Visual Speech Recognition:
Lip Segmentation and Mapping

Alan Wee-Chung Liew
Griffith University, Australia

Shilin Wang
Shanghai Jiaotong University, China

Medical Information Science REFERENCE

MEDICAL INFORMATION SCIENCE REFERENCE

Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Amanda Appicello
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Liew, Alan Wee-Chung, 1968-

Visual speech recognition : lip segmentation and mapping / Alan Wee-Chung Liew and Shilin Wang, Editors.

p. cm.

Includes bibliographical references and index.

Summary: "This book introduces the readers to the various aspects of visual speech recognitions, including lip segmentation from video sequence, lip feature extraction and modeling, feature fusion and classifier design for visual speech recognition and speaker verification"--Provided by publisher.

ISBN 978-1-60566-186-5 (hardcover) -- ISBN 978-1-60566-187-2 (ebook)

1. Automatic speech recognition. 2. Speech processing systems. I. Wang, Shilin. II. Title.

TK7895.S65L44 2009

006.4--dc22

2008037505

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter XVII
Lip Motion Features for Biometric Person Recognition

Maycel Isaac Faraj
Halmstad University, Sweden

Josef Bigun
Halmstad University, Sweden

ABSTRACT

The present chapter reports on the use of lip motion as a stand alone biometric modality as well as a modality integrated with audio speech for identity recognition using digit recognition as a support. First, the auhtors estimate motion vectors from images of lip movements. The motion is modeled as the distribution of apparent line velocities in the movement of brightness patterns in an image. Then, they construct compact lip-motion features from the regional statistics of the local velocities. These can be used as alone or merged with audio features to recognize identity or the uttered digit. The author's present person recognition results using the XM2VTS database representing the video and audio data of 295 people. Furthermore, we present results on digit recognition when it is used in a text prompted mode to verify the liveness of the user. Such user challenges have the intention to reduce replay attack risks of the audio system.

INTRODUCTION

The performance of multimodal systems using audio and visual information in biometrics is superior to those of the acoustic and visual subsystems (Brunelli and Falavigna (1995)), (Tang and Li (2001)), (Bigun et al. (1997b)), and (Ortega-Garcia et al. (2004)) because these systems have a high potential for delivering noise robust biometric recognition systems compared to the corresponding single modalities. This is the general motivation for why there has been increased interest in multimodal biometric iden-

tity recognition. For example in audio based person recognition, phoneme sounds can be acoustically very similar between certain individuals and therefore hard to differentiate. By adding information on lip-motion, the discrimination of identities can be improved.

Speaker recognition using visual information in addition to acoustic features is particularly advantageous for other reasons too. It enables interactive person recognition which can be used to reduce impostor attacks that rely on prerecorded data. Raising antispoofing barriers, known as liveness detection, e.g. to determine if the biometric information being captured is an actual measurement from the live person who is present at the time of capture, for biometric systems is becoming increasingly necessary.

In this chapter extraction of lip-motion features that takes advantage of the spatiotemporal information in an image sequence containing lip-motion is discussed. Motion features are suggested for recognition of human identities and word (digit) recognition which can be used for liveness detection. The discussions include filtering, feature extraction, feature reduction, feature fusion and classification techniques.

Section 2 presents a review of some previous studies relevant to the chapter. The emphasis is on audio-visual systems rather than the massive research body existing in the individual recognition technologies. In particular, lip features suggested previously are discussed in greater detail.

Section 3 presents the theory of three different concepts of motion estimation which is directly relevant to this chapter. The motion estimation techniques based on texture translations and line translations are explicitly contrasted against each other. A further quantification of the speed accuracy of the used motion estimation that assumes moving lines or edges is given. How motion is exploited in other audio-visual recognition studies is also discussed.

In Section 4 we present a discussion on how one can use estimated velocities to produce compact feature vectors for identity recognition and liveness detection by uttered digits. A technique for quantization and dimension reduction is presented to reduce the amount of extracted features. The section also presents the audio and visual features concatenated at the feature level allowing, the integration of different audio and video sampling rates. The visual frames come at one fourth pace of the audio frames do, but contain more data. Yet the final concatenated feature vector must come at the same pace and contain approximately the same amount of data each, to avoid favoring one over the other. The section also presents the performance of visual information as an audio complement feature in speaker recognition and speech recognition using the XM2VTS database. We present a single and multimodal biometric identity recognition system based on the lip-motion features using a Gaussian Mixture Model (GMM) and a Support Vector Machine (SVM) as model builders. Furthermore, we present the experimental test using only one word (digit) to recognize the speaker identity. A discussion on related studies exploiting different techniques for audio-visual recognition is also included.

