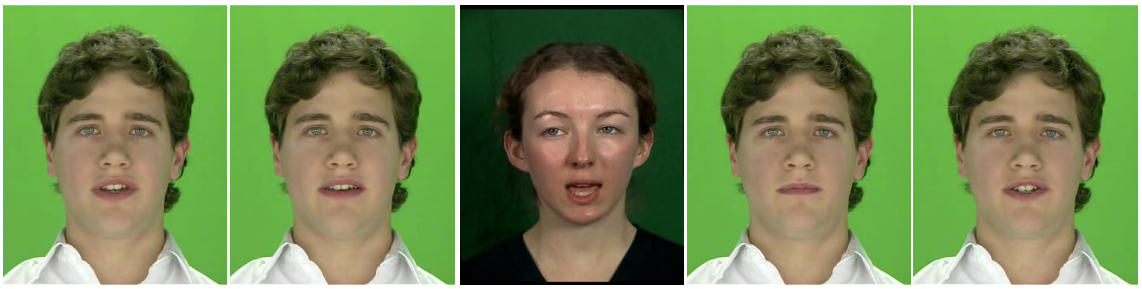


Figure 2.19: Snapshots of the talking heads extracted from 5 video pairs.

Our hypothesis is that original and animated videos are indistinguishable from each other. If the hypothesis is true, the value for NCI is binomially distributed. The probability mass function of binomial distribution is defined in the following way:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2.18}$$

with parameters $n = NTU = 30$, $k = NCI$ and $p = 0.5$ for our subjective test. Fig. 2.20 shows the binomial distribution of the subjective test. The 95% confidence interval is estimated in the zone between 10 and 20. The video pair 2, 3, and 4 is kept in the confidence interval, which means the video pairs are indistinguishable. The video pair 1 and 5 is outside of the confidence interval, but they are very close to the confidence level. In fact, the animations of 1 and 5 also look as realistic as the real ones according to the feedback of the viewers.

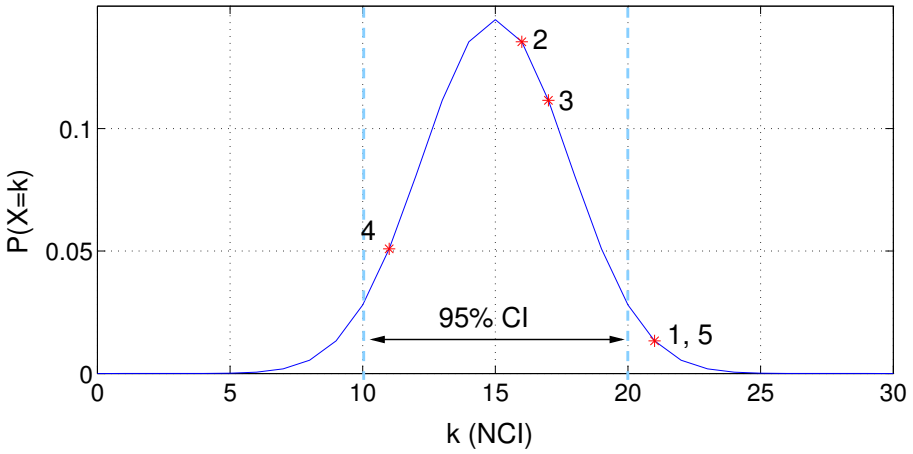


Figure 2.20: Binomial distribution ($n = 30$, $p = 0.5$) of the subjective test. The video pairs are marked with red on the distribution.

The generated talking heads using the LIPS 2008 database were evaluated on the conference of Interspeech 2008. Using the common database [4], the presented competing animation systems were evaluated subjectively. The participants include the research teams from Microsoft Research Asia, TU Berlin, University of Edinburgh, University of Grenoble, University of East Anglia, etc. In comparison to these systems, our proposed talking head system [74] achieved the best audio-visual consistency in terms of naturalness [75] [76]. The Mean Opinion Score (MOS) of our system was about 3.7 in the subjective test evaluated by a 5-point grading scale (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad). The original videos were scored with about 4.7. Fig. 2.21 shows the subjective test in the LIPS 2008 challenge, where the scores along the horizontal axis are average percentages of phonemes correctly identified in the SUS (semantically unpredictable sentences) and were obtained from the intelligibility test, and the scores along the vertical axis are the MOS (1-5) for the audio-visual consistency test (Naturalness Test).

The subjective test carried out in our institute shows that the talking head generated by using the database of TNT performs better than the talking head generated by using the database of LIPS2008. A reason for the better animation results is the designed light settings resulting in a high quality recording. All viewers think the videos from TNT look better, since the lighting contrast of the image has a large impact on the perception

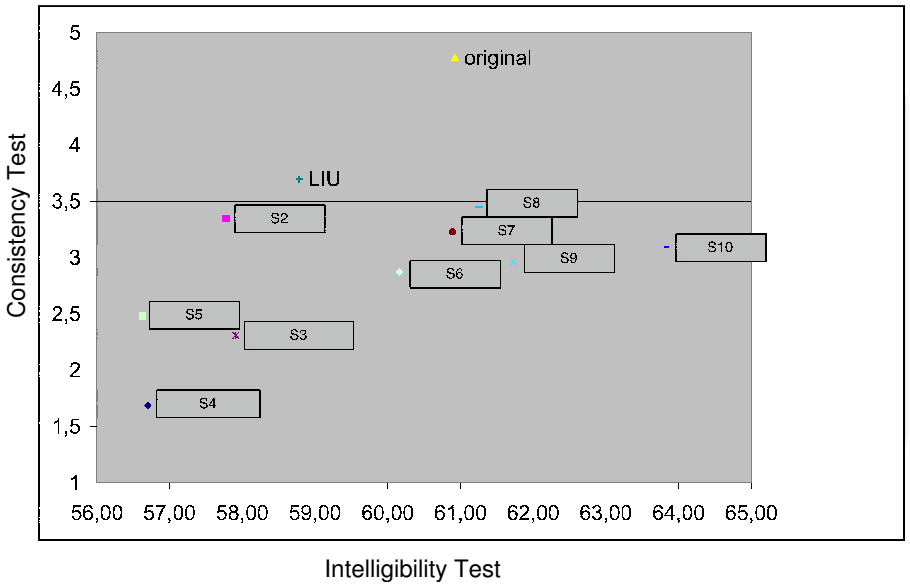


Figure 2.21: Subjective test in the LIPS 2008 challenge. 10 talking head systems were involved in the subjective test. Our system is labeled with "LIU", and "original" means the recorded videos.

of overall quality of talking heads in the subjective test. Furthermore, the shadow and the illumination changes on the face cause problems in motion estimation, which makes the final animations jerky and blinking. Therefore, talking heads generated by using the database of LIPS2008 do not look as realistic as those heads by using the database of TNT.

Based on the facial animation system, Web-based interactive services such as E-shop and Newsreader were developed. The demos and related Website are available at http://www.tnt.uni-hannover.de/project/facialanimation/demo. In addition, the video pairs used for the subjective test can be downloaded from http://www.tnt.uni-hannover.de/project/facialanimation/demo/subtest.

3 Facial Expression Synthesis

The general approach we use to model facial expressions is to build an expressive database, which is composed of different expressions, such as smile, angry, and surprising. The reference talking head system is extended with facial expressions as shown in Fig. 3.1. The reference system can generate only neutral faces without any facial expression. To simplify the problem of facial expression synthesis, a smile is conducted as an example in this thesis. The expressive database of Fig. 3.1 includes two expression states: neutral and smile. For each expression, a number of mouth images are recorded. Depending on the input expression tags, the unit selection selects appropriate mouth images from the expressive database by switching between neutral and other expressive mouths.

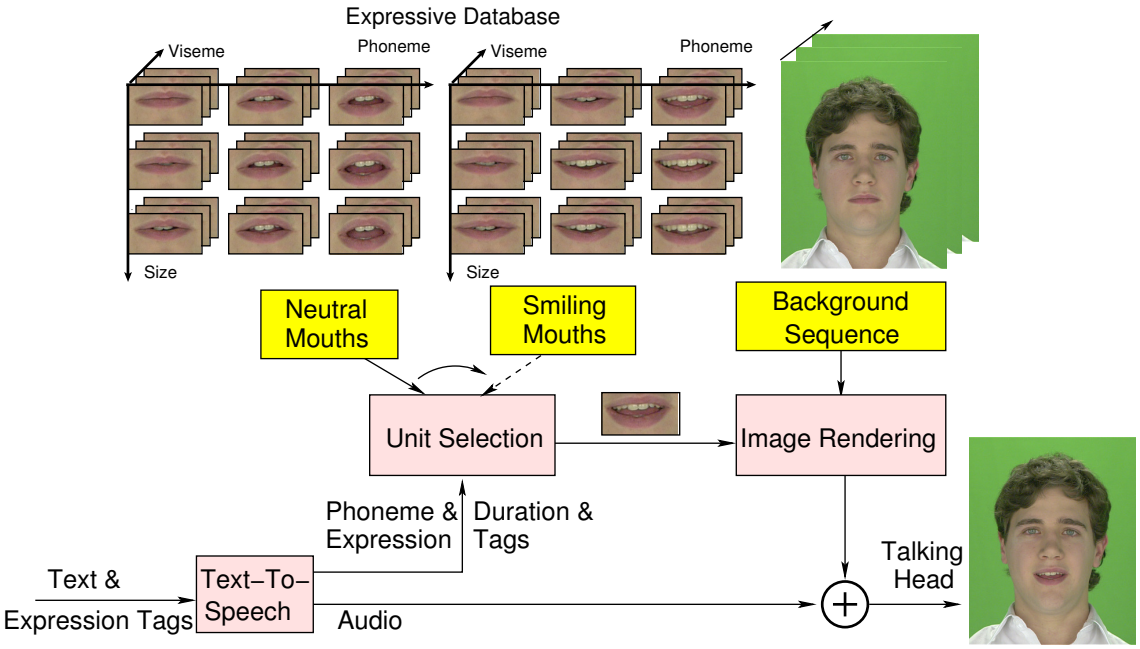


Figure 3.1: Image-based talking head extended with facial expressions.

In this chapter, we first describe the creation of the expressive database. In order to synthesize realistic and smooth facial expressions, natural expression changes and viseme transitions while changing expressions are analyzed in section 3.2. In section 3.3, a modified unit selection for facial expressions is introduced.

3.1 Creation of an Expressive Database

Database building is one of the most important components for generating a realistic talking head with expressions, since image-based talking head synthesis depends heavily on the mouth images in the database. In this thesis, we are animating smiles as an example, since other facial expressions, like sadness and anger, will be straightforward to integrate.

In our studio, a native speaker is recorded while reading a corpus of 150 sentences. These sentences are designed to find a tradeoff between the English phoneme coverage and the size of the corpus. A lighting system is developed for the audio-visual recording, which minimizes the shadow on the face of a subject and hence reduces the change of illumination on the moving head. In our experiment, each sentence is recorded three times with different emotions and expressions, i.e. speaking without any expression (neutral speaking), smiling after speaking, and smiling while speaking. The recording is captured by using a HD camera (Thomson LDK 5490). The video format is originally 1280×720 at 50 fps, which is cropped to 576×720 pixels at 50 fps. The audio signal is sampled at 48 kHz. A total of 450 utterances are recorded to build the database, which contains $78,797$ normalized mouth images with a resolution of 288×304 pixels. A snapshot of example images extracted from the sequence is shown in Fig. 3.2.

(a) Neutral face (b) Smiling face

Figure 3.2: Snapshot of example images extracted from recorded videos. The neutral face (a) and the smiling face (b) are labeled with the same phoneme “ng" of the word “morning".

After recording, the audio data and the texts are aligned by an aligner, which segments the phonemes of the audio data. The aligner is first trained for specific voices in order to

increase the accuracy of the alignment given the recorded audio and the spoken text. Once the aligner has been trained, it produces a timed sequence of phonemes. Therefore, for each frame of the recorded video, the corresponding phoneme and its phoneme context are known. The phoneme context is required in order to capture coarticulation effects.

Using a 3D head scan of the recorded speaker, the recorded videos are processed by model-based motion estimation [66], which estimates the head motion parameters for each frame. The 3D head scan is a 3D face representation, which is a polygon mesh consisting of a collection of vertices and polygons that define the shape of a face in 3D. The model-based estimation algorithm stores texture information of the object and tries to find the motion parameters of the rigid head in a new frame by minimizing the difference between the textured 3D model and the face in the new frame. Using the motion parameters, all frames are normalized to a reference position and the mouth regions of all frames are cropped to build an expressive database containing two expression states: neutral and smile.

Normalized mouth images are transformed into the PCA space [77], so that few parameters are needed to describe the appearance of the mouth image. The shape of the mouth images is extracted by AAM [25], from which geometric parameters such as mouth width and height are derived.

The expressive database is built with a large number of normalized mouth images. Each image is labeled with the following parameters: an expression tag, appearance parameters (PCA), geometric parameters (mouth width and height, the visibility of teeth and tongue), phonetic context, original sequence and frame number.

3.2 Analysis of Viseme Transitions while Changing Expressions

In order to analyze viseme transitions while changing expressions, we firstly analyze the natural expression change of humans, and then analyze viseme transitions in the expressive database. The analytical results are used in the expressive unit selection.

3.2.1 Natural Expression Change of Humans

In order to control facial expressions, the behavior of humans has to be analyzed. In this thesis, we conduct facial expression “smile” as an example, and for this reason, we analyze natural smiles of human from recorded videos. The goal of this analysis is to measure when humans begin and end a smile and to which extent humans smile in which situation.

Our facial animation is aiming to generate realistic facial expressions. For this purpose, the recorded video requires natural facial expression while speaking. The recorded videos for database building are unsuitable for analysis of expressions, since the speaker is aware

of the recording and their expression is controlled. Therefore, videos of spontaneous speech and conversations are required.

In order to analyze natural smiles, three video sets are collected:

1. Recording of special texts (about 10 minutes): These special texts are able to make the speaker laugh when he is reading, the special texts are like "schuettelverse", i.e., poems of two lines embedding spoonerisms.

2. Spontaneous speech and conversation (about 10 minutes): The most natural facial expression is from the spontaneous conversation, where the speaker is unaware of expression change.

3. Videos of a news reporter (10 minutes): A news reporter is trained to give facial expressions and know when and how to smile. A news reporter is a typical application for talking head.

Based on the collected videos, we are able to find some simple rules for facial expressions. In the first video set, the speaker cannot smile at the beginning of the sentence. For example, sometimes the speaker begins to smile at the end of the schuettelverse sentence, because the speaker understands the meaning after reading the sentence. In the second video set of conversation, facial expression depends tightly on current topic and partner's reactions. Different partners have different facial expressions for the same topic. It is also very difficult to be determined when a smile begins while speaking. Starting a smile is very individual realistic, since one person will laugh earlier while another laughs later. However, several rules are also general in the conversational video. A smile begins before speaking and ends after the whole sentence. This happens when the speaker tries to answer a funny question. In some cases, a smile begins somewhere in the sentence, when the meaning changes from neutral to funny. Generally, these changes are almost with a short pause in the sentence. These pauses are always between words or clauses. Someone smiles when he is listening, which gives the partner some information whether the story is funny or he has understood the words very clearly. In the third video set, the news reporter smiles in most cases at the end of the sentence, which gives audience some hints or make the news report friendly.

Another issue is the degree of the smile. A smile or a neutral speaking appears at the end of news, which depends on the speaker. Generally, smiles from the news speaker videos are small and begin from the last sentence of news. Depending on the content, a big or small smile could be at the end of speaking. We have observed that these smiles always start before the last syllable. In most cases, big smiles happen at the listening state without speaking in the conversational videos. Smiles of different degrees are shown in Fig. 3.3.

The spontaneous conversational video is used to analyze the statistics in the above mentioned different kinds of smiles. The results are summarized in Table 3.1. Given Table 3.1, 43% of smiles appear at the last syllable of a sentence. 24% of smiles are

Figure 3.3: Smiles with different degree. The left image is a neutral speaking face. The middle image (small smile) is a speaking face with smile. The right image (big smile) is a smiling face without speaking.

accompanied with speaking. 20% of smiles happens between clauses. Big smiles always happen without any speaking. In a few cases (5% of smiles), smiles are occurring between words.

Table 3.1: *Frequency of different kinds of smiles in a conversational video.*

No.	kind of smile	frequency	%
1	smile between words	2	5%
2	smile between clauses	7	20%
3	smile during speaking	9	24%
4	smile from the last syllable	16	43%
5	big smile without speaking	3	8%

In order to generalize our smiling talking head to all other facial expression synthesis, we have observed that the transition between different facial expressions, such as smile and angry, are through neutral state. In this way, our smiling talking head is easy to be extended to synthesize all other facial expressions. Transition of any two facial expressions is realized by finding a smooth path through neutral mouths.

Facial expression synthesis is not simply defined as mouth animation. Smiles sometimes have an impact on the deformation of the eye parts, even though a smile is formed by flexing the muscles near both ends of the mouth. These kind of smiles are called a Duchenne smile, which involves contraction of both the zygomatic major muscle (which raises the corners of the mouth) and the orbicularis oculi muscle (which raises the cheeks and forms crow's feet around the eyes). Many researchers believe that Duchenne smiles indicate genuine spontaneous emotions since most people cannot voluntarily contract the

outer portion of the orbicularis oculi muscle [78]. In this chapter, we are focusing on the expressive mouth animation. Eye animations, such as eye gazing and eye blinding, are developed by Weissenfeld et al [28], which can be directly integrated in our system.

In the analysis of natural smiles, we can summarize general cases for a smile as follows:

- smile begins before speaking and ends after speaking.
- smile begins at the end of speaking and before the last syllable.
- smile begins between words or clauses in the speaking, where the contents change from neutral to funny, generally a short pause occurs before expression changes.
- big smiles are more often the reaction to the conversational partner without speaking, and at the same time eye animation is required for big smiles.

3.2.2 Viseme Transitions of Expressive Database

In order to generate smooth transitions between the neutral and the smiling mouths by using the expressive database, the relationship between the neutral and the smiling mouths has to be analyzed. The goal of this analysis is to show how well the viseme transition can be done for different visemes.

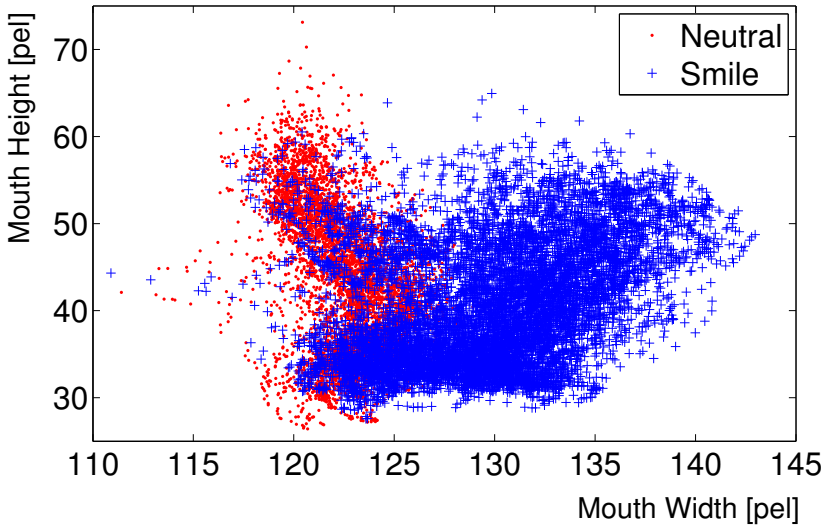
The neutral and the smiling mouths have different mouth shapes and mouth appearances. Fig. 3.4 shows the distribution of mouth width and height for the viseme 1 and 7. Even though the average mouth height of neutral viseme and smiling viseme is similar, the average width of the neutral mouth images is clearly smaller than the average width of smiling mouth images. The mean value and standard deviation of mouth width and height for each viseme are depicted in Fig. 3.5. Given Fig. 3.5, mouth height is similar for both neutral and smiling mouths, hence the mouth height will have little impact on the optimal switching position. However, smiling mouths are broader than neutral mouths. Therefore, the availability of mouth images with similar width will have a major impact on the switching position between the neutral and the smiling mouths. Given the non-overlapping intervals of average mouth width plus/minus standard deviation in Fig. 3.5(a), only few images are suitable for switching.

The appearance features for viseme 1 and 7 are plotted in Fig. 3.6, where only the first two principal components are shown. The area covered by smiling images appears to be a superset of the area of the neutral images in this PCA diagram. We assume that the N-dimensional volume spanned by the PCA components of smiling images has a large overlap with the volume of the neutral images. Hence, smooth transitions between visemes are possible, when considering PCA features.