Section 5 discusses the conclusions of the chapter and presents directions for future work.

REVIEW

In speech recognition, two widely used terms are *phoneme* and *viseme*. The first is the basic linguistic unit and the later is the visually distinguishable speech unit (Luettin (1979)).[1] Whereas the use of *visemes* has been prompted by machine recognition studies, and hence it is in its start stage, the idea of phonemes is old. The science of Phonetics has for example been playing a major role in human language studies. The consonant letters complemented with vocals are approximations of phonemes and the alphabet belongs to greatest inventions of humanity.

Early work from (Petajan (1984)) and (Mase and Pentland (1991)) introduced visual information by the use of lip information as an important aid for speech recognition. (Yamamoto et al. (1998)) proposed visual information semi automatically mapped to lip movements through the aid of sensors put around the mouth to highlight the lips. The experimental results showed that significant performance could be achieved even by only using visual information. (Kittler et al. (1997)) presented a study using geometric features of the lip shapes from model based lip boundary tracking confirming the importance of lip information in identity recognition.

(Luettin et al. (1996)) presented a speaker identification system based only on dynamic visual information from video sequences containing the lip region. The geometrical features of the lips contained information about the shape and intensity information of the lips. The experiments were carried out by 12 speakers uttering digits and were later extended to the M2VTS database (37 speakers) by (Jourlin et al. (1997)). The person identification system based on Hidden Markov Model (HMM) achieved 72.2% using labial information and 100% using merged acoustic and visual features. They achieved good performance with joint systems utilizing a score fusion (late integration) method. They used 14 lip shape parameters, 10 intensity parameters, and the scale as visual features, resulting in a 25 dimensional visual feature vector. The speaker verification system score is computed as a weighted sum of the audio and visual scores.

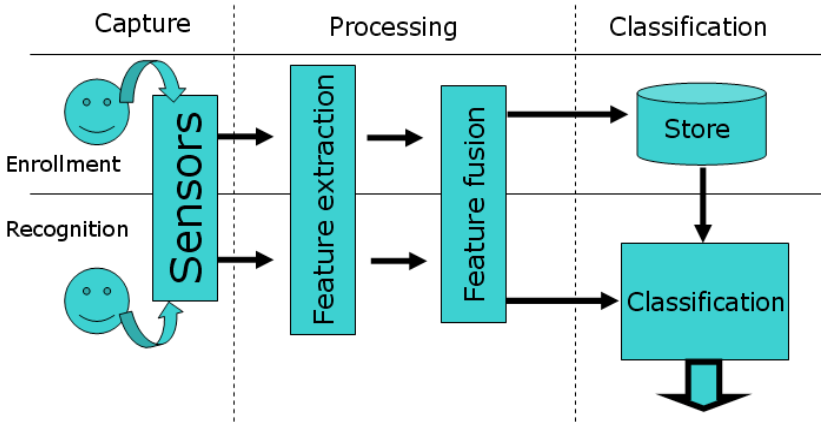
(Brunelli and Falavigna (1995)) developed a text-independent speaker identification system exploiting acoustical information in combination with visual information from static face images. The system is based on several experts: two acoustic modalities (static and dynamic), containing derived features from short time spectral analysis of the speech signal, and three visual experts containing information from the eyes, nose and mouth. By using weighted function to classify the experts, the system performed well on approximately 90 speakers. Other studies using static visual information in recognition systems are (Tistarelli and Grosso (2000)) utilizing morphological filtering for a facial/eye localization followed by a simple matching algorithm for identity verification, (Duc et al. (1997) and Ben-Yacoub et al. (1999)) using Gabor filter responses on sparse graphs on faces but in the context of an audio-visual speaker verification system, (Sanderson and Paliwal (2004)) using Principal Component Analysis (PCA) for face feature extraction for identity verification and (Hazen et al. (2003)) using visual information from the different components in the face in a speaker identification system.

(Wark and Sridharan (1998)) developed a speaker verification system based on dynamic lip contour features extracted by Linear Discriminant Analysis (LDA) in combination with principal component analysis, yielding favorable results. This study was extended to merge audio-visual information by late integration using mixed densities of Gaussians, (Wark et al. (1999)).

(Dieckmann et al. (1997)), proposed a system using multimodal visual information from a video sequence. The modalities, face, voice and lip movement, were fused utilizing voting and opinion fusion. A minimum of two experts had to agree on the opinion and the combined opinion had to exceed the predefined threshold. Other related work has exploited dynamic visual information for speaker recognition (Frischholz and Dieckmann (2000)) and (Kittler et al. (1997)), and used multimodal information for speaker identification (Bigun et al. (1997a) and Bigun et al. (1997b)).

(Nakamura (2001)) proposed a method based on HMMs to integrate multimodal information considering synchronization and weights for different modalities. He built compound HMMs, each including a large number of states, incorporating states in an audio HMM and a visual HMM for all possible combinations. The system showed improved performance of speech recognition when using multimodal information.

Figure 1. The figure illustrates a block diagram of audio-visual biometric system used for speech and speaker recognition studies of this chapter



By utilizing the lip contour, such as the contour height, width and area, (Chen 2001)) presented a speech recognition system based on these features with real-time tracking using the multi-stream HMM without automatic weight optimization.

Biometric Recognition Framework

The generic framework describing biometric recognition systems is useful to understand the present work. We describe it in **Fig. 1** and it consists of three main blocks that of capturing, processing (feature extraction and feature fusion) and classification. In the case of offline recognition one does not need to take into account the capturing methods. This chapter falls under the category of processing block, proposing novel methods mainly for visual feature extraction. The classification block, also known as matching, is based on already developed systems such as GMM (using HTK toolkit) and SVM using SVM library. Below, we outline the existing methods used for feature extraction and feature fusion.

Visual Feature Extraction

The main benefit, in speech recognition, of using visual cues is that they are complementary to the acoustic signal: some phonemes that are difficult to understand acoustically in noisy environments can be easier to distinguish visually, and vice versa.

We distinguish two challenges in lip features processing, **Fig. 2**, i) detection of face/mouth/lips and ii) extraction of features. The first problem amounts to finding and tracking a specific facial part (mouth,

Figure 2. The figure illustrates visual feature representation approaches

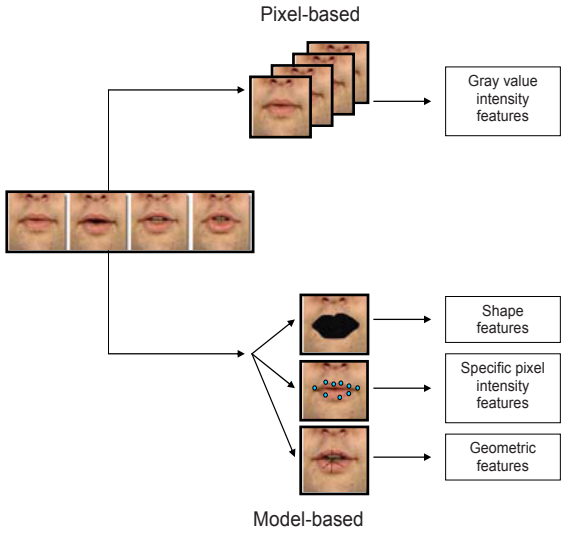
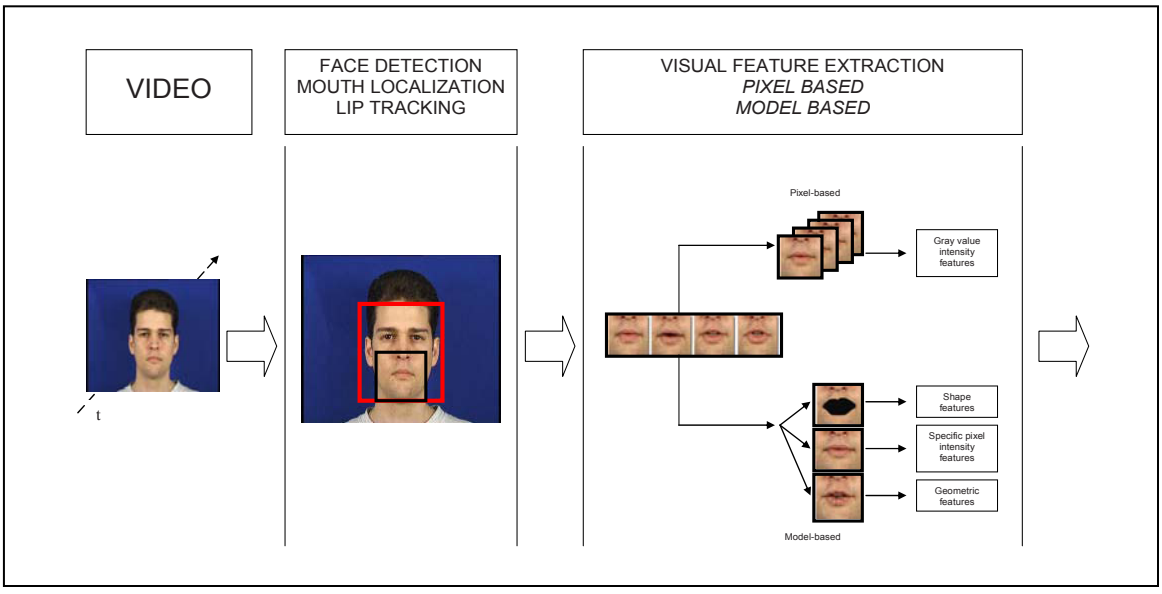