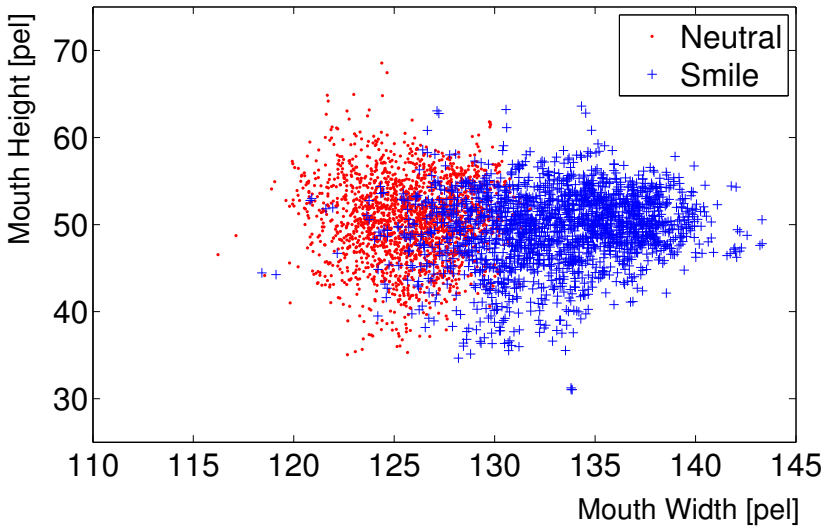
The switching ability from viseme v_i to viseme v_j, is measured in the following way:

$$p_{v_i,v_j} = \frac{m_{i,j}}{N_{v_i}} \tag{3.1}$$

$$p_{v_j,v_i} = \frac{m_{j,i}}{N_{v_j}} \tag{3.2}$$

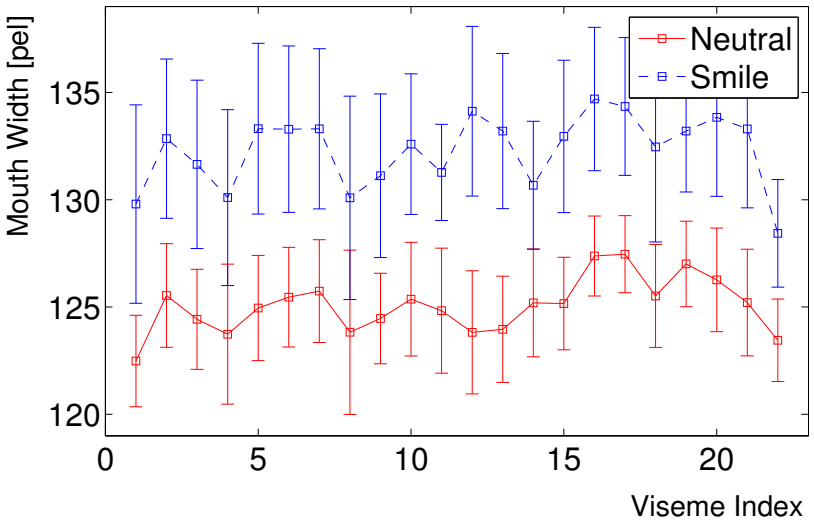


(a) Viseme 1 (silence)

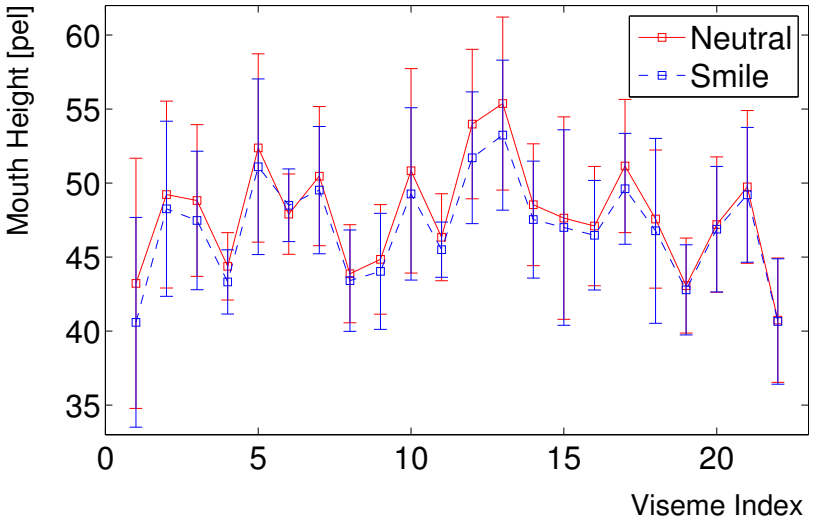


(b) Viseme 7 (iy, y, ih, ix)

Figure 3.4: Distribution of mouth width and height in geometric feature space. (a) is for viseme 1. (b) is for viseme 7. Red points are from the neutral mouths, blue points are from the smiling mouths.

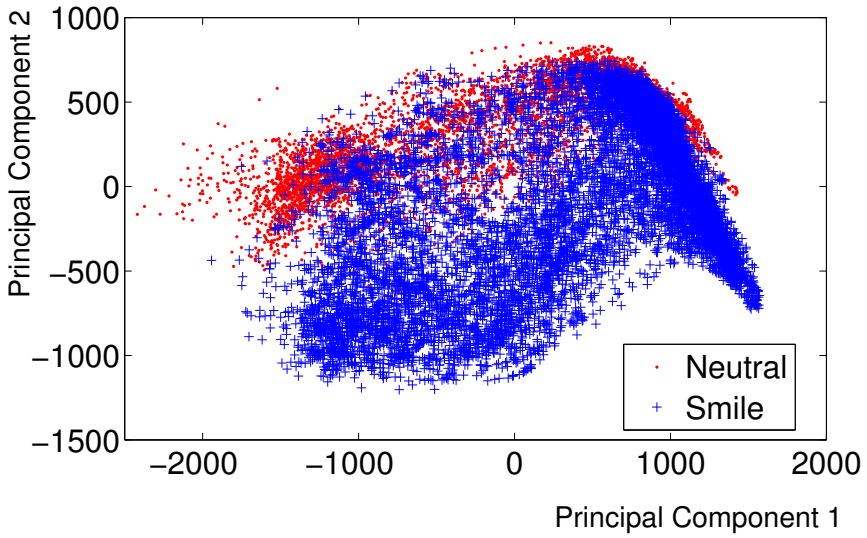


(a) Mouth width

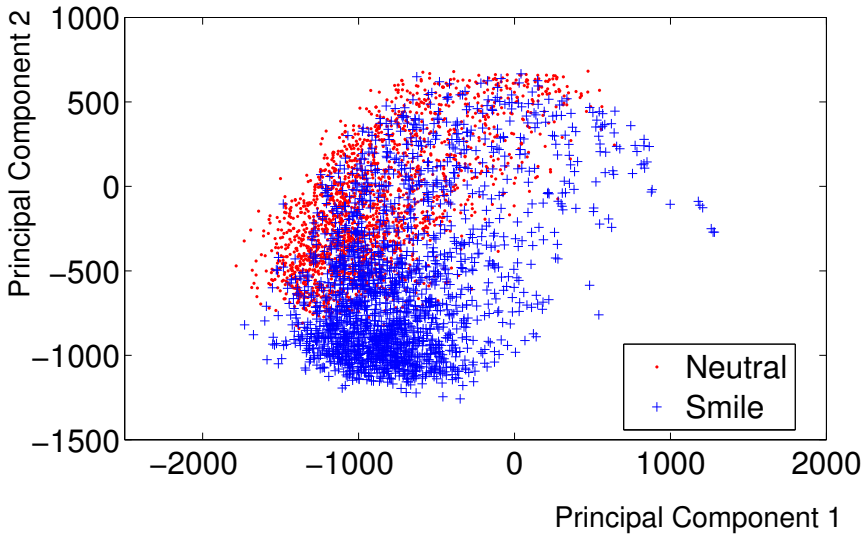


(b) Mouth height

Figure 3.5: Analysis of geometric features of neutral and smiling mouths. The mean and standard deviation are calculated for each viseme. The red line is the mean for neutral mouths, and the blue line is for a smile. The mouth width for a smile is significant in the geometric feature space.



(a) Viseme 1 (silence)



(b) Viseme 7 (iy, y, ih, ix)

Figure 3.6: Distribution of mouth images in PCA space. (a) is for viseme 1. (b) is for viseme 7. Red points are from the neutral mouths, blue points are from the smiling mouths.

where N_{v_i} is the number of mouth images of viseme v_i and N_{v_j} is the number of mouth images of viseme v_j. $m_{i,j}$ is the number of mouth images in v_i, which can find a neighbor from v_j so that the distance is smaller than a predefined threshold. The threshold is chosen manually in a subject test. If the distance between two images is smaller than the threshold, transitions between two images are smooth. Switchable images are images with one expression, which have a similar neighbor with another expression. Their distance is less than the threshold. In our experiments, the threshold is manually determined in the PCA space. Fig. 3.7 shows the switching matrix of viseme transitions. Due to the higher p_{v_i,v_j} value, switching from neutral to smile is easier than reverse.

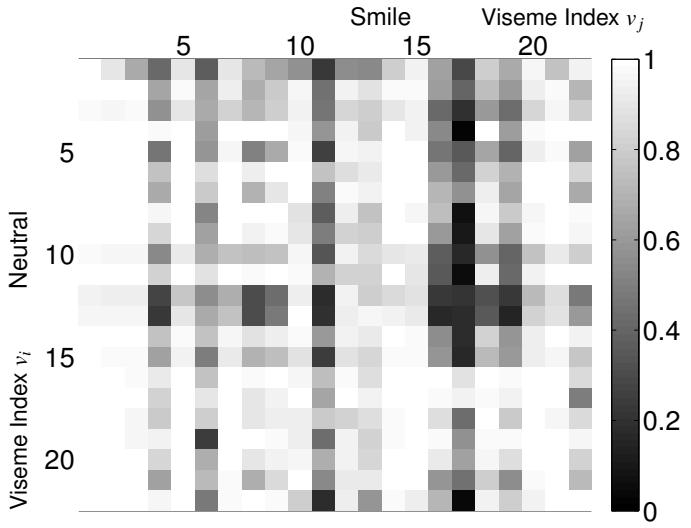
Evaluating the whole database, we have measured that 65% of the neutral mouth images are able to find at least one similar smiling mouth to switch to, while only 25% of the smiling mouth images can switch to the neutral mouths smoothly. For each viseme the geometric features of these mouth images are plotted in Fig. 3.8. In Fig. 3.8(a), the average mouth widths of the switchable neutral and smiling mouths approach closely. Furthermore, the large overlap intervals of the average mouth width with plus/minus standard deviation indicate that the smiling mouth images with small widths can be switched to the neutral mouths smoothly. In Fig. 3.8(b), we can see that the average mouth height of the switchable neutral mouths and the switchable smiling mouths is very low, compared to the average mouth height of the whole neutral and smiling mouths (gray curves). Therefore, the switch tends to happen at the boundary of a viseme, when the switchable neutral and smiling mouths are almost closed.

3.3 Expressive Unit Selection

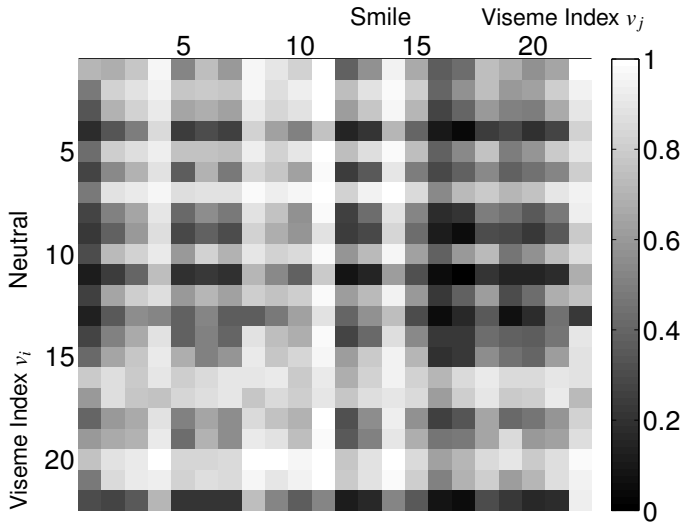
The expressive unit selection uses weighted target costs and concatenation costs. The target cost measures the lip synchronization of the mouth image. The concatenation cost measures the smoothness of the transition from one image to another. The target cost is computed by calculating the distance between the phoneme context of the mouth image and the input phoneme context. The concatenation cost of two images is computed by calculating the weighted geometric and appearance distance of these two mouth images. Thus the larger the database is, the more units the unit selection can choose from, which make it easier to find matching units for a given phoneme sequence. A Viterbi search finds optimal units from the database by minimizing the two types of costs.

Expression tags are used to control when the talking head begins and ends a smile while speaking. The syntaxes for facial expression synthesis are listed in Table 3.2.

For expression mode 1 the syntax is used to generate animation that speaks the text with smile expression. The expression tag can be inserted at any position inside the sentence or for the whole sentence. By using the expression tag, the unit selection is able to select the mouth images by switching between neutral and smiling mouths. In expression mode 2, the syntax is designed for appending a smile pattern at the end of speaking, or for a smile without speaking. The intensity of the desired smile pattern can be specified by

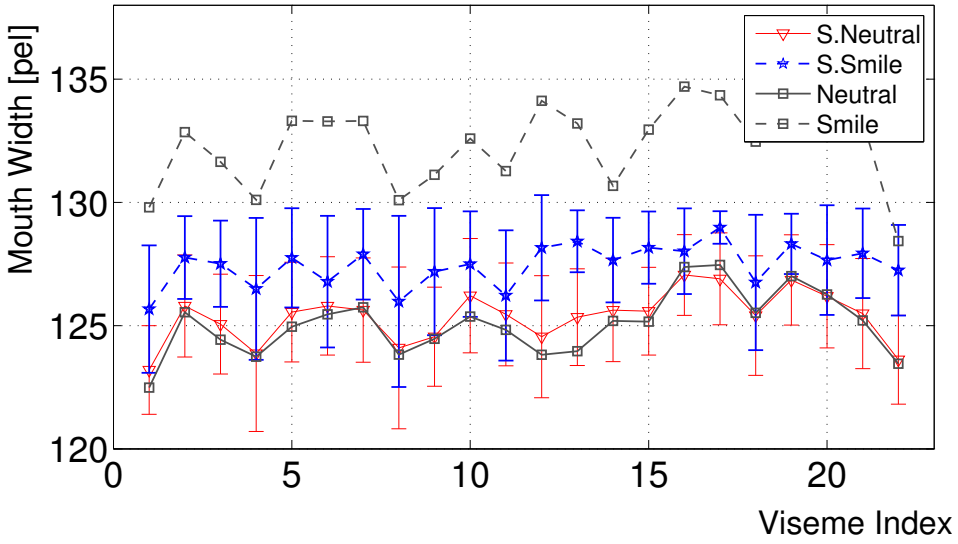


(a) Viseme switching ability p_{v_i,v_j} from neutral to smile

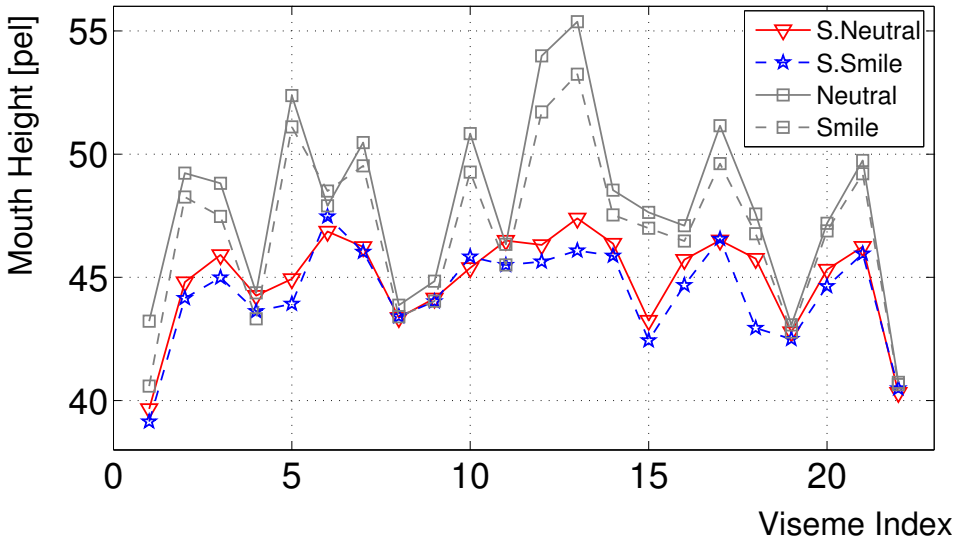


(b) Viseme switching ability p_{v_j,v_i} from smile to neutral

Figure 3.7: Switching matrix between neutral and smiling visemes. (a) Viseme switching from neutral to smile, (b) Viseme switching from smile to neutral. Switching from neutral viseme to smiling viseme is generally higher.



(a) Mouth width



(b) Mouth height

Figure 3.8: The average mouth width and height of the switchable mouth images of each viseme measured for the neutral and the smiling mouths, respectively. The red line is the average width and height of the switchable images from the neutral mouths, the blue dash line for the average width and height of the switchable images from the smiling mouths. As a reference, the gray lines are the average width and height of all images of a viseme from the neutral and the smiling mouths, respectively (as in Fig. 3.5).

Table 3.2: Syntaxes of facial expression synthesis.

expression mode	syntax
1. speaking with smile	<prosody expression = “smile”> text </prosody>
2. smile without speaking	<smilepattern size = “small | medium | large”/>

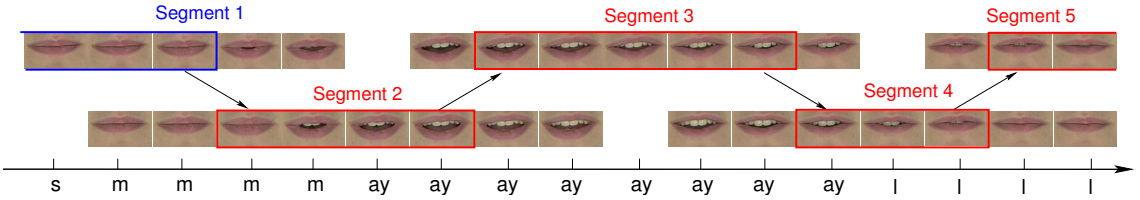


Figure 3.9: Part of the unit selection for the last word “smile” of the sentence “I can also speak with an expression, like a smile.”. The mouth image video segments are selected from the neutral and smiling mouths, depending on the input expression tag. The position of transition from one segment to the next is determined by the concatenation costs of the unit selection. Segment 1 is selected from neutral mouths, the segments after that are selected from smiling mouths. In this example, switching occurs from the neutral phoneme ‘m’ to the smiling phoneme ‘m’, when the mouth is closed.

the parameter “size” of the smile pattern tag. Smile patterns are categorized into three classes: small smile, medium smile and large smile. The syntax can also be extended to synthesize other facial expressions easily.

Based on the results of analyzing natural expression change, four possible natural talking heads can be synthesized by the expressive database for a given sentence:

- Case 1: Talking head speaks without any expression.
- Case 2: Smile is accompanying speaking.
- Case 3: Facial expression is changed from neutral to smile during speaking.
- Case 4: A smile pattern is appended to an animation for non-verbal communication.

For case 1, the unit selection selects mouth images only from the neutral mouths and for case 2, from the smiling mouths only with no need for switching the database. Case 4 is for non-verbal communications, which is easily done by appending one of the smile patterns to an animation. The focus of the expressive unit selection is on case 3, where the database has to be switched from the neutral mouths to the smiling ones while speaking, or vice versa.

For case 3, mouth images are selected from the expressive database by switching between neutral and smiling mouth images. Based on the switching matrix of viseme tran-

sitions, the viseme transition with lowest value around the input expression tag is determined as an initial switching position. In order to achieve a very smooth transition while speaking and changing expressions, units from both expressions with small concatenation costs have to be found. With low concatenation cost, the unit selection can generate smooth animations.

Fig. 3.9 shows a sequence of mouth images that the unit selection considers for the mouth animation. The time line is labeled with phonemes. According to the target cost, a segment is selected from the database. According to the concatenation cost, optimal transitions between segments are determined. Depending on the input expression tag, the unit selection selects appropriate mouth images from the database containing the corresponding expression. We have observed that viseme switching from neutral to smile is always occurring when the mouth is almost closed.