Figure 3. The figure illustrates the stages of information extraction in a biometric recognition system utilizing lip features. The video stream is first processed by a tracking and detection technique and in the second block the lip features are extracted from the tracked object (mouth region, lip contours, lips, etc.)



lips, lip contours etc.) whereas the second problem comprises the extraction of the visual information in terms of a small number of informative variables or measurement entities.

Successful mouth tracking is still challenging in cases where the background, head pose and lighting vary greatly, (Iyengar and Neti (2001)). After successful face detection, the region is processed further to obtain lip features. Though not very detailed in terms of lip motion description, even the bounding boxes of lip regions can reveal useful lip features if they are estimated for every frame independently because such rectangles reveal the dynamic evolvement of the height and width (Zhang et al. (2002)) and (Luettin et al. (1996)) during speech production. However, the lip information within the mouth region is most commonly extracted. Visual features are then extracted either from single frames (static) or from a set of consecutive images (dynamic). The visual features can be categorized into two groups: pixel based approaches and model based approaches, **Fig. 3**, regardless of whether they model the static or dynamic information.

- *Pixel based approach*: Each pixel in the image participates into computations of features such as Fourier transform, discrete cosine transformation, optical flow, etc. The features are directly pixel driven without form constraints between pixels which are to be contrasted to for example lip contour models. However, even pixel driven techniques presuppose the extraction of at least a sufficiently narrow region containing the mouth. The extracted lip region is often processed further with normalization techniques in an attempt to improve resilience against disturbances caused by head pose and lighting information. The pixels of the found mouth region can be mapped to a different space for lip features extraction. A popular mapping is projection to an image basis obtained by PCA. Such methods model static information in single image frames explicitly, even though they implicitly can represent the dynamic information between the frames, such as motion. Motion estimation (optical flow) which can capture the lip velocity and acceleration information in each pixel over time is by contrast an approach that explicitly models motion information. (Chan (2001)) presented a combined geometric lip features utilizing the PCA projection. The PCA are determined by a subset of pixels contained within the mouth. (Chiou and Hwang (1997)) on the other hand, presented combination of a number of snake lip contour vectors with PCA features from the color pixel values of a rectangle mouth region of interest. Furthermore, (Neti et al. (2000)) and (Matthews et al. (2001)) uses joined model of PCA techniques for estimating dimensionalities of shape models and appearance vectors. In this chapter we pursue the motion modeling approach to extract lip-motion features. The features are undoubtedly pixel driven, yet it describes but the allowable motions are restricted in direction to conform to what can be expected from a lip-motion. Using the motion estimation technique, requirements for accurate mouth state or lip contour extraction may be eased since a rough detection of the mouth region is sufficient to obtain visual features.
- *Model based approach*: Geometric and shape based methods represent the dynamic visual lip images by lip contour information and shape information of lips (Chan (2001)) and (Chiou and Hwang (1997)). These features are normally extracted from the region-of-interest equipped with a lip tracking preprocessing algorithm. Excluding the preprocessing part (lip contour tracking), these methods require less computation since they only work with a few control points. However, lip contour detection can be computationally demanding and prone to errors.