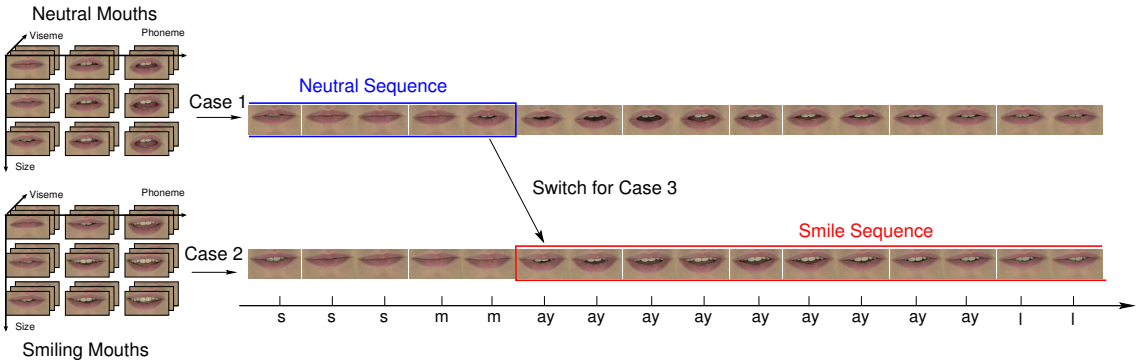


Figure 3.10: Illustration of the expressive unit selection. The top row shows the animated sequence is selected from the neutral mouths. The bottom row shows the animated sequence is selected from the smiling mouths. The expressive unit selection tries to switch from the neutral mouths to the smiling ones at the optimal position. This is an another view of the expressive unit selection for Fig. 3.9.

The relationship of the expressive unit selection for Case 1, 2 and 3 are shown in Fig. 3.10. The top row of Fig. 3.10 is for Case 1, where a neutral mouth image sequence is selected, and the bottom row gives the smiling mouth image sequence of Case 2. The mouth image sequence for Case 3 is the sequence that switching from Case 1 to Case 2 depending on the input expression tag.

4 Head Motion Synthesis

Generating realistic talking heads is not limited to the field of lip synchronization and facial expressions. Especially for longer animations, a lack of natural head movements limits the naturalness of the animations. In this section, we will introduce a novel approach to generate head movements.

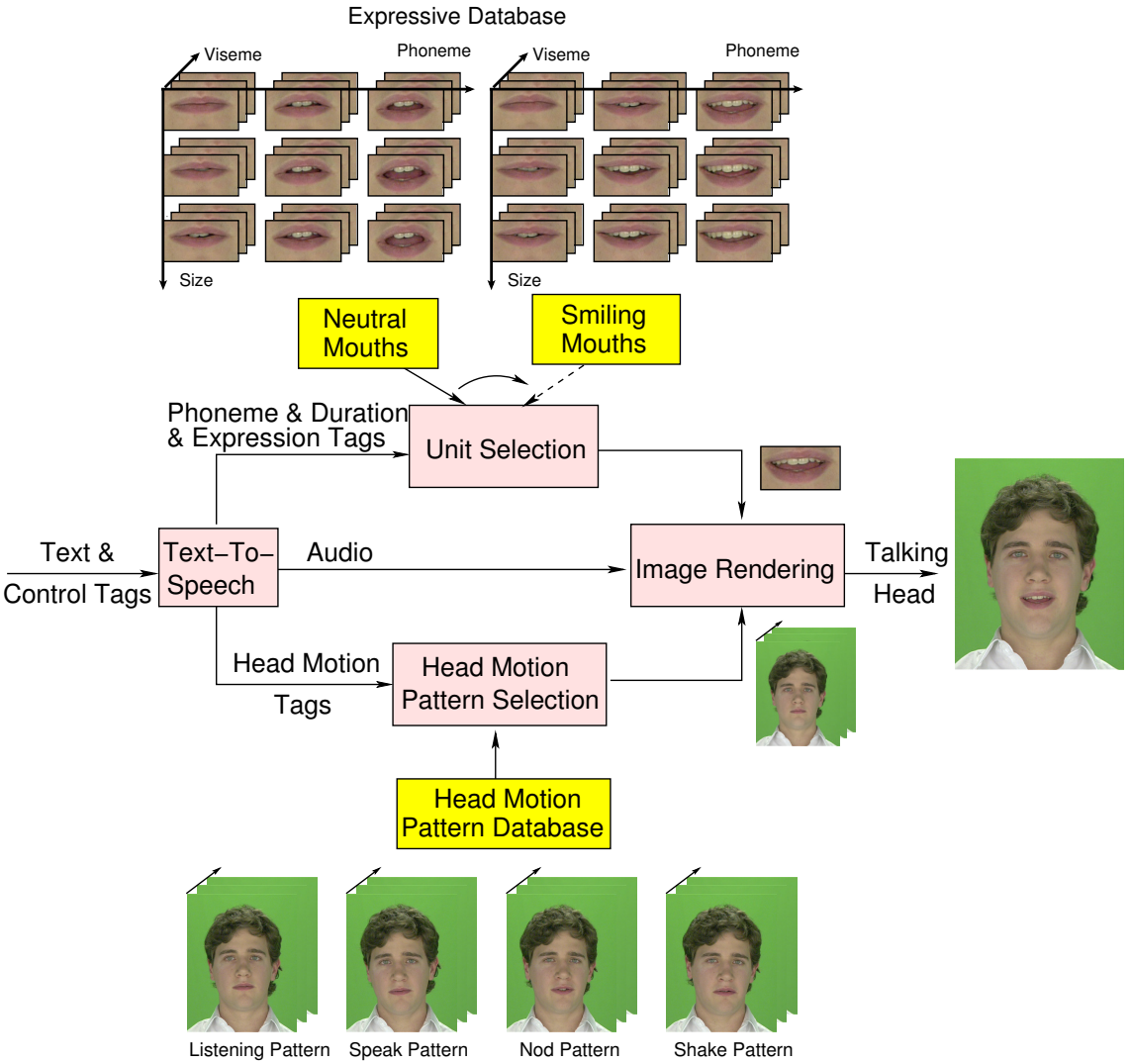


Figure 4.1: Expressive talking head extended with head motion synthesis.

Fig. 4.1 shows the expressive talking head extended with a flexible head motion synthesis. The head motion synthesis is driven by the input control tags. According to the head motion tags like nodding and shaking, head movements can be synthesized by selecting and concatenating appropriate head motion patterns from the head motion pattern database. The transitions between these patterns are smoothed by the optical flow based morphing technique.

4.1 Collection of Head Motion Patterns

In order to analyze realistic head motion of the speaker, we have recorded a face to face conversation of about 20 minutes with spontaneous head and facial movements. The speaker was unaware that the head movements were relevant.

To identify head motion patterns, rotation and translation parameters are estimated by the model-based approach [66]. We have observed that during the conversation, the head moves away from and back to the straight forward looking position. The first frame of one pattern is manually defined. The end frame is found by computing the minimal distance to the first frame. Instead of using the whole images as in [79], the distance of two frames is measured by calculating the Euclidean distance between the two frames in the pose space in the following way:

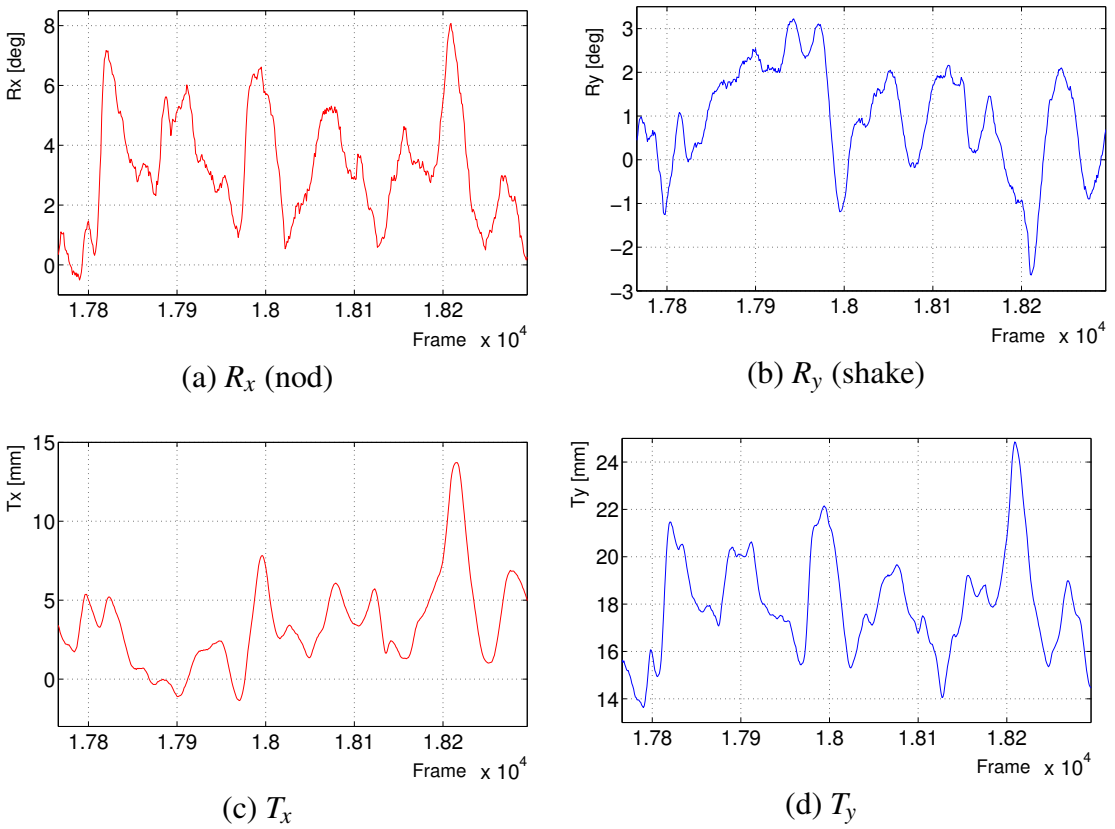
$$D_{ij} = w_{rt} \left\| R_i - R_j \right\|_{L_2} + \left\| T_i - T_j \right\|_{L_2} \tag{4.1}$$

where R_i and R_j are rotation parameters of the head in the image i and j, T_i and T_j are translation parameters for the image i and j. R is represented for the vector (R_x, R_y, R_z), T for (T_x, T_y, T_z). w_{rt} is a scale factor that controls the impact of the two terms.

Distance in the pose space can be computed accurately and quickly. The local deformation of the mouth part is not considered in the measurement because the mouth part is replaced by a new mouth in the animation rendering. Finally, the head motion patterns are extracted from the video sequence.

In head motion analysis, we have collected 29 head motion patterns, which are classified as listening patterns (6), speaking patterns (15), and idle patterns (8). The average duration of these patterns is about 14 seconds long. Four additional head motion patterns are also collected as semantic patterns, such as nods (2) and head shaking (2).

Fig. 4.2 shows the pose parameters of a speaking pattern, which is extracted from the conversation. Fig. 4.2 (a) and Fig. 4.2 (b) are the trajectories of the rotation around axis x (head nod) and y (head shake). The motion around z is almost zero during the conversation. Fig. 4.2 (c) is the trajectory of the translation of the head along x in the world coordinate system parallel to the image plane. In this figure we can see that the rotation of the face begins and ends almost at the zero orientation. This speaking pattern is about 11 seconds long.



(a) R_x (nod) (b) R_y (shake)

(c) T_x (d) T_y

Figure 4.2: Examples of pose parameters (R, T) of a speaking pattern. (a) rotation around x, (b) rotation around y, (c) translation in x, (d) translation in y.

Parameterization of Head Motion Patterns

The head motion patterns should be parameterized, so that the head motion pattern selection is able to select the background sequence with appropriate head movements. The motion parameters, like speed v and acceleration a of the head motion, can be calculated from the pose parameters (R, T). Moreover, each head motion pattern is labeled with a head motion tag, such as speaking, listening, idleness, nodding and shaking. Two extra head motion patterns (start and end head motion patterns) are the sequences with a forward-looking head pose. The start and end head motion patterns are used as the beginning and the end of the background sequence, which makes the head look forward after animation.

Since the sampling rate of our recording is 50 Hz, the speed can be approximated very well by using numerical differentiation. The speed of the rotation around the x-axis is

given by:

$$v_{R_x}(i) = (R_x(i+1) - R_x(i))/\Delta t \tag{4.2}$$

where $R_x(i)$ is the rotation R_x of frame i, $R_x(i+1)$ for frame $i+1$. Δt is the time interval of the sampling rate, here $\Delta t = 20ms$. The speeds of rotation around the y-, and z-axis, respectively, are computed in the same way:

$$v_{R_y}(i) = (R_y(i+1) - R_y(i))/\Delta t \tag{4.3}$$
$$v_{R_z}(i) = (R_z(i+1) - R_z(i))/\Delta t \tag{4.4}$$

The acceleration of the rotation R_x, R_y, and R_z is given by:

$$a_{R_x}(i) = (v_{R_x}(i+1) - v_{R_x}(i))/\Delta t \tag{4.5}$$
$$a_{R_y}(i) = (v_{R_y}(i+1) - v_{R_y}(i))/\Delta t \tag{4.6}$$
$$a_{R_z}(i) = (v_{R_z}(i+1) - v_{R_z}(i))/\Delta t \tag{4.7}$$

The speed and acceleration of the translation are calculated as:

$$v_{T_x}(i) = (T_x(i+1) - T_x(i))/\Delta t \tag{4.8}$$
$$v_{T_y}(i) = (T_y(i+1) - T_y(i))/\Delta t \tag{4.9}$$
$$v_{T_z}(i) = (T_z(i+1) - T_z(i))/\Delta t \tag{4.10}$$
$$a_{T_x}(i) = (v_{T_x}(i+1) - v_{T_x}(i))/\Delta t \tag{4.11}$$
$$a_{T_y}(i) = (v_{T_y}(i+1) - v_{T_y}(i))/\Delta t \tag{4.12}$$
$$a_{T_z}(i) = (v_{T_z}(i+1) - v_{T_z}(i))/\Delta t \tag{4.13}$$

Database of Head Motion Patterns

The head motion pattern database contains a number of head motion patterns. Each of them is labeled with a head motion tag, like speaking, listening, nodding and shaking of the head. Furthermore, each frame of these patterns is characterized by the pose and motion parameters (R, T, v, a). A start pattern and an end pattern are also required for the initial and final states.

4.2 Generation of Background Sequences with Flexible Head Motions

In order to generate background sequences, the head motion patterns are selected from the database of head motion patterns according to the input head motion tags. Then these

patterns are concatenated by a morphing technique, which smooths the transitions and does not introduce any visual discontinuities.

4.2.1 Selection of Head Motion Patterns

We select head motion patterns according to the input head motion tags and the desired pattern lengths. Table 4.1 gives the syntaxes that are used in the selection of head motion patterns. According to the database of head motion patterns, a “strong” tag is used to select head motion patterns labeled with speaking which have a strong head motions when speaking. A “medium” tag is used to select head motion patterns labeled with listening, which have medium head motions. An “idle” tag instructs the unit selection to insert a pause in the animation. “Nod” and “shake” tags are used for selecting head motion patterns in a semantic level.

Table 4.1: Syntaxes of selection of head motion patterns.

attribute	syntax
strong	<prosody motion=“strong”> text </prosody>
medium	<prosody motion=“medium”> text </prosody>
idle	<head_motion_break>
nod	<prosody motion=“nod”> text </prosody>
shake	<prosody motion=“shake”> text </prosody>

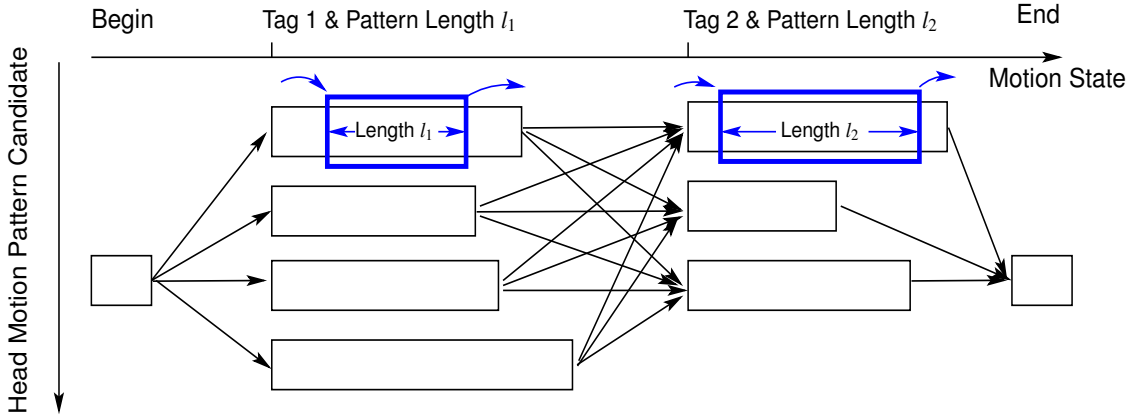


Figure 4.3: Unit selection of head motion patterns for generation of background sequence.

Head motion patterns are selected in the following way as shown in Fig. 4.3. The inputs are the tags and the pattern lengths. The pattern length is derived from the TTS output, matching the length of the output audio data. According to the input head motion tag, a list of candidate patterns, labeled with the same tag as input, are selected from the

database. The beginning and end motion patterns are added to the initial and final states. All the patterns are connected from left to right, so that a search graph is built. In order to determine the optimal path, costs are assigned to transitions. A window is moving over each pattern. The length of the window matches the input pattern length. The cost function for transition considers the pose and motion difference between the last frame of the current moving window and the first frame of the succeeding moving window in the search graph. If these two frames have similar parameters (R, T, v, a), their transition cost is small. The concatenation cost of two patterns is computed as:

$$CC_{motion} = w_{rt} \left\| R_i - R_j \right\|_{L_2} + \left\| T_i - T_j \right\|_{L_2} + w_v \left\| v_i - v_j \right\|_{L_2} + w_a \left\| a_i - a_j \right\|_{L_2} \quad (4.14)$$

where (R_i, T_i, v_i, a_i) is the pose and motion parameters for the last frame of the current window i, (R_j, T_j, v_j, a_j) for the first frame of the succeeding window j. Weights w_{rt}, w_v, and w_a are the discrimination factors that influence the pose and motion difference. The concatenation cost between two patterns is the Euclidean distance in the pose and motion space. A Viterbi search is used to find the optimal path with minimal costs.

The selection of head motion patterns is summarized in three steps:

- Step 1: Inserting head motion pattern candidates into each motion state, according to the input motion tags.
- Step 2: Assigning CC_{motion} for each transition between the moving windows and building a search graph.
- Step 3: Finding the optimal path with minimal costs by using Viterbi search.

4.2.2 Transition of Head Motion Patterns

Once the unit selection of head motion patterns has found the appropriate path, the selected head motion patterns should be concatenated to generate the background sequence for facial animation. In order to concatenate these head motion patterns without creating noticeable discontinuities, the technique of morphing the whole face is used. Given two images I_0 and I_1, morphing is used to interpolate intermediate images I_α, where α is a parameter ranging from 0 to 1. Since the forward looking head at the joint of the patterns has similar position and orientation, only a few frames are required to smooth the head motion when concatenating these patterns. The number of interpolated frames depends on the difference of the head pose in the joint position of two patterns. In order to create morphing sequences without blending artifacts, pixels in image I_0 have to be accurately mapped onto corresponding pixels in image I_1. We identify these correspondences using a high accuracy optical flow estimation algorithm based on a theory for warping [61], which integrates three assumptions into an energy function for computing optical flow. These are

a gray value constancy assumption, a gradient constancy assumption, and a discontinuity-preserving spatio-temporal smoothness constraint. This optical flow approach is suitable for images with large displacements.

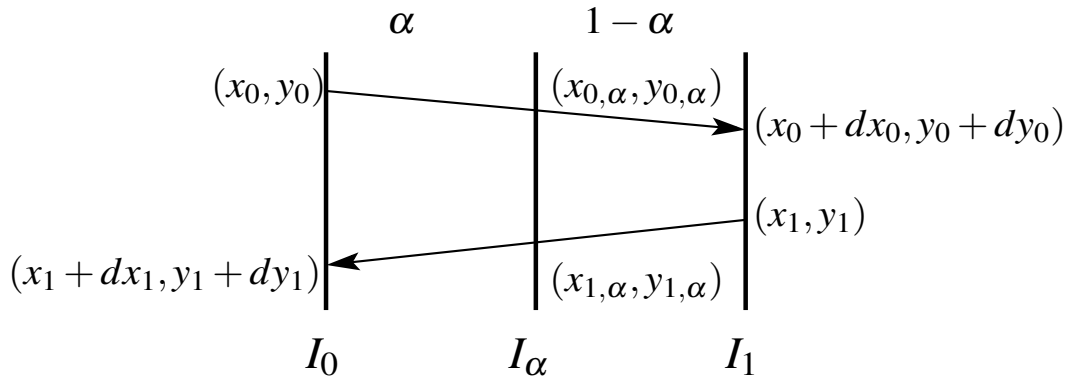


Figure 4.4: Image interpolation by morphing.

The interpolated images are generated in three steps and are shown in Fig. 4.4:

- Step 1: Finding correspondences in two images by optical flow.
- Step 2: Forward warping I_0 towards I_1 and backward warping I_1 towards I_0 along the computed flow vectors.

  The pixel value of the forward warped image $I_\alpha^{Forward}$ is calculated as:

$$x_{0,\alpha} = x_0 + \alpha d_{x_0} \tag{4.15}$$
$$y_{0,\alpha} = y_0 + \alpha d_{y_0} \tag{4.16}$$
$$I_\alpha^{Forward}(x_{0,\alpha}, y_{0,\alpha}) = I_0(x_0, y_0) \tag{4.17}$$

  where (d_{x_0}, d_{y_0}) is the motion displacement of the point (x_0, y_0).

  The pixel value of the backward warped image $I_\alpha^{backward}$ is calculated as:

$$x_{1,\alpha} = x_1 + (1-\alpha) d_{x_1} \tag{4.18}$$
$$y_{1,\alpha} = y_1 + (1-\alpha) d_{y_1} \tag{4.19}$$
$$I_\alpha^{Backward}(x_{1,\alpha}, y_{1,\alpha}) = I_1(x_1, y_1) \tag{4.20}$$

  where (d_{x_1}, d_{y_1}) is the motion displacement of the point (x_1, y_1).

- Step 3: Cross-dissolving the warped images to produce the final image I_α.

  The final morphed image is computed in the following way:

$$I_\alpha = (1-\alpha) I_\alpha^{Forward} + \alpha I_\alpha^{Backward} \tag{4.21}$$

In case of head rotation, the points on the face move along an arc on the surface of a sphere. Fig. 4.5 shows the linear approximation of head motion. The point A on the face moves to point B while the head rotates around the axis Z that is vertical to the plane $X-Y$. The real trajectory of the motion from A to B is the arc $\widehat{AB}$. If the rotation angle is small enough, the line $\overline{AB}$ can approximate the arc $\widehat{AB}$. Since the face has a front view and is recorded at 50 frames per second, the rotation between frames is small and the approximation is valid.

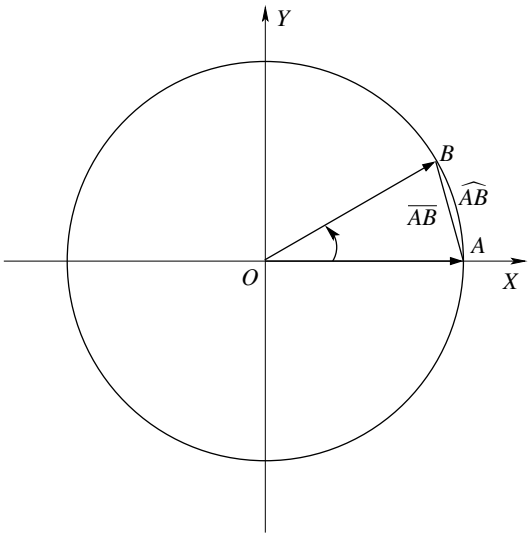


Figure 4.5: Linear approximation of head motion.

We select two images with large head motion displacements from the background video database to evaluate the proposed morphing approach. Fig. 4.6 shows the results of optical flow and interpolated image.

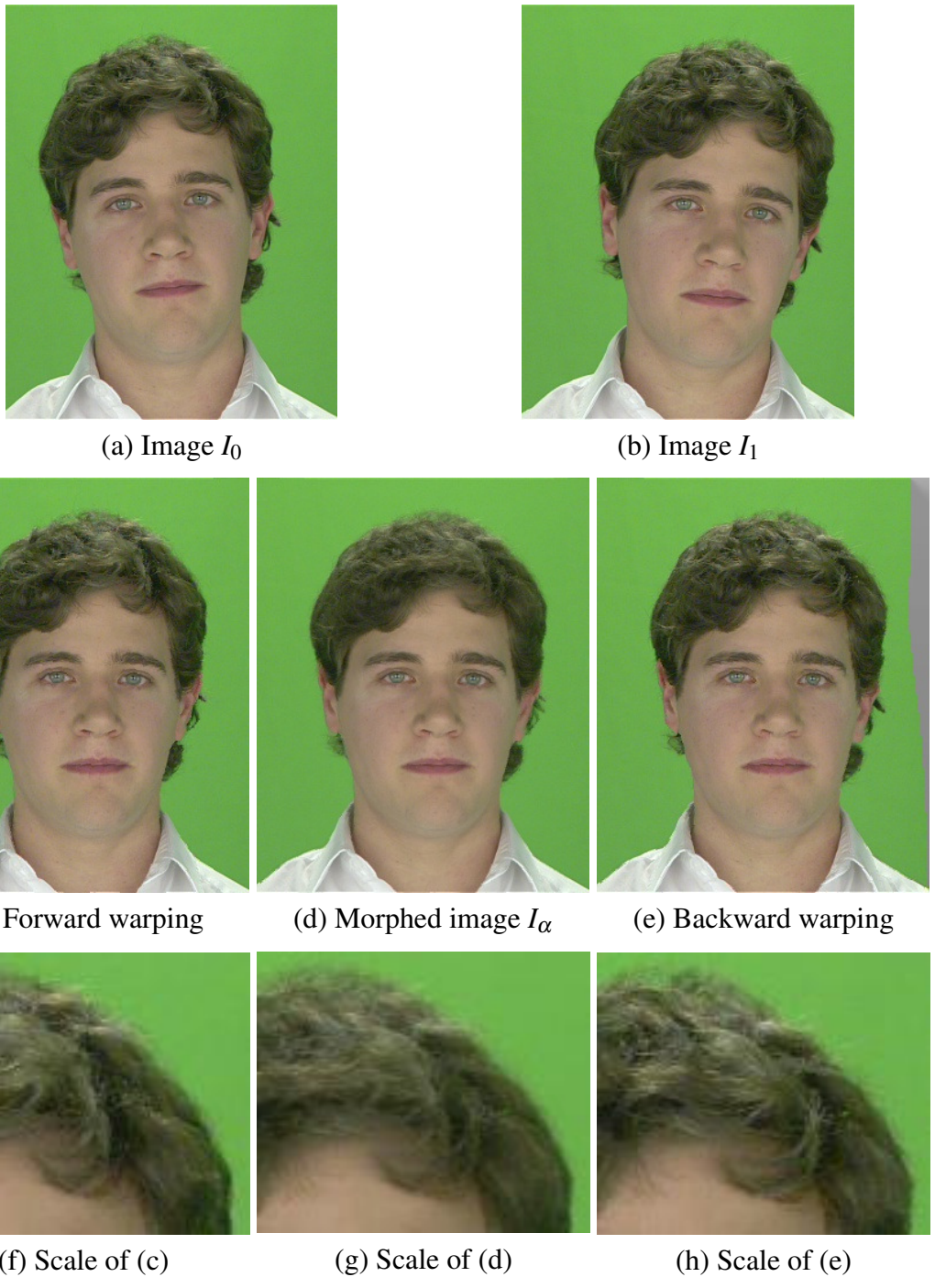
The vector field of the optical flow is evaluated by local consistency error (LCE), which measures the squared Euclidean distance at each pixel after composing the forward and backward warping between two images as shown in Fig. 4.7. The LCE is calculated in the following way:

$$LCE(x,y) = \sqrt{(x-x')^2+(y-y')^2} \tag{4.22}$$

with

$$x' = x+dx+dx' \tag{4.23}$$
$$y' = y+dy+dy' \tag{4.24}$$

where (x,y) is the pixel position at the image I_0, (dx,dy) is the motion vector of the forward warping, (dx',dy') is the motion vector of backward warping. The pixel (x',y') is the correspondence of the pixel (x,y) after forward and backward warping.



(a) Image I_0 (b) Image I_1

(c) Forward warping (d) Morphed image I_α (e) Backward warping

(f) Scale of (c) (g) Scale of (d) (h) Scale of (e)

Figure 4.6: Examples of optical flow based morphing. (a) Image I_0, (b) Image I_1, (c) Forward-warped image, (d) Morphed image I_α with $\alpha = 0.5$, (e) Backward-warped image. (f), (g) and (h) are the scaled hair part of (c), (d) and (e), respectively.

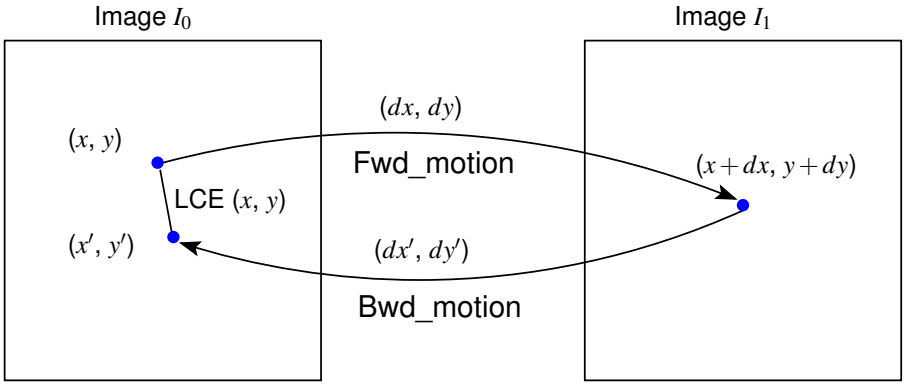


Figure 4.7: The local consistency error (LCE) measurement. Fwd_motion is the forward motion field calculation of the optical flow from image I_0 to image I_1, Bwd_motion is the backward motion field calculation from image I_1 to image I_0.

Fig. 4.8 shows the values of local consistency error for the image I_0 and I_1 of Fig. 4.6. The mean and standard deviation of the LCE is 6.69 ± 4.64 for image I_0 and 6.82 ± 4.65 for image I_1. In this example, the large consistency errors are brought mostly by the hairs, since the hair has a homogeneous texture and the light on the hair changes while the head is moving.

The difference between the image and the warped image is used to measure the motion vectors. The difference is qualified by PSNR (Peak Signal to Noise Ratio):

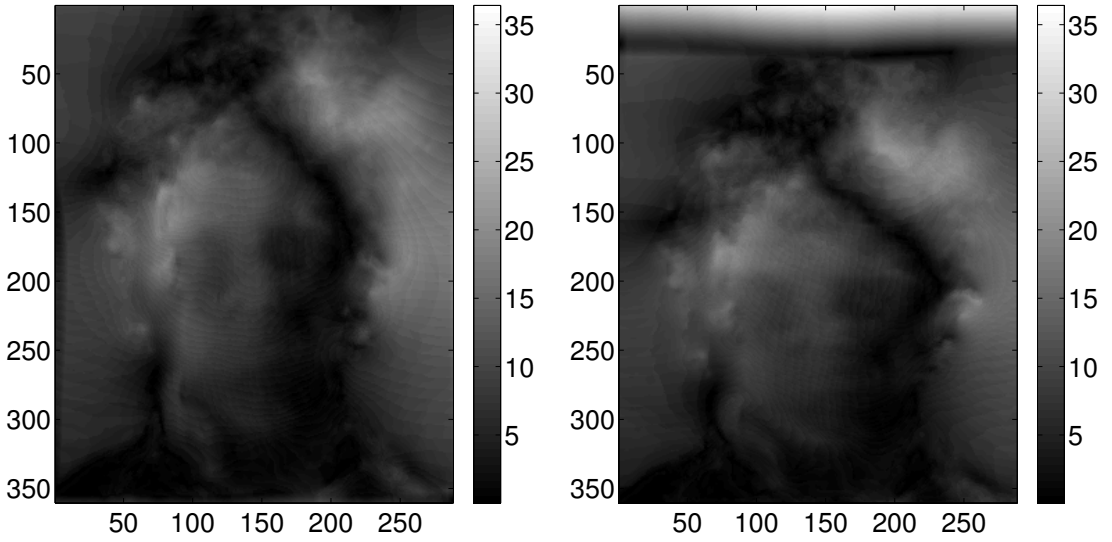
$$PSNR = 10 \cdot log_{10} \frac{255^2}{MSE} \tag{4.25}$$

where MSE (mean square error) is the average of the square differences of each pixel between the two images.

Fig. 4.9 shows the difference of the image and the warped image. Given in Fig. 4.9, (b) is the warped image from image (e) I_1 with $\alpha = 0$ by using the backward motion vector, (c) is the logarithmic difference between image (a) and (b). (d) is the warped image from image (a) I_0 with $\alpha = 1$ by using the forward motion vector, (f) is the logarithmic difference between image (d) and (e). The PSNR of image (c) and (f) is $29.29dB$ and $28.66dB$, respectively. From image (c) and (f), we can see that hair has the large difference between the original and warped images (white color means large value, black means small value).

4.3 Analysis of Head Motion Dynamics

Using the optical flow based morphing technique, the images between two head motion patterns can be interpolated without introducing noticeable artifacts. However, the speed



(a) LCE of Image I_0 (b) LCE of Image I_1

Figure 4.8: The local consistency error (LCE) of image I_0 and I_1. The mean and standard deviation of LCE for image I_0 and I_1 is 6.69 ± 4.64 and 6.82 ± 4.65 respectively.

and acceleration of the head movements are not smooth and make animations look unrealistic. In this section, we will analyze the dynamics of head motion at the transition from one head motion pattern to another.

Fig. 4.10 shows an example of the pose parameters of concatenated head motion patterns. From Fig. 4.10, we can see the jumps from one pattern to another in the pose parameter space, especially for rotation Ry, R_z and translation Tx, Ty. The influence of T_z is very weak. R_x gives the impact on the head pose change when the head nods.

In order to eliminate jerky head motions and smooth the transition between the patterns, an interpolation approach is required. Linear interpolation and sine function interpolation are two basic methods.

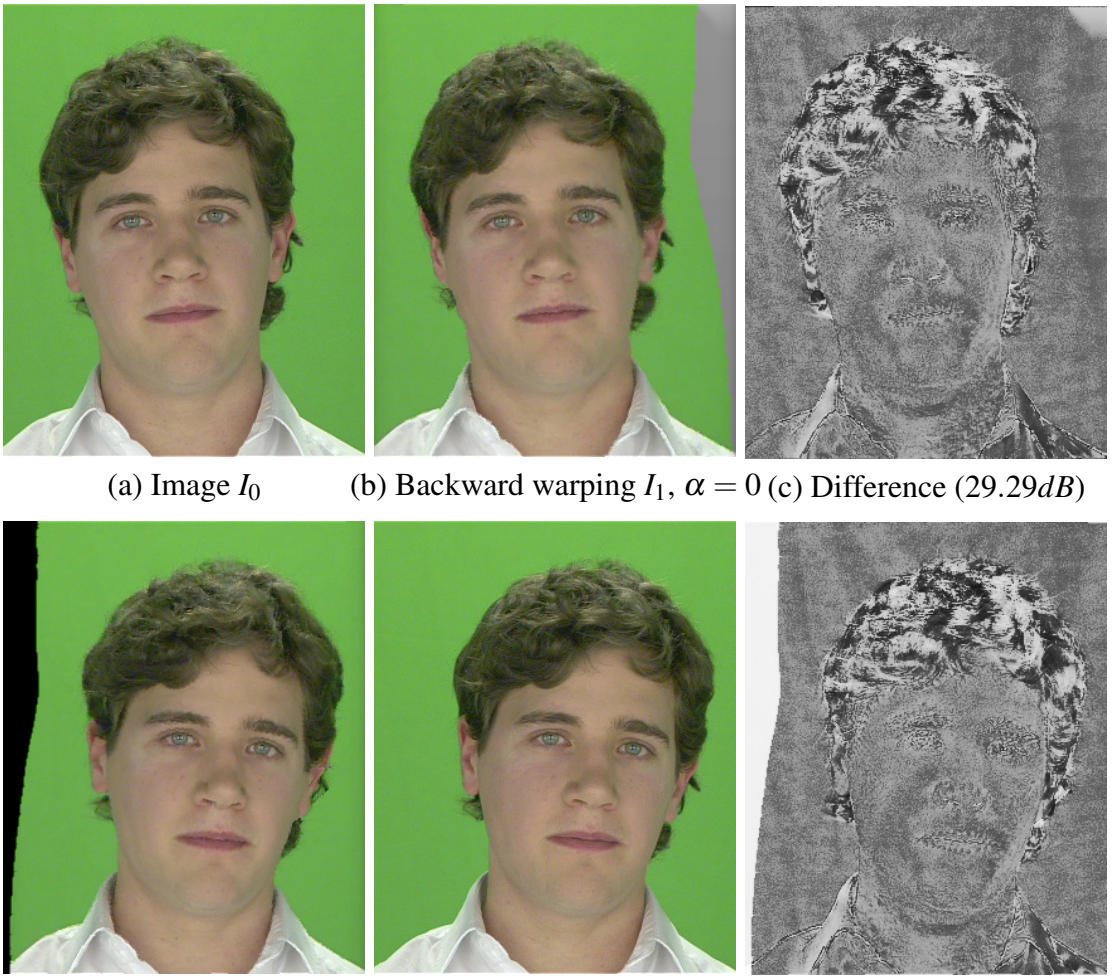
The linear interpolation, shown in Fig. 4.11(a), is calculated as follows:

$$\alpha = \frac{n}{N} \tag{4.26}$$

where n is the interpolated frame number in the interval of $[0,N]$, N is the total number of frames to be interpolated. α is the interpolation parameter for morphing in Eq. (4.15).

The sine function interpolation, shown in Fig. 4.11(b), is calculated:

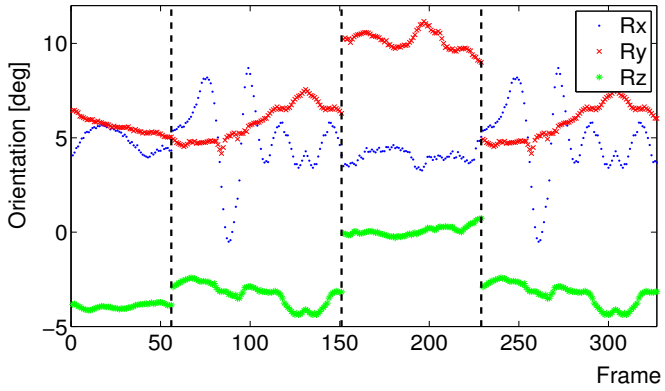
$$\alpha = \frac{1}{2}(1 + \sin(\frac{\pi}{N}n - \frac{\pi}{2})) \tag{4.27}$$



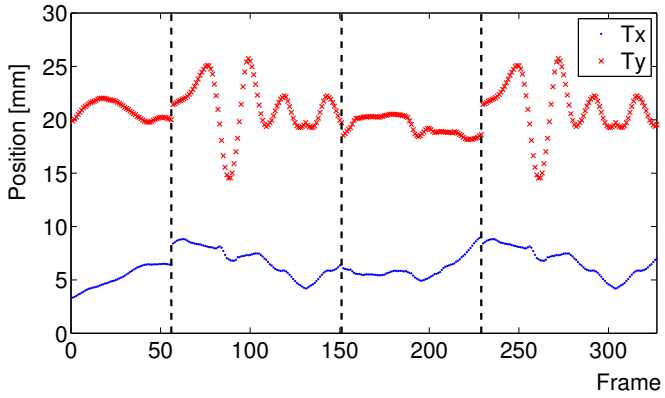
(a) Image I_0 (b) Backward warping I_1, $\alpha = 0$ (c) Difference ($29.29dB$)

(d) Forward warping I_0, $\alpha = 1$ (e) Image I_1 (f) Difference ($28.66dB$)

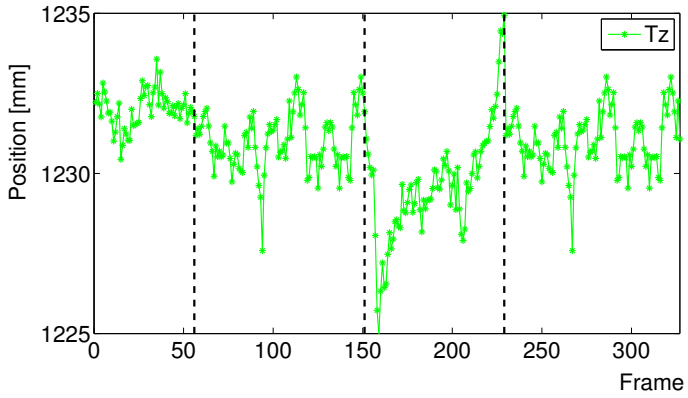
Figure 4.9: The difference between the image and the warped image. (a) Image I_0, (e) Image I_1, (b) Backward warped image from I_1 with $\alpha = 0$, (c) Logarithmic difference between image (a) and (b), $PSNR = 29.29dB$, (d) Forward warped image from I_0 with $\alpha = 1$, (f) Logarithmic difference between image (d) and (e), $PSNR = 28.66dB$.