Next, we will present three mouth region features, proposed by (Jourlin et al. (1997)), (Dieckmann (1997)) and (Liang et al. (2002)) as they represent well the two categories outlined above.

Lip Feature Representation Approaches

Jourlin et al. (1997)

The lip feature extraction proposed by (Luettin et al. (1996)) is model based and assumes that most relevant information is contained in the *shape* (contours) of the speaker's lips.

The approach consists of a combination of lip contour information and the gray level distribution around the mouth area. During speech production the lip shape varies. For each speaker a spatiotemporal model that describes the mouth shape of the speaker and its temporal change is built.

They use a shape model that describes the outer and inner lip contour and a deformable gray level model to describe intensity values around the lip contours. Active shape models are used to locate, track and parameterize the lips over an image sequence. The principal modes of deformation are obtained by performing PCA on a labeled training set. A shape model is then approximated by a linear combination of the first few principal modes of deviation from the average lip curve.

Gray levels representing the intensities perpendicular to the contour at each control point of the model are concatenated to form a *profile* vector. The profile vectors of speakers in a training set are subjected to PCA to capture the profile variation modes. A profile is then represented as a linear combination of deviation modes (PCA basis) from the average gray profiles. The concatenated vectors of all model points represent a profile model for a speaker. PCA is performed on all profiles to obtain the principal modes of the profile variation.

In a lip sequence unseen by the system, the profile model is used to enable tracking whereby the curve parameters (weights, or basis coefficients) corresponding to the curves defined by the tracked control points are subsequently computed. The found contour and profile parameters are used as lip features for speaker recognition.

Dieckmann et al. (1997)

(Dieckmann et al. (1997)) presented a speaker recognition system (SESAM) using lip features that are pixel based. Beside the lip information, facial information from the speaker was added.

Their approach is based on the optical flow analysis using the (Horn and Schunck (1981)) method applied to mouth sequences. The Horn and Schunck method is a differential motion estimation method based on two frame differences. The main difference between Horn and Schunck technique and Lukas and Kanade technique is the weighting function to enforce the spatial continuity of the estimated optical flow. If we set the weighting function to zero we will have the method suggested by Lucas and Kanade, which we will discuss in detail in Section Performance measurement.

The lip movement estimation of the SESAM system calculates a vector field representing the local movement of each two consecutive frames in the video sequence. An averaging is used to reduce the amount of velocity vectors to 16 (one fourth of the original size), representing velocities in 16 sub regions. 3D fast Fourier transforms are applied on the velocity vectors to represent the movement of identifiable points from frame to frame.

Zhang et al. (2002)

Automatic speech reading as well as speaker recognition by visual speech has been studied by (Zhang et al. (2002)). In this pixel based work, the authors suggest a primarily color driven algorithm for automatically locating the equivalents of 3 bounding boxes for the i) mouth region, outer lip contour, inner lip contour. Though this is not part of the features, motion is also modeled but after that the points of the bounding boxes have been identified. The used fusion is a decision fusion consisting of averaging the audio and lip expert scores. The study presents visual feature performance comparisons and confirms that the visual information is highly effective for improving recognition performance over a variety of acoustic noise levels.

Liang et al. (2002)

The technique proposed by (Liang et al. (2002)) is categorized as pixel based because its features are extracted using all pixels without an explicit constraint on the shape of the lips.

The visual observation vector is extracted from the mouth region using basically two algorithms in cascade. The gray pixels in the mouth region are mapped to a 32 dimensional eigenvectors produced from a PCA decomposition of gray value deviations from the average mouth region. This is computed from a set of approximately 200000 mouth region images. Temporally, the feature are up sampled and normalized to match the acoustic sample rate. The visual observation vectors are concatenated and projected on a 13 class linear discriminant space, using Linear Discriminant analysis (LDA). By this the visual features are reduced to a new set of dimension 13.