(a) Rotation in x, y, z

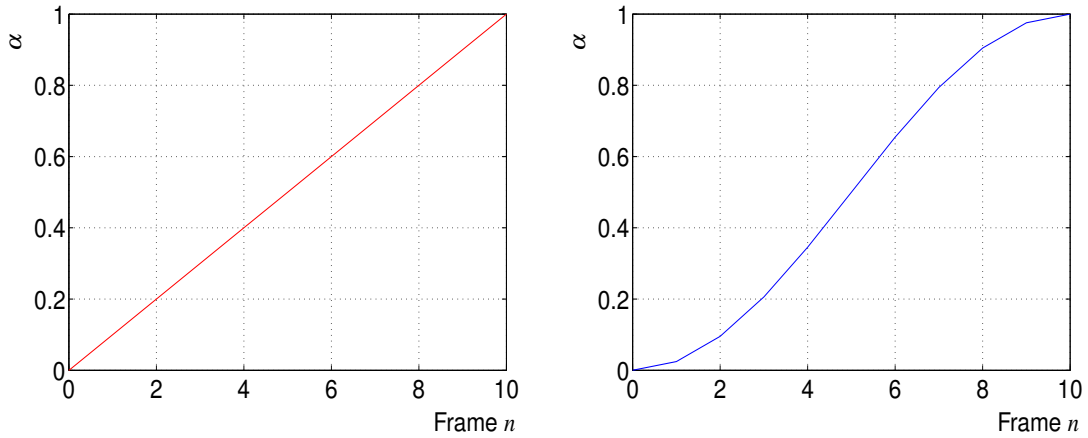


(b) Translation in x, y



(c) Translation in z

Figure 4.10: Pose parameters of a concatenation of 3 different head motion patterns, the second and fourth sentences use the same head motion pattern. (a) R_x, R_y, R_z, (b) T_x, T_y, (c) T_z.



(a) Linear interpolation (b) Sine function interpolation

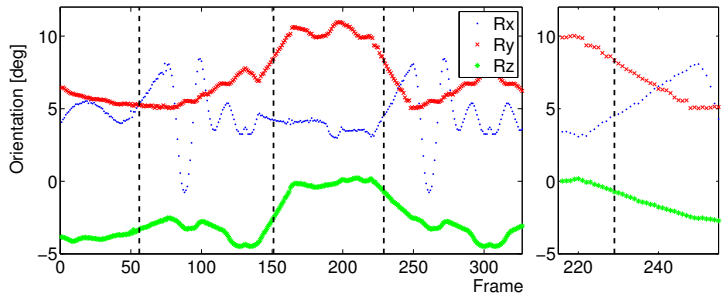
Figure 4.11: Linear and sine function interpolation.

We use the linear and the sine functions to morph the images between the motion patterns in the animation by Eq. (4.21). The head motion of the animated sequences is estimated by our model-based approach.

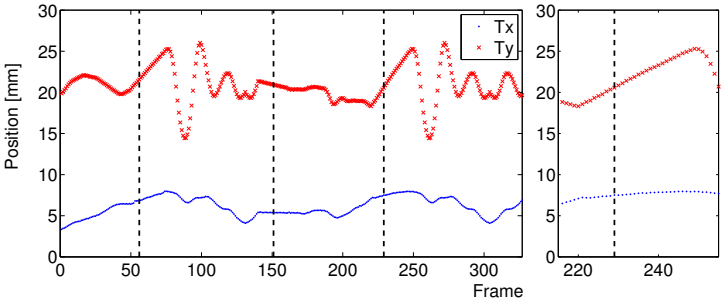
The transitions between head motion patterns of the animation in Fig. 4.10 are interpolated by linear interpolation and sine interpolation, respectively. In order to verify the smooth transition between patterns, the pose parameters of the interpolated sequences are estimated by our model-based approach, which are shown in Fig. 4.12. Fig. 4.12 (a) and (b) show the linear interpolation of transitions between head motion patterns, and Fig. 4.12 (c) and (d) show the sine interpolation of transitions between these head motion patterns. As an example, the transition of translation T_y between head motion patterns in the zoomed image of Fig. 4.12 (d) is more soft than the one in Fig. 4.12 (b).

The speed and acceleration of pose parameters for the animation of Fig. 4.10 are computed by Eq. (4.2)~Eq. (4.13). As an example, the speed of R_x and T_y in Fig. 4.13 are derived from the motion parameters of Fig. 4.10. We can see that the speed at the transition between patterns has an impulse-like shape in Fig. 4.13. On the contrary, the speed of R_x and T_y in Fig. 4.12 (c) and (d), which are interpolated by sine function, become continuous as shown in Fig. 4.14. Comparing Fig. 4.13 with Fig. 4.14, we can see that the pose parameters are estimated differently. The reason for this is that the pose parameters of the recorded sequence (Fig. 4.13) are estimated through the whole database with texture updates of face model. However, the pose parameters of the animated sequence (Fig. 4.14) are estimated with only a few texture update.

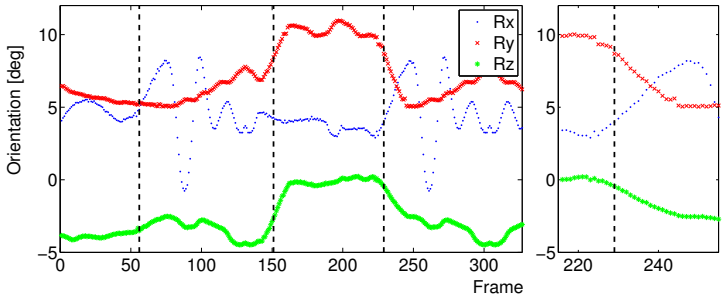
By analyzing the dynamics of the interpolation methods, we have found that both linear and sine interpolations can smooth the transitions between head motion patterns. The sine interpolation outperforms the linear interpolation in our informal subjective tests, since linear head motion is unrealistic. However, both methods have not considered the



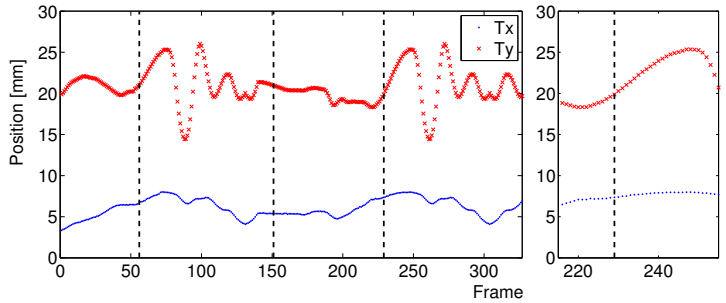
(a) Rotation trajectory of linear interpolation



(b) Translation trajectory of linear interpolation

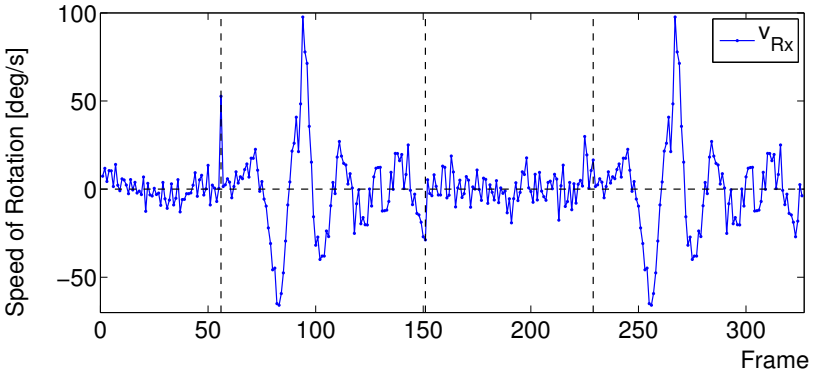


(c) Rotation trajectory of sine interpolation

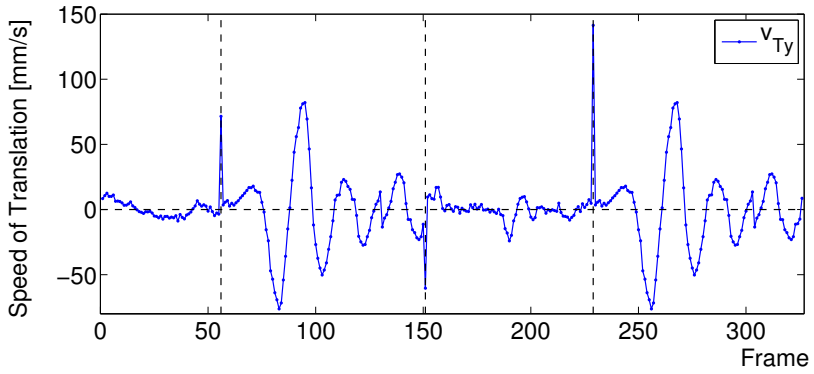


(d) Translation trajectory of sine interpolation

Figure 4.12: Linear and sine interpolation of pose parameters for the animated sequences. The right part is the zoomed view of the left plot.



(a) Speed of R_x



(b) Speed of T_y

Figure 4.13: Speed of R_x and T_y of the concatenated head motion patterns without interpolation. Transitions are at frame 63, frame 152, and frame 230.

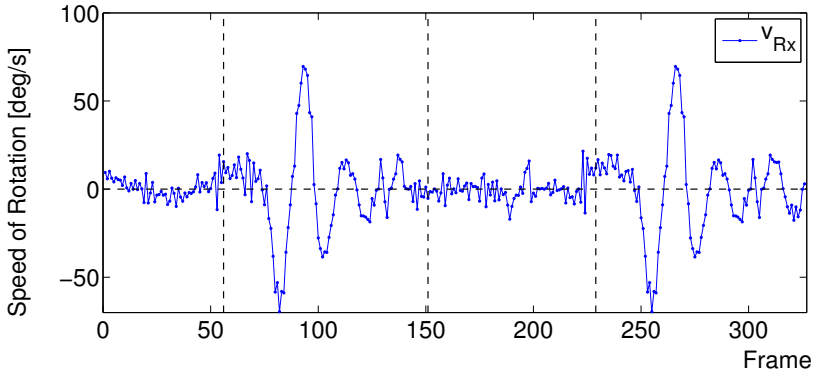
dynamics of the head motion before and after transitions.

Since the speed and direction of the curve are not considered in the linear interpolation and the sine function interpolation, a cubic Hermite interpolation is able to solve this problem [40].

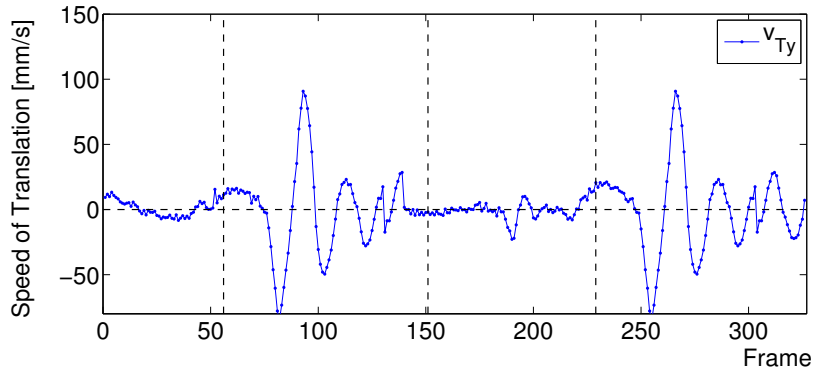
The cubic Hermite polynomial $p(x)$ has the interpolative properties

$$p(x_0) = y_0 \tag{4.28}$$
$$p(x_1) = y_1 \tag{4.29}$$
$$p'(x_0) = T_0 \tag{4.30}$$
$$p'(x_1) = T_1 \tag{4.31}$$

where points $P_0(x_0, y_0)$ and $P_1(x_1, y_1)$ are the start and end points, T_0 and T_1 are the tangents at the start and end points, respectively. Both the function values and their derivatives are known at the end points of the interval $[x_0, x_1]$.



(a) Speed of R_x



(b) Speed of T_y

Figure 4.14: Speed of R_x and T_y of the concatenated head motion patterns with sine interpolation. Morphed images are frames from 49 to 76, from 140 to 165, and from 220 to 249.

The cubic Hermite polynomial $p(x) = ax^3 + bx^2 + cx + d$ satisfies:

$$p(x_0) = ax_0^3 + bx_0^2 + cx_0 + d = y_0 \tag{4.32}$$
$$p(x_1) = ax_1^3 + bx_1^2 + cx_1 + d = y_1 \tag{4.33}$$
$$p'(x_0) = 3ax_0^2 + 2bx_0 + c = T_0 \tag{4.34}$$
$$p'(x_1) = 3ax_1^2 + 2bx_1 + c = T_1 \tag{4.35}$$

Now represent the linear equations in matrix format:

$$\begin{bmatrix} x_0^3 & x_0^2 & x_0 & 1 \\ x_1^3 & x_1^2 & x_1 & 1 \\ 3x_0^2 & 2x_0 & 1 & 0 \\ 3x_1^2 & 2x_1 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ T_0 \\ T_1 \end{bmatrix} \tag{4.36}$$

Now find the solution and set to Hermite polynomial:

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} x_0^3 & x_0^2 & x_0 & 1 \\ x_1^3 & x_1^2 & x_1 & 1 \\ 3x_0^2 & 2x_0 & 1 & 0 \\ 3x_1^2 & 2x_1 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} y_0 \\ y_1 \\ T_0 \\ T_1 \end{bmatrix} \tag{4.37}$$

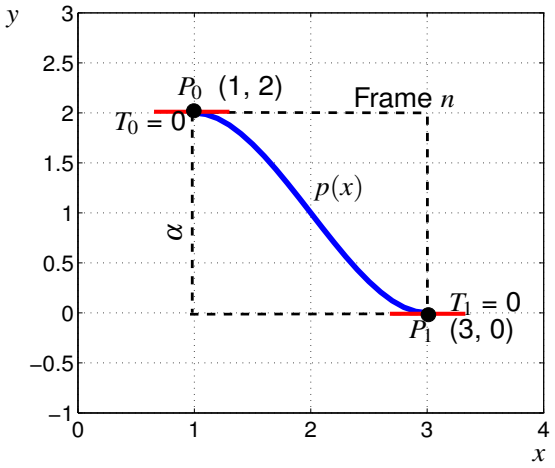


Figure 4.15: An example of cubic Hermite interpolation $p(x) = 0.5x^3 - 3x^2 + 4.5x$. α value is interpolated by the Hermite polynomial.

Fig. 4.15 shows an example of cubic Hermite polynomial interpolation, which smooths not only the transition, but also the tangents at the start and end points. In Fig. 4.15, start point P_0 maps to $\alpha = 0$, and end point P_1 maps to $\alpha = 1$. Other α values are interpolated by the cubic Hermite polynomial. In this example, the Hermite polynomial approximates sine function.

In order to calculate the parameter α, the pose parameter trajectories of the head motion patterns are interpolated by using the Hermite polynomial. α value is derived by normalizing the interpolated pose parameter to the interval of $[0,1]$, where the parameter of start point is mapped to 0, the parameter of end point is mapped to 1. Finally, the head motion patterns are morphed by using these α values. α value is calculated in this way:

$$\alpha = \frac{p(x) - y_0}{y_1 - y_0} \tag{4.38}$$

with

$$x = \frac{n}{N}(x_1 - x_0) + x_0 \quad ; \quad n \in [0, N] \tag{4.39}$$

where $p(x)$ is given by the Hermite polynomial, (x_0, y_0) is the start point and (x_1, y_1) is the end point.

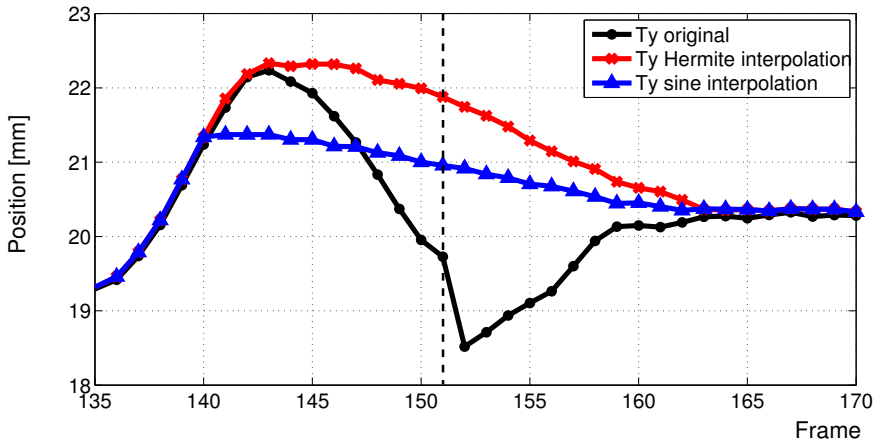
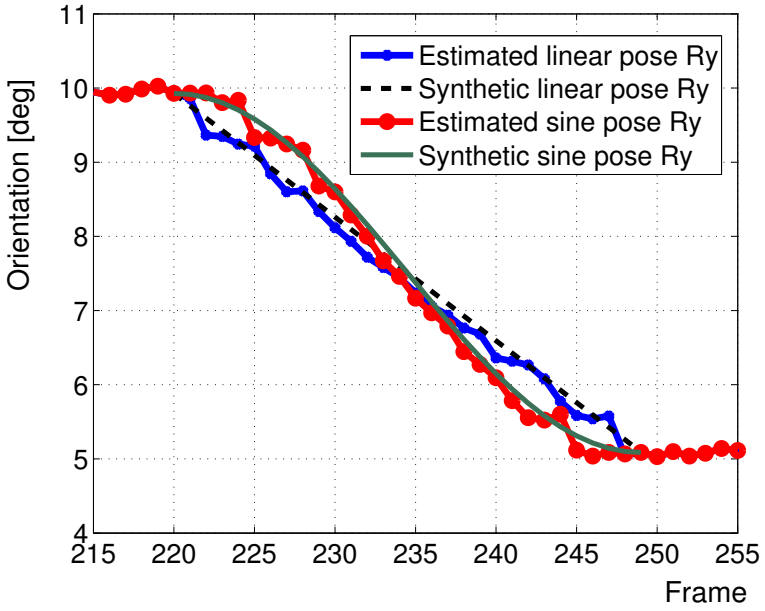


Figure 4.16: Comparison of Hermite and sine interpolations. The black curve is the pose parameters of the head motion patterns extracted from Fig. 4.10(b). The pose parameters of red and blue curves are estimated from the animations.

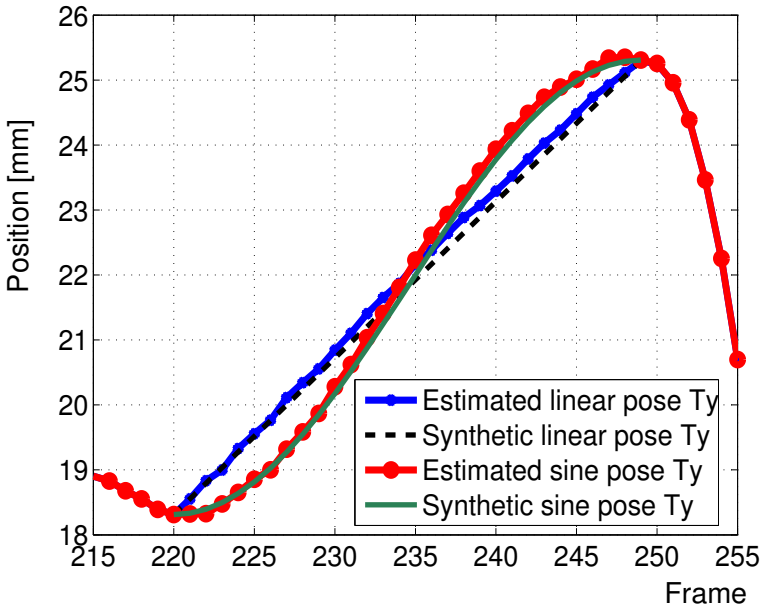
Fig. 4.16 compares the sine interpolation and Hermite interpolation. The black curve is original T_y trajectory of two head motion patterns cropped from Fig. 4.10(b). The blue curve is the trajectory of T_y, which is estimated from the sequence morphed by using the sine interpolation, and the red curve is the one estimated from the sequence morphed by using the Hermite interpolation. Given in Fig. 4.16, the transition by using Hermite interpolation is smoother than the one by using sine interpolation.