Other Relevant Studies

PCA has been used by (Luettin et al. (1996)), (Potamianos et al. (1998)) and (Sanderson and Paliwal (2004)) to represent mouth movements in speaker and speech recognition systems. The PCA data projection achieves optimal information compression in the sense of minimum square error between the original vector and its reconstruction based on its projection. The achieved dimension reduction serves to reduce the massive image data. Here, however, it serves an even more important purpose, to prevent the intra class covariance matrix, needed at the next stage (LDA), from being singular. This is because there is typically never enough data to compute a reliable estimation of the intra class covariance matrix for a high dimensional dataset, as is the case in a mouth region image (which has approximately 40000 gray values, and there are a couple of hundreds of frames available per class/person, typically). LDA transform maps the feature space to a new space for improved classification, i.e. features that offer a clear separation between the pattern classes. In image pattern classification, it is common that LDA is applied in a cascade following the PCA projection of a single image frames.

Integration of Audio and Visual Information

Feature fusion is used here fuse different information sources with the ultimate goal of achieving superior recognition results. Fusion techniques are divided into three categories: feature fusion, intermediate fusion and decision fusion, (Sanderson and Paliwal (2004)) and (Aleksic and Katsaggelos (2006)).

- *Feature fusion*: Because it occurs early in the information processing chain leading to the decision, feature fusion can intuitively be perceived as the simplest fusion method as it can be implemented by concatenation. Though used in other fields frequently, there are few studies using feature fusion in audio-visual pattern recognition, (Liang et al. (2002)), essentially because of increased dimension and different data rates and types, causing modeling difficulties if carried out in a straight forward manner. There are even fewer studies reporting results on large, publicly available audio-visual databases. Other studies using feature fusion is (Chaudhari et al. (2003)) and (Fox et al. (2007)).
- *Decision fusion*: Some form of recognition is performed separately for each modality and these results are fused at the decision level. When there are more than 2 machine experts, ranked lists can be utilized (Brunelli and Falavigna (1995)), (Kittler et al. (1997)), (Wark et al. (1999)), (Luettin and Thacker (1997)) and (Chibelushi et al. (2002)). This is relevant even in multiple algorithms or multiple classifier decision making strategies too. The latter strategy has been specifically used here, when we identified people and the digits they uttered, where numerous 2-class SVM are combined to obtain a decision on n-class (persons or digit identities) problems. The majority voting and combined voting are commonly utilized techniques in decision fusion. Majority voting refers to that the final decision is made by taking the (common) decision of most sub classifiers. For ranked lists, each sub classifier provides a ranked list that is combined with other classifiers' lists in the final stage. The method requires less computation but can be more complex to grip and implement because some combinations will not work for some users and an automatic selection and combination rules will be needed.
- *Intermediate fusion*: Information from audio and visual streams is processed during the procedure of mapping from feature space into opinion/decision space. In a decision fusion process this mapping for audio and video streams would run in parallel and without influence on each other. In intermediate fusion HMMs are used often to couple and extend these two processing strands ending in a common decision (Liang et al. (2002)), (Aleksic and Paliwal (2002)), (Chaudhari et al. (2003)), (Bengio (2003)) and (Fox et al. (2007)). Complex intermediate fusion schemes promise to take into account for the different reliability of the two streams dynamically, and even for different temporal sampling rates.

Decision fusion can be complex quickly if it is user adaptive. This is because for large number of classes (users), a large amount of training is needed.

Many biometric systems support multiple experts even within one modality as they apply decision fusion. However, with increased number of machine experts, the complexity of the classifier increases because in addition to training individual experts the training of supervisors will be mandatory as the experts will differ in their recognition skills (performance). Accordingly, it is not self evident that decision fusion will yield a more efficient decision making as compared to feature fusion with increased number of independent experts. Feature fusion might have a higher computational entry cost in terms of implementation because modality specific issues need to be tackled e.g. the audio and video feature rates as well as the amount of data are significantly different between audio and video. On the other hand, feature fusion gives a better opportunity to design an effective synergy between the audio and video signals reducing the need for more complex decision making rules later.

Databases

There exist only few databases suitable for recognition systems using audio-visual information. Databases usually vary in the number of speakers, vocabulary size, number of sessions, scenarios, evaluation measures or protocols. The way of collection of databases can influence the methods or scenarios for which they can be useful. For example, the database could have been collected with a specific scenario in mind whereas the real scenario for which a biometric system needs to be developed could be very different from this scenario. This makes the comparison of different visual features and fusion methods, with respect to the overall performance of a biometric system difficult. Here, we present an overview of some of the datasets that are currently publicly available and have been utilized in several published literatures aiming audio-visual biometric systems.