To evaluate the interpolation methods, we can compare the motion trajectories of morphed sequence with a synthetic one. Synthetic trajectories are ground truth. Fig. 4.17(a) shows the estimated and synthetic linear and sine pose of R_y. In this example, the sine function approximates the Hermite polynomial. The blue curve is interpolated by linear function, extracted from frame 215 to 255 in Fig. 4.12(a). The red one is interpolated by sine function, extracted from frame 215 to 255 in Fig. 4.12(c). The black dashed line is the synthetic linear pose and the green curve is the synthetic sine pose. Fig. 4.17(b) shows the estimated and synthetic linear and sine pose of T_y. From Fig. 4.17, we can find that the trajectories of estimated and synthetic translation T_y are matched better than the trajectories of rotation R_y. The reason is that the morphed images have blurred textures of hair (see Fig. 4.6), since the optical flow cannot find good correspondences for hair between two images. This influences the rotation estimation more than the translation estimation. The head motion in the morphed videos is smooth in our informal subjective tests, even though small mismatch of these rotation curves exist.

The expressive unit selection finds smooth transitions between head motion patterns, according to the concatenation cost measurement. The pose and motion parameters of the connecting images are similar, so that the α value for morphing is always in the interval of [0, 1].



(a) Rotation R_y



(b) Translation T_y

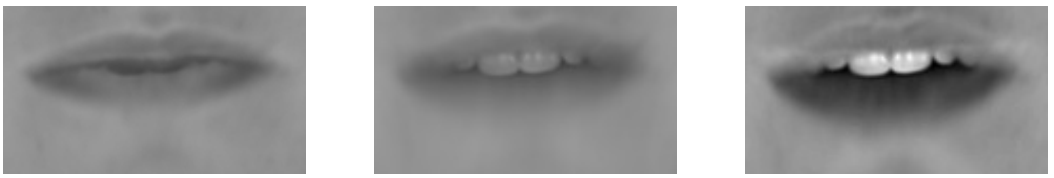
Figure 4.17: Estimated and synthetic linear and sine pose R_y and T_y. The images are morphed from frame 220 to 249 of Fig. 4.10.

5 Quality Assessment

A subjective test is the ultimate measure of quality. However, this is time consuming and expensive. An automatic objective assessment is required to accelerate the development and also increase the efficiency of subjective tests. Usually, the feedback of the viewers can be explained by measurable characteristics of animations. However, an objective measure will never replace subjective tests. For any subjective tests, it is important to carefully separate the different factors influencing the perceived quality of animation. Therefore, the synthesis of facial expression and the synthesis of head motion are evaluated separately in this chapter. Part of the evaluation results is presented in [62].

5.1 Objective Evaluation

Animations are driven by the phonemes and durations of the real audio, so that the comparison between real and synthesized sequences is possible. Objective evaluation is performed by directly comparing the visual parameters of the synthesized sequence with the real one. The visual parameters are the geometric and appearance features of the mouth image sequence. Mouth height is the geometric parameter and the first significant PCA component is the appearance parameter for objective evaluation. The first PCA component of the mouth image database represents the mouth height. Fig. 5.1 shows the texture variation of the mouth images in the first PCA component. However, the mouth height cannot replace the first PCA component, because different mouth textures could have the same mouth height.



(a) -3 std. dev. (b) Mean mouth (c) +3 std. dev.

Figure 5.1: Texture deformation by varying the first PCA component. (b) is the mean mouth texture of the database, (a) is the deformation of (b) with -3 standard deviation in the first PCA component, (c) is the deformation of (b) with 3 standard deviation in the first PCA component.

Experiments indicate that human viewers are very sensitive to closures, and getting the closures at the right time may be the most important criterion for providing the impression

that lips and sound are synchronized. Closures are easy to identify visually, simply by finding whether the inner lips are touched. The precise shapes of the openings are less important for the perceived quality of the articulation [9]. The match rate of closures $R_{closures}$ between animated and real sequence is used to measure lip synchronization. The match rate of closures is calculated as:

$$R_{closures} = \frac{Number\ of\ closures\ of\ animated\ sequence}{Number\ of\ closures\ of\ real\ sequence} \tag{5.1}$$

Another criterion to measure lip synchronization is the match rate of turning points (TP) between real and animated sequence. The turning point is defined as a position where the mouth changes the movement from opening to closing, or vice versa. In other words, the turning points are peaks and troughs in the curve. These are typically points of high acceleration of the jaw and of strong muscle action. The precise placement of a single turning point does not seem to be important. If several turning points in one animated sequence are different from those in the corresponding recorded sequence, viewers get the impression of poor lip synchronization. The match rate R_{TP} of valid turning points is calculated as follows:

$$R_{TP} = \frac{Number\ of\ valid\ turning\ points\ of\ animated\ sequence}{Total\ number\ of\ turning\ points\ of\ real\ sequence} \tag{5.2}$$

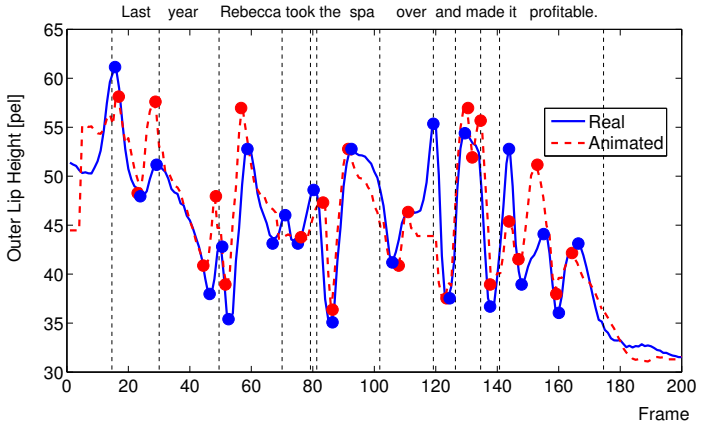
Valid turning points are the turning points of the animated sequence, which have the same timing as the turning points of real sequence have. The invalid turning points of the animated sequence could be the missing turning points and the additional turning points, so that the mismatch rate F_{TP} of the invalid turning points is computed in the following way:

$$F_{TP} = \frac{Number\ of\ invalid\ turning\ points\ of\ animated\ sequence}{Total\ number\ of\ turning\ points\ of\ real\ sequence} \tag{5.3}$$

Since the measured trajectories of lip height are always noisy, several unexpected peaks and troughs exist. In order to determine the correct turning points, we measure the difference between lip height amplitude of succeeding peak and trough. If the difference is smaller than 2 pixels, the peak and the trough are not regarded as real turning points.

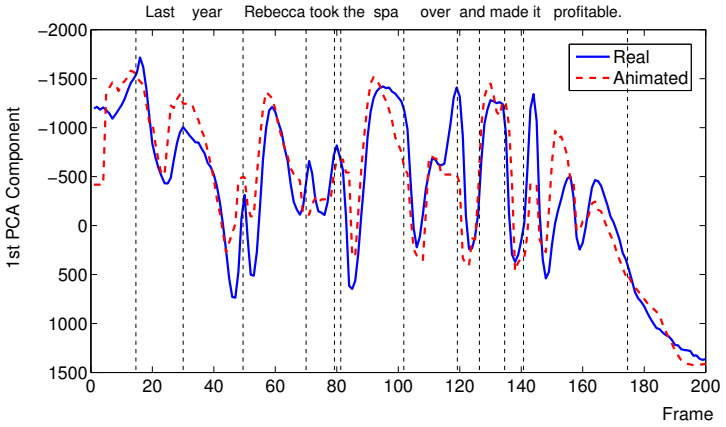
The finding of the Advanced Television Systems Committee(ATSC) states that lip sync errors become noticeable if the audio is early by more than 15*ms* or late by more than 45*ms* [80]. Hence, if the offset of the corresponding turning points between the real and animated sequences are smaller than 2 frames, the two turning points are synchronized subjectively.

We select 9 recorded sentences with a smile from the expressive database as ground truth. The talking head smiles at the end of each sentence before the last syllable. These 9 sequences are left out of the database. For each real sequence, an animated sequence is generated by the expressive unit selection.



(a) Outer lip height trajectory

(b) Inner lip height trajectory



(c) PCA trajectory

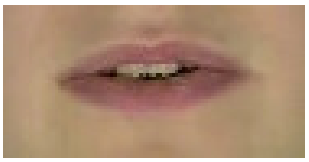
Figure 5.2: Trajectories of the animated sequence and the recorded sequence of a sentence. (a) shows the outer lip height trajectories and the marked points are turning points, (b) shows the inner lip height trajectories, (c) shows the 1^{st} PCA component trajectories. The sentence is "Last year Rebecca took the spa over and made it profitable.". The dashed lines are word boundaries.

The trajectories of geometric and appearance parameters for the animated and the real sequences of a sentence are shown in Fig. 5.2. Fig. 5.2 (a) shows the trajectories of the outer lip height of the real (blue curve) and the animated (red curve) sequences. Fig. 5.2 (b) shows the trajectories of the inner lip height of the real (blue curve) and the animated (red curve) sequences. Fig. 5.2 (c) shows the trajectories of the first PCA component of the real (blue curve) and the animated (red curve) sequences.

The closures can be measured from the inner lip height in Fig.5.2 (b). If the curve reaches the bottom, the mouth is closed. Both the animated and the real sequences have identical mouth closures in terms of timing. The match rate of closures $R_{closures}$ is 100%. We have measured the maximal frame offset of closures between the animated and the real sequences are 2 frames (audio delayed video 40ms at 50Hz). Hence, the audio and video of animations are synchronized.

In Fig. 5.2 (a), there are 12 peaks and 11 troughs in the speaking part in the real sequence. 10 peaks of 12 peaks in the animated sequence appear at the right time as the real ones. Two peaks, at frame 70 and frame 120, are missing and one peak at frame 134 is not expected. There are 11 troughs in the animated sequence, of which one trough at frame 131 is additional and one trough is missing at frame 67. The valid turning points have a maximum of 1 frame offset between the real and the animated sequences, which is in the tolerance of lip synchronization. The match rate R_{TP} is 87% and the mismatch rate F_{TP} is 22%. The amplitudes of these peaks and troughs are less important in the perception of coarticulation.

(a) Mouth cropped from real sequence



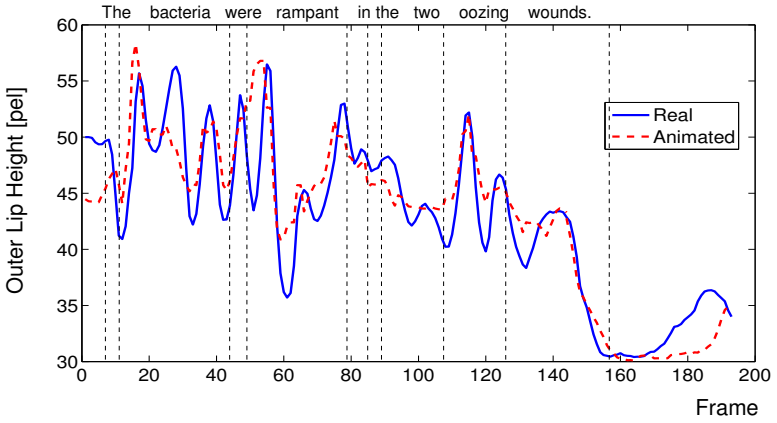
(b) Mouth cropped from animated sequence

Figure 5.3: The mouth images (size 93×47) of frame 120 in the real and the animated sequences in Fig. 5.2.

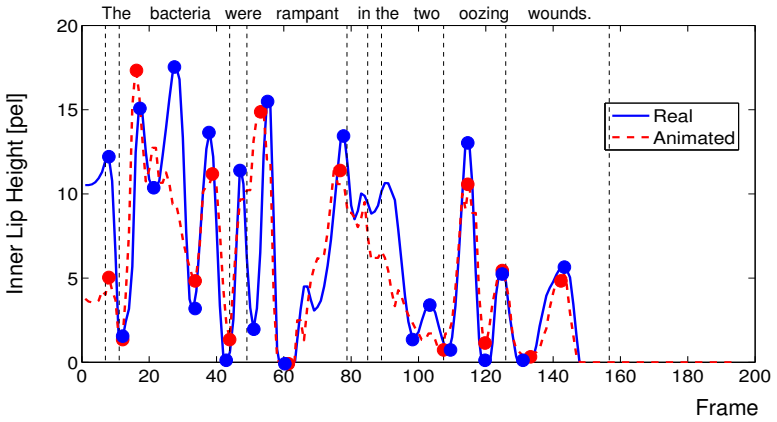
One noticeable position in Fig. 5.2 (a) is around frame 120. The difference between outer lip height of the real and the animated sequences at frame 120 is 11 pixels. The mouths of frame 120 are cropped and shown in Fig. 5.3. The viseme “ax” in the word “and”, shown in Fig. 5.3 (a), is synthesized by the first viseme “ax” in the word “Nebraska”, shown in Fig. 5.3 (b), since the unit selection balances lip synchronization and smoothness. The phonemes of “and” are “ax n” and the phonemes of “Nebraska” are “n ax b r ae s k ax”.

In Fig. 5.2 (c), the curves of PCA parameters run in the way like the outer lip height, except that the sign of mouth height is opposite to the sign of PCA space.

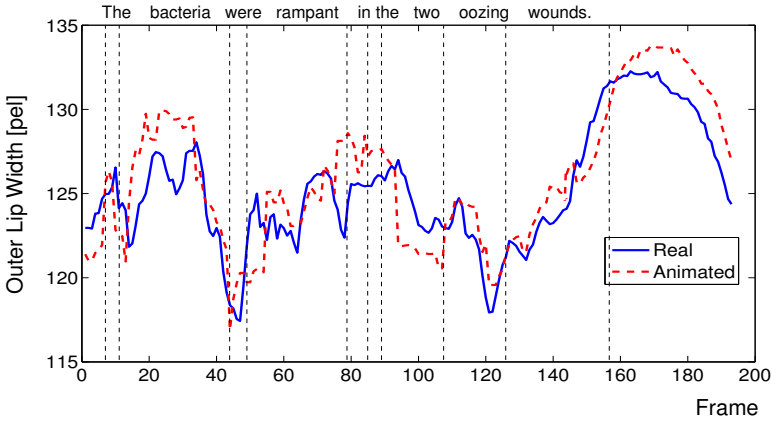
The trajectories of the animated and the real sequences of another sentence are shown



(a) Outer lip height trajectory



(b) Inner lip height trajectory



(c) Outer lip width trajectory

Figure 5.4: Trajectories of the animated sequence and the recorded sequence of a sentence. (a) shows the outer lip height trajectories, (b) shows the inner lip height trajectories and the marked points are turning points, (c) shows the outer lip width trajectories. The sentence is "The bacteria were rampant in the two oozing wounds." The dashed lines are word boundaries.

in Fig. 5.4. Fig. 5.4 (a) shows the trajectories of the outer lip height of the real (blue curve) and the animated (red curve) sequences. Fig. 5.4 (b) shows the trajectories of the inner lip height of the real (blue curve) and the animated (red curve) sequences. Fig. 5.4 (c) shows the trajectories of the outer lip width of the real (blue curve) and the animated (red curve) sequences. When smiling, the mouth width becomes larger and the mouth height becomes smaller after frame 140. Furthermore, the transition from neutral to smile in the animated sequence is also as smooth as the real one. The smile starts before the last syllable.

(a) Mouth cropped from real sequence (b) Mouth cropped from animated sequence

Figure 5.5: The mouth images (size 93×47) of frame 27 in the real and the animated sequences in Fig. 5.4.

One noticeable position in Fig. 5.4 (a) is around frame 30. The difference between outer lip height of real and animated sequences at frame 27 is 7 pixels. The mouths of frame 27 are cropped and shown in Fig. 5.5. Even though the opening of the synthesized mouth is smaller than the opening of the real mouth, the quality of the animation is as good as the quality of the real sequence subjectively. Since a phoneme could be spoken with different openings, the unit selection selects best mouths, which fulfill the two subjective requirements of humans in terms of synchronization and smoothness. The phonemes of "bacteria" is "b ae k t ia r ia" and the phonemes of "Australia" are "oh s t r ey ia". The first viseme "ia" in the word "bacteria", shown in Fig. 5.5 (a), are synthesized by the viseme "ia" in the word "Australia", shown in Fig. 5.5 (b).

(a) Mouth cropped from real sequence (b) Mouth cropped from animated sequence

Figure 5.6: The mouth images (size 93×47) of frame 60 in the real and the animated sequences in Fig. 5.4.

The mouth has different outer lip height when it is closed. In order to measure closure accurately, the inner lip height is used. As an example, the outer lip of frame 60 has different heights in the real and animated sequences of Fig.5.4(a), while the inner lip of

frame 60 has closed at this frame in both sequences of Fig. 5.4(b). The mouths are cropped from this frame in the real and animated sequences and are shown in Fig. 5.6. The width and the height of the outer lip, the inner lip height of mouth (a) of Fig. 5.6 is 123 pel, 36 pel and 0 pel, respectively. The width and the height of the outer lip, the inner lip height of mouth (b) of Fig. 5.6 is 124 pel, 41 pel and 0 pel, respectively. The match rate of mouth closures between the real and the animated sequence is also 100%. Another special case is the borderline closure as shown in Fig. 5.7. The mouth in the real sequence is just to be closed at frame 42, while the mouth in the animated sequence is almost closed. The distance of the inner lip height between the animated and the real sequence is less than 1 pixel.

(a) Mouth cropped from real sequence (b) Mouth cropped from animated sequence

Figure 5.7: The mouth images (size 93×47) of frame 42 in the real and the animated sequences in Fig. 5.4.

In Fig. 5.4 (a), the outer lip height trajectory of the animated sequence looks very noisy. In this example, the inner lip heights are better matched than the outer lip heights, so that we measure the turning points from Fig. 5.4 (b). There are 11 peaks and 10 troughs in the real sequence. The 8 peaks and 7 troughs in the animated sequence are matched with the real ones, even though 6 turning points are missing. The match rate R_{TP} is 71% and the mismatch rate F_{TP} of invalid turning points is 29%. The match rate of Fig. 5.2 is higher than the match rate of Fig. 5.4, although they have good closures.

The closures and the turning points of the 9 pairs of real and animated sequences are measured. The average match rate of the closures is 100%. The average match rate of the turning points is 81%, while the average mismatch rate of the turning points is 23%. The 9 pairs of real and animated sequences are further used for subjective tests.

5.2 Subjective Evaluation

The standard approach [81] to assess naturalness of a talking head is to conduct subjective tests where viewers score animations on a scale from 1 to 5. The distribution of actual subjective data is approximated by a symmetrical logistic function.

The analysis of the MOS scores is summarized in two steps [81]. The first step of the

analysis of the results is the calculation of the mean score, $\bar{u}$, for each video.

$$\bar{u}_k = \frac{1}{N}\sum_{i=1}^{N} u_{i,k} \tag{5.4}$$

where $u_{i,k}$ is the score of observer i for video k, N is the number of observers.

The second step is to calculate the confidence interval. It is proposed to use the 95% confidence interval which is given by:

$$[\bar{u}_k - \delta_k, \bar{u}_k + \delta_k] \tag{5.5}$$

where

$$\delta_k = 1.96\frac{S_k}{\sqrt{N}} \tag{5.6}$$

$$S_k = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_k - u_{i,k})^2}{N-1}} \tag{5.7}$$

In the subjective tests of synthesis of facial expression and synthesis of head motion, the standard mean opinion score approach is used. This evaluation method is also used in the first visual speech synthesis challenge (LIPS2008) [4].

Thirty subjects participated in the experiments. Almost all were students and staff in Leibniz Universität Hannover. Subject characteristics are summarized in Table 5.1.

Table 5.1: Subjects.

by gender	by age	by language	by major
Female: 10	19-22: 9	English good: 20	EE: 10
Male: 20	23-26: 9	English normal: 10	CS: 10
	27-30: 5		other: 10
	31-35: 3		
	over 36: 4		

5.2.1 Subjective Evaluation of Facial Expression Synthesis

We evaluate the smiling talking head subjectively in terms of naturalness. The viewers should give scores to the synthesized videos and the corresponding real ones.

Data Collection

Nine recorded sentences are collected as testing sequences, which are not included in the expressive database. The audios and the texts of these 9 sentences are aligned and the phonemes and durations are obtained, which are used to drive the smiling talking head. Each animation is generated by using the recorded video as the background video, so that the comparison is focusing on the mouth part. Finally, 9 video pairs, including 9 recorded videos and 9 synthesized videos, are collected for the subjective test.

Figure 5.8: A snapshot of cropped mouth images from recorded smile sequence (row A1 and A2) and animated smile sequence (row B1 and B2). The text of the sentence is "Upton saw a disaster coming.". The animated sequence uses the recorded sequence as background video.

The texts of the 9 sentences are listed as follows:

1. Upton saw a disaster coming.
2. The smell of ether overpowered the smell of coffee in the room.
3. I want to divide the talcum powder into two piles.

4. Last year Rebecca took the spa over and made it profitable.
5. Don't scoff at the strong work ethic Mister Hoffa had all his life.
6. The military junta had trouble within its ranks.
7. Mister Dewey played bassoon in the orchestra.
8. The bacteria were rampant in the two oozing wounds.
9. You need a periodic check of your intrauterine device.

A snapshot of cropped mouth images from the recorded sequence and the animated sequence is shown in Fig. 5.8.

Naturalness Test

Even though only the mouth is replaced by the new mouth, the quality of the whole face is influenced by color and pose. Therefore, a naturalness test is able to evaluate the overall quality of talking heads.

Before running the subjective test, the basic technologies of facial animations are introduced to the viewers. In order to make the viewers judge the videos with reasonable scores, several videos with different qualities (from bad to good) are shown and explained. Thereafter, the subjective test starts.

Table 5.2: Naturalness test for facial expression synthesis. Average MOS scores (5=Excellent, 1=Bad) and the confidence intervals ($\bar{u}_k \pm \delta_k$) are analyzed.

Pair No.	Recorded video	Animated video
1	3.7 ± 0.4	3.5 ± 0.5
2	4.0 ± 0.3	3.9 ± 0.4
3	4.0 ± 0.3	3.9 ± 0.4
4	3.8 ± 0.3	3.7 ± 0.4
5	4.6 ± 0.3	4.3 ± 0.4
6	4.5 ± 0.3	4.1 ± 0.5
7	4.1 ± 0.3	3.9 ± 0.4
8	4.1 ± 0.4	4.0 ± 0.6
9	4.2 ± 0.3	3.8 ± 0.6
avg. MOS	4.1	3.9

9 video pairs of random order are shown to each viewer only once. Each video pair is played randomly, so that the viewer does not know which one is played first. After showing a video pair, the viewer should give scores immediately. The questionnaire for the subjective test is attached in the appendix A. The MOS scores and the confidence

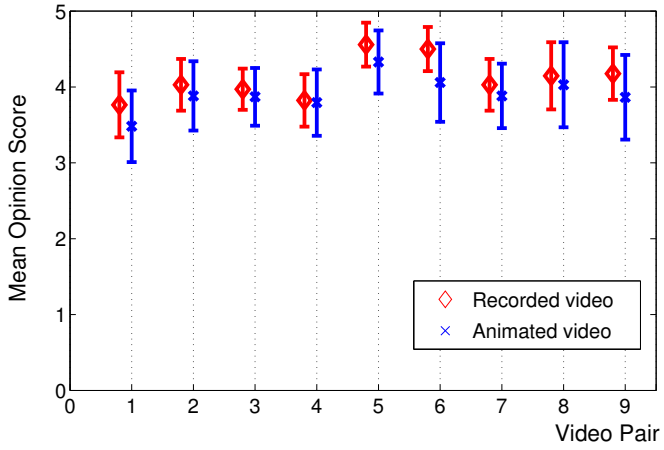


Figure 5.9: Results of the subjective test for facial expression synthesis.

intervals are analyzed and summarized in Table 5.2, which is plotted in Fig. 5.9. The confidence intervals of the MOS scores overlap. The overall MOS score of the animated videos is 3.9 and the overall MOS score of the recorded videos is 4.1.

In addition, we have asked the viewers, which part of face decreases the quality, such as eye, mouth, or unknown. The answers are insignificant. The viewers indicate that they cannot distinguish the animated smiles from real ones.

Supplemental materials of facial expression synthesis can be downloaded from http://www.tnt.uni-hannover.de/project/facialanimation/demo/emotion.

5.2.2 Subjective Evaluation of Head Motion Synthesis

The subjective test is to evaluate the naturalness of head motions of background sequences, which are generated by concatenation of head motion patterns.

Data Collection

In order to evaluate head motion synthesis, talking heads with or without head motion synthesis should be compared. Talking heads without head motion synthesis require only a short head motion pattern as the background sequence. If the target animation is longer than the head motion pattern, the head motion pattern will be replayed inversely. Corresponding to the talking head with repeated head motion, an animation of talking head with flexible head motion is generated by selecting a different head motion pattern for each sentence as the background sequence. The best transition positions between these patterns are found by the Viterbi search. The transitions between these patterns are smoothed by the optical flow based morphing technique.

The text of each video consists of at least two sentences from BBC News Website. The texts with head motion tags are listed as follows:

1. <prosody motion="medium">The US files two new cases against China at the World Trade Organization, amid rising tensions over China's currency.</prosody> <prosody motion="medium">India's central bank raises interest rates by more than expected as it continues to battle high inflation.</prosody>

2. <prosody motion="medium">Are you ok?</prosody>
<prosody motion="nod">I hope so.</prosody>
<prosody motion="strong">Are you happy?</prosody>
<prosody motion="nod">I hope so.</prosody>

3. <prosody motion="medium">In Northern Ireland, lessons in personal finance are compulsory for all children and there are calls for it to be extended across the UK.</prosody>
<prosody motion="strong">Bank of England governor has blamed financial firms and policy-makers for the economic crisis, admitting: "We let it slip."</prosody>

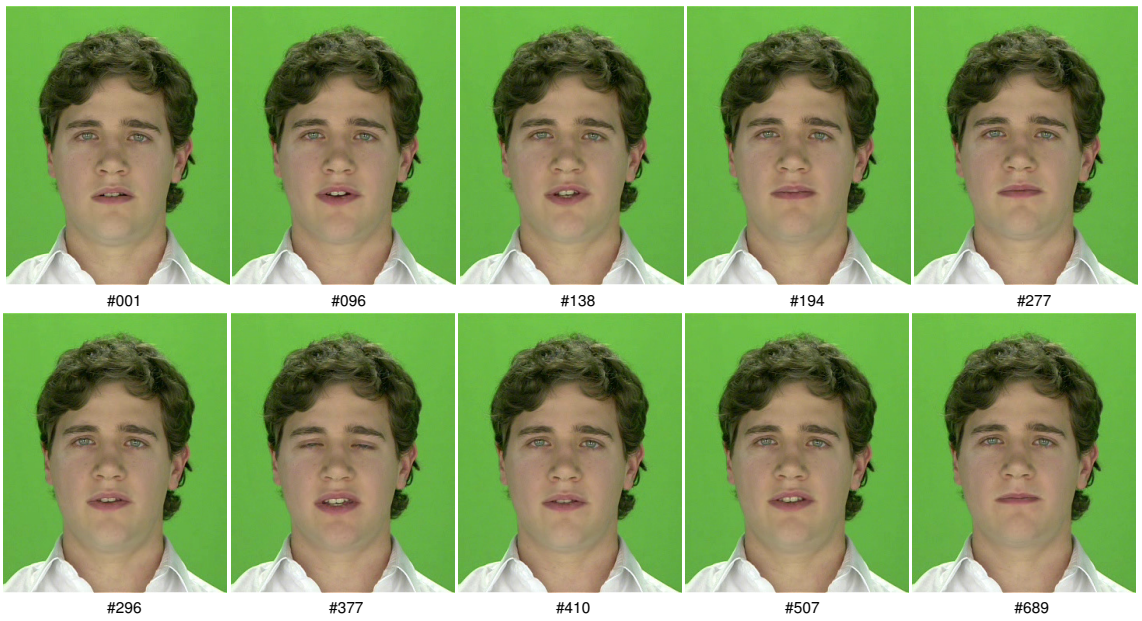
4. <prosody motion="medium">Oil prices fall for the third day in a row as a US government report suggested that demand remains sluggish.</prosody>
<prosody motion="medium">Volunteers can now apply to work at the London 2012 Olympic Games.</prosody>

5. <prosody motion="medium">Falling in love comes at the cost of losing two close friends, a study says.</prosody>
<prosody motion="strong">If you would like the latest BBC News World video stories,</prosody> <prosody motion="strong">please visit the video and audio section of the BBC News Website.</prosody>

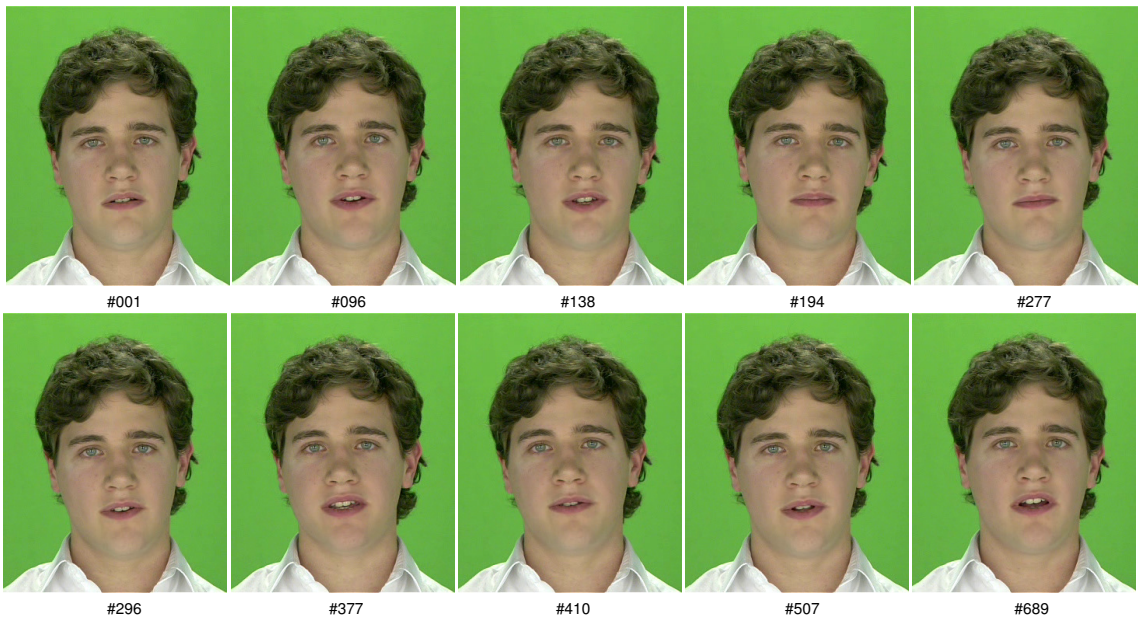
For each text, a pair of videos is generated. One is a talking head with repeated head motion, the other is a talking head with flexible head motion. Therefore, 5 video pairs are collected for the subjective test.

A snapshot of images extracted from the animation with repeated head motion and the animation with flexible head motion is shown in Fig. 5.10.

The pose parameters of the repeated sequence and the flexible sequence from Fig. 5.10 are plotted in Fig. 5.11 and in Fig. 5.12, respectively. There are no transitions in Fig. 5.11. The background sequence (duration about 5 seconds) is replayed in the way of forwards and backwards. In Fig. 5.12, three head motion patterns are concatenated. Two transitions are interpolated by Hermite functions around frame 254 and frame 474, respectively. The large transition speeds of R_y, R_z and T_x are as big as the real ones. For example, the transition of R_y at frame 254 is similar to the natural head motion from frame 330 to frame 350.

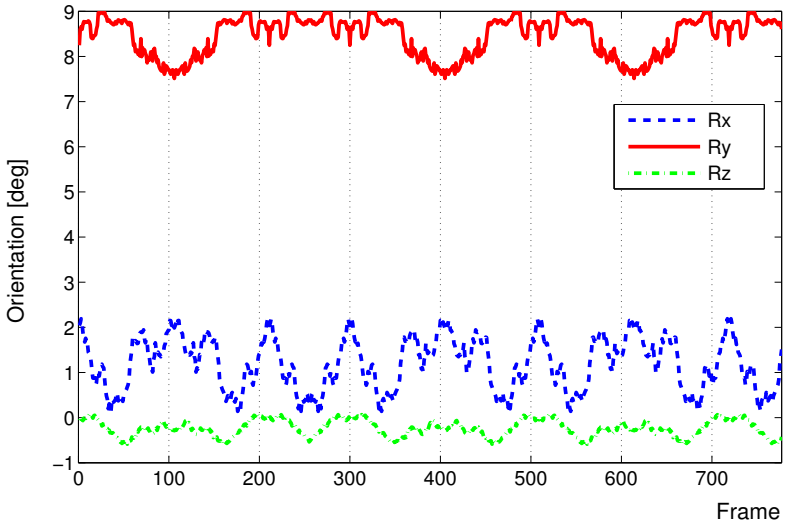


(a) Animation with repeated head motion

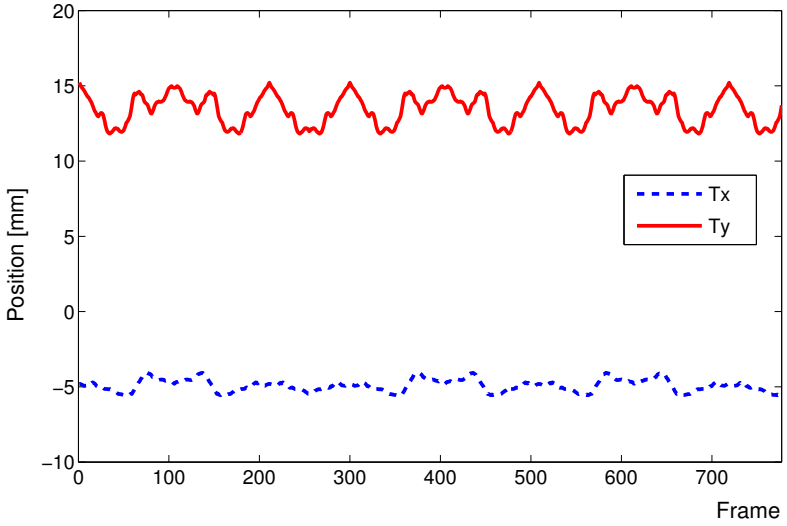


(b) Animation with flexible head motion

Figure 5.10: A snapshot of images extracted from the animation with repeated head motion and the animation with flexible head motion. The utterance is "Falling in love comes at the cost of two close friends, a study says. If you would like the latest BBC News World video stories, please visit the video and audio section of the BBC News Website." Both sequences have the same mouth animation, but with different background video.

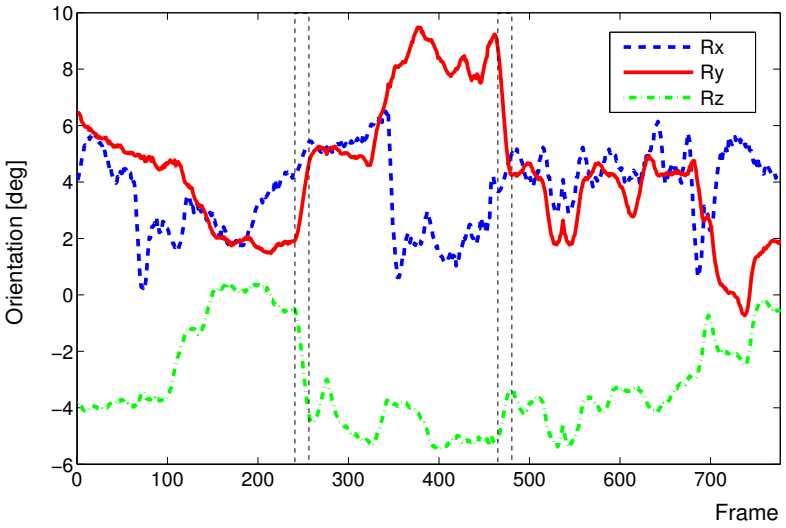


(a) R_x, R_y, R_z of repeated head motion

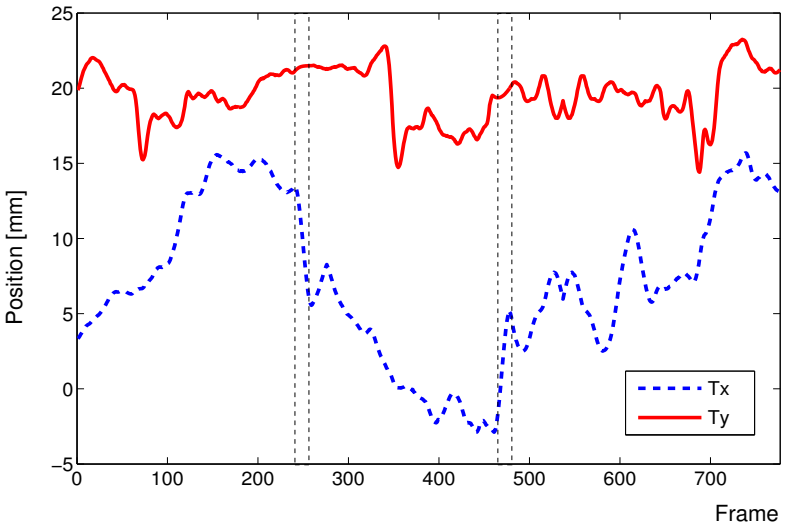


(b) T_x, T_y of repeated head motion

Figure 5.11: Pose parameters of the talking head with repeated head motion. The short head motion pattern has 255 frames.



(a) R_x, R_y, R_z of flexible head motion



(b) T_x, T_y of flexible head motion

Figure 5.12: Pose parameters of the talking head with flexible head motion. Three head motion patterns are concatenated. Two positions are interpolated by Hermite functions. The dashed lines, from frame 244 to 264 and from frame 464 to 484, are two transitions between head motion patterns.

Naturalness Test

Naturalness test of head motion synthesis is to evaluate head motion smoothness. The 5 video pairs are selected in a random order and two videos of a pair are also shown each viewer randomly. Viewers are asked to focus on the head motion and to score each video after a video pair is displayed. The questionnaire for the subjective test is attached in the appendix A.

Table 5.3: Naturalness test for head motion synthesis. Average MOS scores (5=Excellent, 1=Bad) and the confidence intervals ($\bar{u}_k \pm \delta_k$) are analyzed.

Pair No.	Repeated head motion	Flexible head motion
1	2.7 ± 0.5	3.8 ± 0.5
2	2.2 ± 0.4	2.9 ± 0.5
3	2.9 ± 0.3	4.2 ± 0.3
4	3.0 ± 0.4	3.4 ± 0.5
5	2.7 ± 0.4	3.5 ± 0.6
avg. MOS	2.7	3.6

Table 5.3 gives the average MOS scores and the confidence intervals for each animation. The overall MOS score of animations with flexible head motion is 3.6, and the overall MOS score of animations with repeated head motion is 2.7. The feedback of the viewers is that they can recognize repeated head motion immediately. Such talking heads are boring and unrealistic, while the talking heads with flexible head motions are more lifelike and engaging.

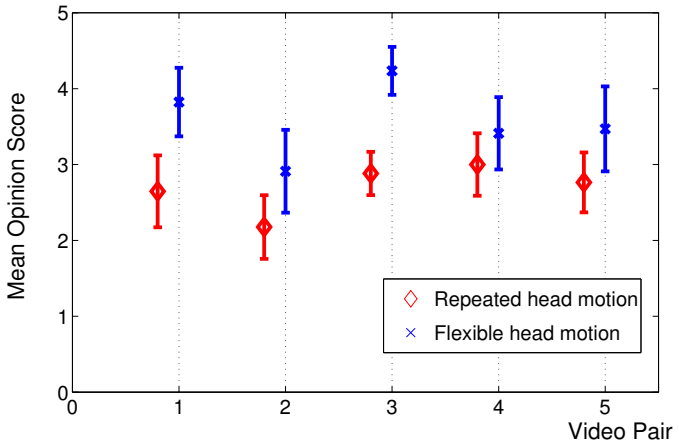


Figure 5.13: Results of the subjective test for head motion synthesis.

The results of the MOS scores are plotted in Fig. 5.13. We can see that the scores of the talking heads with flexible head motions (blue cross marker) are obviously over the

scores of the talking heads with repeated head motions (red diamond marker).