- M2VTS and XM2VTS database: The M2VTS (Multimodal Verification for Teleservices and Security Applications) database consists of audio and video recording of 37 subjects uttering digits in different occasions, (Pigeon and Vandendorpe (1997)). Because it has a small set of different users the work was extended to 295 subjects (Messer et al. (1999)). The resulting XM2VTS (extended M2VTS) database is actually a different database that offers three fixed phrases, two ten digit sequences and one seven word sentence and a side view images of the subjects. All recordings were performed in a studio in four sessions, separated by a lapse of approximately 6 weeks, during a period of five months. The database is intended for researches in areas related to biometric multimodal recognition systems and have been frequently used in the literature. (Teferi and Bigun (2007)) presented recently the DXM2VTS (Damascened XM2VTS) by replacing the background of the speakers in the XM2VTS database with videos of different real scenes. Furthermore, the merged images are offered with appropriate test protocols and different levels of disturbances including motion blur, (translation, rotation, zooming), and noise (e.g. Gaussian and salt/pepper noise) to measure the performance of biometric systems at different scenarios.
- BANCA database: The BANCA (Biometric Access Control for Networked and e-Commerce Applications) database consists of audio and video recordings of 208 subjects recorded by three different scenarios, (Bailly et al. (2003)). The subjects were recorded while they were saying a random 12 digit number, name, address and date of birth. The BANCA database contains four different language recordings. It is aimed for realistic and challenging conditions for real time applications, though there are very few studies which have published results on it.
- AV-TIMIT: The database consists of 223 subjects, (Sanderson (2002)), and its main properties are continuous phonetically balanced speech, multiple speakers, controlled office environment and high resolution video. Speakers were video recorded while reciting sentences from the TIMIT corpus. Each speaker was asked to read 20 to 21 sentences. The first sentence of each round was identical for all speakers and the rest of the sentences were random for each speaker.

Additional datasets, (Potamianos et al. (2003)), are DAVID, containing 100 speakers uttering digits, the alphabet, syllables and phrases, VALID which consists of 106 speakers recording the same sentences as recorded in the XM2VTS database with some additional environment and acoustical noise, and AVICAR (AV speech corpus in a car environment) which consists of 100 subjects uttering isolated digits, isolated letters and phone numbers inside a car.

Our experiments have been performed on XM2VTS because a large number of biometric recognition studies are based on it using standard protocols. A major reason for its popularity is that it is publicly available and that it contains biometric data of a large number of individuals in several modalities across time, allowing for both impostor and client tests.

Performance Measurement

In general, the performance of a biometric recognition system is evaluated by its error rates at various situations. The performance of identification systems is normally reported in terms of identification error, defined as the probability that the correct match of the unknown person's biometric data is correlated to one of the speaker subjects in the dataset. In practice this translates often to give a list of 10 (or any other practicable number) best matches sorted in resemblance order. Such systems could also be equipped with an option to reject to provide a "10 best" list on various valid grounds, e.g. because data quality is too poor, or the likelihood that the queried identity is in the database of clients is below a preset threshold.

For verification systems, two commonly used error measures are the *false acceptance rate* (FAR)—that is an impostor is accepted—and *false rejection rate* (FRR)—where a client is rejected. These error rates are defined by

$$FAR = IM_A / IM$$
$$FRR = CL_R / CL$$

where IM_A and IM denote the number of accepted impostors and the number of impostor claims and CL_R and CL represent the number of rejected clients and the number of client claims, respectively. FAR and FRR curves of a biometric system decrease, respective increase as a function of the threshold (assuming 0 means the identity claim is false, an impostor, and 1 means it is true, a client). Verification systems also present the performance by choosing a threshold where FAR is equal to FRR, called the equal error rate (EER).