Good animations with head motion synthesis depend not only on synthesis of smooth and realistic head movements, but also on the perception of the head motions, whether they are related to the spoken words. If the head motions match the spoken words very well, the overall quality of animations is high, otherwise, the quality is low. Sentence 2 is an example to generate head motion with a "nod". Some viewers find that the nod is unnecessary or unrelated to the spoken words. Most viewers indicate that the quality of TTS audio has also decreased the perception of the animations significantly, since the TTS audio has no emotion, which is unexpected by the viewers. In case of sentence 3, the head motion patterns match the meaning of the sentence very well, so that the perceived quality of the animations becomes higher. From the results of the subjective test, the qualities of sentence 1, 4 and 5 are between the two cases.

The overall MOS scores for head motion synthesis are lower than the ones for facial expression synthesis. This is due to the fact that audios are from the TTS synthesizer in the head motion synthesis, which decreases the quality in perception of the talking heads.

Furthermore, physical modeling of head motion transition is unnecessary, since head motion transition by Hermite interpolation is realistic and smooth according to the results of the subjective test.

Supplemental materials of head motion synthesis can be downloaded from http://www.tnt.uni-hannover.de/project/facialanimation/demo/headmotion.

The sources of the expressive database and all animations presented in this thesis are listed in the appendix B.

6 Conclusions

This thesis presents an image-based talking head which provides a higher level of realism when compared to 3D model-based talking head. We first improve the image-based reference talking head system. Based on the optimized system, the image-based talking head system is extended with facial expression synthesis and head motion synthesis. This makes the talking head appear more realistic and lifelike.

Lip Synchronization

The image-based talking head system consists of an off-line audio-visual analysis and an on-line unit selection synthesis. Compared to the reference system, two improvements are made for lip synchronization. One is the feature detection in the analysis part. Instead of the reference method, color template based approach, AAM (Active Appearance Models) based facial feature detection is used to find geometric parameters of mouth images. By doing so, the accuracy of facial features is improved to sub-pixel. The other improvement is to optimize the unit selection in the synthesis part. Since the unit selection is a nonlinear system, Pareto optimization is introduced to train the unit selection algorithm such that the visual speech synthesis is stable for arbitrary input texts. The optimization criteria include lip synchronization, visual smoothness and others.

Various state-of-the-art talking head systems were evaluated in the LIPS challenge, which focused on lip synchronization with a goal to develop standardized evaluation procedures. This optimized image-based talking head system was presented in the special session “LIPS2008: Visual Speech Synthesis Challenge” at the conference of Interspeech 2008. Using a given dataset in the challenge, the presented competing animation systems were evaluated subjectively. We received the Golden Lips Award for audio-visual consistency for having the most natural talking heads.

Facial Expression Synthesis

The image-based talking head system is extended to enable synthesis of realistic facial expressions accompanying speech, given arbitrary text input and control tags of facial expression. This is the first image-based facial animation system to synthesize facial expressions. As an example of facial expression primitives, a smile is used. First, three types of videos are recorded: a performer speaking without any expressions, smiling while speaking, and smiling after speaking. By analyzing the recorded audio-visual data, an expressive database is built and contains normalized neutral mouth images and smiling

mouth images, as well as their associated features and expressive labels. In order to synthesize realistic facial expressions, natural expression changes are analyzed, and several rules of when and how human smiles are summarized. In order to generate smooth transitions from one expression to another, the viseme transition in the expressive database has been analyzed, from which the viseme switching matrix is derived. Based on these rules and the viseme switching matrix, the unit selection algorithm selects and concatenates appropriate mouth image segments from the expressive database.

In the objective evaluation of the animations, the match rate of closures between animated and real sequences is the most important criterion, since closures provide the impression that lips and sound are synchronized. The precise shapes of the openings are less important for the perceived quality of the coarticulation. Experimental results show that the closures between animated and real sequences are always matched.

In the subjective test of facial expression synthesis, the standard Mean Opinion Score (MOS) approach is used. Thirty viewers are involved in the subjective test and they compare the animations with the real videos. The animations are driven by the original audios, so that the audio has no influence on the subjective test. The subjective test shows that the overall MOS score of the animations is 3.9, while the overall MOS score of the real videos is 4.1. The confidence intervals overlap largely. Furthermore, the viewers indicate that they cannot distinguish real smiles from synthesized ones.

Eye animation is related to mouth animation while smiling. Image-based eye animation [28] can be directly integrated in our system. Other facial expressions, such as anger and surprise, can also be switched through neutral mouths so that smooth animations are possible.

Head Motion Synthesis

Even though animations of short sentences are hard to distinguish from recorded videos, longer sentences are immediately identified as synthetic due to a lack of natural head movements. A novel approach to add flexible head motions to talking heads is developed. This is the first image-based facial animation system to synthesize head motions. First, head motion patterns are collected from original recordings. These head motion patterns are recorded video segments with different head motions, such as the nodding and shaking of the head. The head motion is synthesized by selecting and concatenating appropriate head motion patterns according to the input text or head motion tags. In order to join these patterns, optical flow based morphing is used to smooth transitions without introducing noticeable discontinuities. To generate realistic head motion, the speed and acceleration of the head movements should be considered. The dynamics of head motions are analyzed. The transition between head motion patterns is required. Since the Hermite function considers the orientation of start and end frames, the interpolation of Hermite function achieves smooth head motion transitions subjectively. Therefore, the physical modeling of head motion is unnecessary.

In the subjective test, the standard mean opinion score approach is used to evaluate

head motion synthesis. Talking heads with repeated head motion and with flexible head motion are synthesized and assessed. The subjective test shows that the overall MOS score of animations with flexible head motion is 3.6, and the overall MOS score of animations with repeated head motion is 2.7. The feedback of the viewers is that talking heads with flexible head motion is more realistic and lifelike than the talking heads with repeated head motion.

In this thesis we have improved the basic facial animation system. Furthermore, the facial animation system is extended with realistic facial expression and added with flexible head motion. Combined with a dialog system, the developed expressive talking head, integrated with eye animation, will open a wide space in many applications. These applications include web-based customer service, E-education, and E-care. The expressive talking head is able to give a very personalized and believable interface for human-machine communication. These intuitive and efficient interfaces will be extensively used in the near future.

A Questionnaire for Subjective Tests

Subjective Test Sheet No.___

Information of test person:

- Gender: ☐ Female / ☐ Male
- Age: ☐ 19-22, ☐ 23-26, ☐ 27-30, ☐ 31-35, ☐ over 36
- English level: ☐ Normal or ☐ Good
- Major: ☐ EE / ☐ CS / ☐ Other

Rules:

- Scoring the overall quality of videos with 1-5.
  1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent
- Giving comments, which part is not satisfied.
  1: Mouth, 2: Eye, 3: Head, 4: Unknown

Scoring Test for Facial Expression Synthesis

Pair No.	Video 1	Comment 1	Video 2	Comment 2
1				
2				
3				
4				
5				
6				
7				
8				
9				

Scoring Test for Head Motion Synthesis

Pair No.	Video 1	Comment 1	Video 2	Comment 2
1				
2				
3				
4				
5				

B Animations and Videos

The animations and recorded videos presented in this thesis can be found in our Web server.

- Basic animations:
  http://www.tnt.uni-hannover.de/project/facialanimation/demo/mouth.
- Facial expression synthesis:
  http://www.tnt.uni-hannover.de/project/facialanimation/demo/emotion.
- Head motion synthesis:
  http://www.tnt.uni-hannover.de/project/facialanimation/demo/headmotion.

The expressive database is made available to the facial animation community for scientific purposes.

- Recorded clips:
  http://www.tnt.uni-hannover.de/project/facialanimation/demo/database.

Bibliography

[1] J. Beskow. *Talking Heads - Models and Applications for Multimodal Speech Synthesis*. PhD thesis, Department of Speech, Music and Hearing, KTH, Stockholm, 2003.

[2] H. Prendinger and M. Ishizuka. *Life-Like Characters: Tools, Affective Functions, and Applications*. Springer-Verlag, 2004.

[3] K. Liu and J. Ostermann. Realistic talking head for human-car-entertainment services. In *Proceedings of the Informationssysteme für Mobile Anwendungen IMA 2008*, pages 108–118, 2008.

[4] B. Theobald, S. Fagel, G. Bailly, and F. Elsei. LIPS2008: Visual Speech Synthesis Challenge. In *Proceedings of Interspeech 2008*, pages 2310–2313, 2008.

[5] I.S. Pandzic, J. Ostermann, and D. Millen. User evaluation: Synthetic talking faces for interactive services. *The visual computer*, 15(7):330–340, 1999.

[6] J. Ostermann and A. Weissenfeld. Talking Faces - Technologies and Applications. In *Proceedings of ICPR04*, volume 3, pages 826–833, 2004.

[7] K. Liu, A. Weissenfeld, J. Ostermann, and X. Luo. Robust AAM building for morphing in an image-based facial animation system. In *Proceedings of ICME 2008*, pages 933–936, 2008.

[8] K. Liu and J. Ostermann. Minimized Database of Unit Selection in Visual Speech Synthesis without Loss of Naturalness. In *Proceedings of Computer Analysis of Images and Patterns*, pages 1212–1219, 2009.

[9] E. Cosatto, J. Ostermann, H.P. Graf, and J. Schroeter. Lifelike Talking Faces for Interactive Services. In *Proceedings of the IEEE*, volume 91, pages 1406–1429, 2003.

[10] K. Liu and J. Ostermann. Optimization of an Image-based Talking Head System. *Special issue on animating virtual speakers or singers from audio: Lip-synching facial animation, EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[11] G. Bailly, M. Bérar, F. Elisei, and M. Odisio. Audiovisual Speech Synthesis. *International Journal of Speech Technology*, 6(4):331–346, 2003.

[12] Z. Deng and U. Neumann. *Data-driven 3D facial animation*. Springer-Verlag, 2008.

[13] J. Ostermann. Animation of synthetic faces in MPEG-4. In *Proceedings of Computer Animation*, pages 49–55, 1998.

[14] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin. Synthesizing Realistic Facial Expressions from Photographs. In *Proceedings of SIGGRAPH 1998*, pages 75–84, 1998.

[15] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of SIGGRAPH 1999*, Computer Graphics Proceedings, Annual Conference Series, pages 187–194. ACM, ACM Press / ACM SIGGRAPH, 1999.

[16] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *EUROGRAPHICS 2003*, volume 22 of *24th Annual Conference Series*, pages 641–650. the European Association for Computer Graphics, 2003.

[17] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. The Digital Emily Project: Photoreal Facial Modeling and Animation. In *SIGGRAPH Course*, 2009.

[18] D. Terzopoulos and K. Waters. Physically-based facial modeling analysis and animation. *Journal of Visualization and Computer Animation*, 1(4):73–80, 1990.

[19] K. Waters and J. Frisbie. A coordinated muscle model for speech animation. In *Proceedings of Graphics Interface*, pages 163–170, 1995.

[20] K. Kaehler, J. Haber, H. Yamauchi, and H.P. Seidel. Head Shop: Generating animated head models with anatomical structure. In *Proceedings of SIGGRAPH 2002*, pages 55–63, 2002.

[21] P. Ekman and W.V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[22] C. Bregler, M. Covell, and M. Slaney. Video rewrite: driving visual speech with audio. In *Proceedings of SIGGRAPH 1997*, pages 353–360, 1997.

[23] T. Ezzat and T. Poggio. Miketalk: a talking facial display based on morphing visemes. In *Proceedings of Computer Animation*, pages 96–102, 1998.

[24] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of SIGGRAPH 2002*, pages 388–397, 2002.

[25] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[26] B. Theobald, J.A. Bangham, I. Matthews, and G. Cawley. Near-videorealistic synthetic talking faces: Implementation and evaluation. *Speech Communication*, 44:127–140, 2004.

[27] M. Banf and V. Blanz. Example-Based Rendering of Eye Movements. *Computer Graphics Forum*, 28(2):659–666, 2009.

[28] A. Weissenfeld, K. Liu, and J. Ostermann. Video-realistic image-based eye animation via statistically driven state machines. *The Visual Computer, International Journal of Computer Graphics*, 2009.

[29] M.M. Cohen and D.W. Massaro. Modeling coarticulation in synthetic visual speech. *ACM Transactions on Graphics*, pages 139–156, 1993.

[30] N. Hewlett and W.J. Hardcastle. *Coarticulation: Theory, Data and Techniques*. Cambridge University Press, 2000.

[31] F.I. Parke. Parameterized Models for Facial Animation. *IEEE Computer Graphics and Applications*, 2(9):61–68, 1982.

[32] J. Beskow. Rule-based Visual Speech Synthesis. In *Proceedings of Eurospeech 95*, pages 299–302, 1995.

[33] M. Brand. Voice puppetry. In *Proceedings of SIGGRAPH 1999*, pages 21–28, 1999.

[34] K. Choi, Y. Luo, and J.N. Hwang. Hidden Markov Model Inversion for Audio-to-visual Conversion in an MPEG-4 Facial Animation System. *Journal of VLSI Signal Processing*, 29(1-2):51–61, 2001.

[35] E. Cosatto. *Sample-based talking-head synthesis*. PhD thesis, EPFL, Switzerland, 2002.

[36] E. Zitzler, M. Laumanns, and S. Bleuler. A tutorial on evolutionary multiobjective optimization. In *Proceedings of the Multiple Objective Metaheuristics (MOMH 03)*, 2003.

[37] J. Von Livonius, H. Blume, and T.G. Noll. Flexible Umgebung zur Pareto-Optimierung von Algorithmen - Anwendungen in der Videosignalverarbeitung. In *ITG*, 2007.

[38] P. Ekman and W.V. Friesen. *Unmasking the face. A guide to recognizing emotions from facial clues*. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

[39] P. Ekman. *Emotion in the human face*. New York: Cambridge University Press, 1982.

[40] A.M. Tekalp and J. Ostermann. Face and 2D mesh animation in mpeg-4. *Signal Processing: Image Communication*, 15(4):387–421, 2000.

[41] I.S. Pandzic and R. Forchheimer. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. Wiley, 2002.

[42] I.A. Essa, S. Basu, D. Trevor, and A.P. Pentland. Modeling, tracking and interactive animation of faces and heads using input from video. In *Proceedings of Computer Animation*, pages 68–79, 1996.

[43] I.A. Essa and A.P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[44] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[45] Y. Cao, P. Faloutsos, and F. Pighin. Unsupervised learning for speech motion editing. In *SIGGRAPH Symposium on Computer Animation*, pages 225–231. Eurographics Association, 2003.

[46] Y. Cao, W.C. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics*, 24(4):1283–1302, 2005.

[47] E. Chuang and C. Bregler. Mood Swings: Expressive Speech Animation. *ACM Transactions on Graphics*, 24(2):331–347, 2005.

[48] F. Bettinger and T.F. Cootes. A model of facial behaviour. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, pages 797–806, 2004.

[49] K.G. Munhall, J.A. Jones, D.E. Callan, T. Kuratate, and E. Vatikiotis-Bateson. Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science*, 15(2):133–137, 2004.

[50] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversation agents. In *Proceedings of SIGGRAPH 1994*, pages 413–420, 1994.

[51] C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, 1996.

[52] M. Brkic, K. Smid, T. Pejsa, and I.S. Pandzic. Towards Natural Head Movement of Autonomous Speaker Agent. In *Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems*, pages 73–80, 2008.

[53] H.P. Graf, E. Cosatto, V. Strom, and F.J. Huang. Visual Prosody: Facial Movements Accompanying Speech. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, pages 96–102, 2002.

[54] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S.S. Narayanan. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 15, pages 1075–1086, 2007.

[55] C. Busso, Z. Deng, U. Neumann, and S.S. Narayanan. Natural head motion synthesis driven by acoustic prosodic features. In *Proceedings of Computer Animation and Virtual Worlds 2005*, volume 16, pages 283–290, 2005.

[56] D. Heylen. Listening Heads. In *Modeling Communication with robots and virtual humans*, volume 4930 of *Lecture Notes in Artificial Intelligence*, pages 241–259. Springer Verlag, 2008.

[57] M.B. Stegmann, B.K. Ersboll, and R. Larsen. Fame - a flexible appearance modeling environment. *IEEE Transactions on Medical Imaging*, 22(10):1319–1331, 2003.

[58] H.P. Graf, E. Cosatto, and G. Potamianos. Robust Recogniton of Faces and Facial Features with a Multi-Modal System. *IEEE Systems, Man and Cybernetics*, pages 2034–2039, 1997.

[59] K. Liu and J. Ostermann. Realistic Facial Expression Synthesis for an Image-based Talking Head. In *Proceedings of ICME 2011*, 2011.

[60] K. Liu and J. Ostermann. Realistic Head Motion Synthesis for an Image-based Talking Head. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, 2011.

[61] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, volume 3024, pages 25–36. Springer, May 2004.

[62] K. Liu and J. Ostermann. Evaluation of an Image-based Talking Head with Realistic Facial Expression and Head Motion. In *Proceedings of CASA (Computer Animation and Social Agents) Workshop on Emotion-based Interaction*, 2011.

[63] D. Oberkampf, D. Dementhon, and L. Davis. Iterative Pose Estimation Using Coplanar Feature Points. In *Internal Report, CVL, CAR-TR-677*. University of Maryland, 1993.

[64] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T Next-Gen TTS System. In *Joint Meeting of ASA, EAA AND DAGA*, pages 18–24, 1999.

[65] A.J. Hunt and A.W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP 1996*, pages 373–376, 1996.

[66] A. Weissenfeld, O. Urfalioglu, K. Liu, and J. Ostermann. Robust Rigid Head Motion Estimation based on Differential Evolution. In *Proceedings of ICME 2006*, pages 225–228, 2006.

[67] T. Beier and S. Neely. Feature-Based Image Metamorphosis. *Computer Graphics*, 26(2):35–42, 1992.

[68] A. Weissenfeld, K. Liu, S. Klomp, and J. Ostermann. Personalized unit selection for an image-based facial animation system. In *Proceedings of MMSP*, 2005.

[69] S.E. Chen and L. Williams. View interpolation for image synthesis. In *Proceedings of SIGGRAPH 1997*, pages 279–288, 1993.

[70] D.E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley Publishing Company, 1989.

[71] X. Hue. *Genetic Algorithms for Optimization*. Edinburgh, 1997.

[72] F. Alias and X. Llora. Evolutionary Weight Tuning for Unit Selection Based on Diphone Pairs. In *IlliGal Report No. 2003018, University of Illinois at Urbana-Champaign*, 2003.

[73] R. Kumar. A Genetic Algorithm for Unit Seletion based Speech Synthesis. In *Interspeech - ICSLP*, pages 1233–1236, 2004.

[74] K. Liu and J. Ostermann. Realistic Facial Animation System for Interactive Services. In *Proceedings of Interspeech 2008*, pages 2330–2333, 2008.

[75] S. Fagel, G. Bailly, and B. Theobald. Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[76] http://www.lips2008.org. LIPS2008: Visual Speech Synthesis Challenge.

[77] I. Jolliffe. *Principal Component Analysis*. Springer Verlag, New York, 1989.

[78] P. Ekman, W.V. Friesen, and M. O'Sullivan. Smiles when lying. *Journal of Personality and Social Psychology*, 54:414–420, 1988.

[79] A. Schoedl, R. Szeliski, D.H. Salesin, and I. Essa. Video Textures. In *Proceedings of SIGGRAPH 2000*, pages 489–498, 2000.

[80] Advanced Television Systems Committee(ATSC). *ATSC Implementation Subcommittee Finding: Relative Timing of Sound and Vision for Broadcast Operations Advanced Television*. Doc. IS-191, 2003.

[81] ITU-R BT.500 11. Methodology for the subjective assessment of the quality of television pictures. 2002.

Lebenslauf

Kang Liu
geboren am 10.10.1977 in Sichuan, V.R. China
verheiratet, drei Kinder

Ausbildung

1984	–	1990	Grundschule in Tianjin, China
1990	–	1996	Gymnasium (Tianjin No.45 Middle School) in Tianjin, China Abschluss: Abitur
1996	–	2001	Studium der Mechatronik an der Zhejiang Universität, Hangzhou, China Abschluss: Bachelor of Science (B.Sc.)
2001	–	2004	Fortsetzung des Studiums der Mechatronik an der Zhejiang Universität, Hangzhou, China Abschluss: Master of Science (M.Sc.)

Beruf

2004	–	2009	Doktorand des Instituts für Informationsverarbeitung, Leibniz Universität Hannover
seit		2010	wissenschaftlicher Mitarbeiter am Institut für Informationsverarbeitung, Leibniz Universität Hannover