In the Lausanne protocol, (Luettin and Maitre (1998)), the system performance is tested at two levels after the training. Although there is a fully functional system at hand at the end of the training, one has yet to set a threshold to make it operational. The evaluation test presents the performance of the system by plotting FAR and FRR curves for all possible thresholds (in practice a discrete set of thresholds) by using images that the recognition system has not seen during the training (evaluation set). The FAR and FRR in all our experiments are obtained on the evaluation set of the Lausanne protocol. However, a system owner is yet to decide at which point (or points) on the ROC curve the system should be operated, and determine the corresponding threshold. In our publications we reported the (correct) verification rate (VR) which is

$$VR = (1 – (FAR + FRR))$$

to represent the successful or correct decision rate.

MOTION ESTIMATION TECHNIQUES

A fundamental problem in image sequence processing is the measurement of optical flow (or image velocity). The aim is to determine (approximate) the 2D motion field from spatiotemporal patterns of image intensity. The computed measurements of the image sequence are used, here, providing a close approximation to the lip motion in the 2D field.

Several methods for computing optical flow have been proposed (Barron et al. (1992)). Here, we will give an overview of the differential method proposed by (Lucas and Kanade (1981)) and the structure tensor based technique proposed by (Bigun et al. (1991)) in addition to our proposed method. Next we present the motion of two image patches that contain fundamentally different patterns.

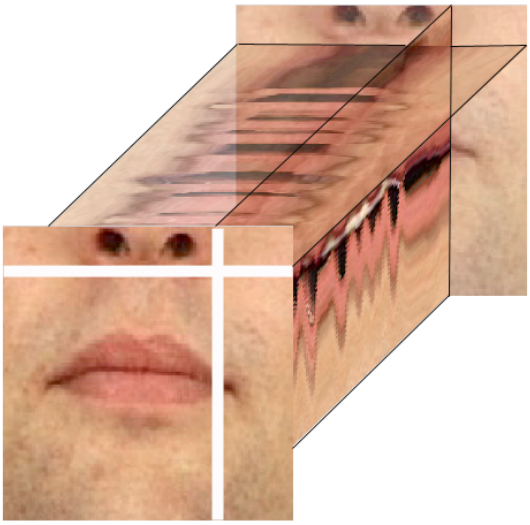
Point and Line Motion

We can study motion in image sequences by making assumptions on the contents of the (local) 2D image patches on the move. Although patch types are many and therefore difficult to enlist, two types, the motion of lines and the motion of points, are particularly important for motion estimation. When an image patch consisting of points translates (the dots move in a group) relative to a fixed camera, the image plane of the camera registers continuously images of the motion which can be stacked to generate the 3D volume. The interest in this motion has been uncertain in image analysis community mainly because it is possible to establish automatically by "walking" along a fixed direction (the direction of the parallel bunch of lines in the figure) in the stack of images. Because every point in the original patch can be tracked without ambiguity in the next frame, it is this type of motion that is used to "track" patches and even real objects in image sequences. The (common) direction of the lines generated by the moving dots represents the velocity of the 2D patch in motion, which is known as the *Brightness Constancy Constraint* (BCC).

However, images are also full of other patches (local images) that do not contain points and they too move. A particularly important class of patches is those that contain lines, edges and other patterns that have a common direction patches that possess linear symmetry, (Bigun and Granlund (1988)). When such a patch translates, this motion generates a tilted plane (or several parallel planes if there are parallel lines in the patch. Here the motion does not generate a bunch of lines that can be utilized to establish correspondence between points belonging two different image frames any more. One can track lines between image frames but not the individual points (of the lines) since it is not possible to observe a difference between a line that translates perpendicular to its direction, and the same line when it translates along its direction in addition to the translation it performs in the perpendicular direction[2]. This non uniqueness (at point level) in tracking is generally not desirable in image analysis applications and is therefore called as the aperture problem with emphasis on "problem". When optical flow is computed patches containing directions (linear symmetries) are typically avoided whereas patches containing points (texture) are promoted. There is a BCC assumption even in this scenario but the brightness constancy is now at the line level.

The question is whether the motion of lines, normally not desirable in image sequence analysis, can be useful for lip-motion quantification. This is significant from resource utilization point of view because the moving patches that contain lines outnumber greatly those that contain dots[3] in lip sequences. Before

Figure 4. The figure illustrates a lip sequence for a speaker uttering digits zero to nine. The vertical and horizontal cross section indicates the existing lip movements.



discussing how to proceed to obtain lip-motion, we outline two methods, (Lucas and Kanade (1981)) and (Bigun et al. (1991)) that represent the distinction between point-motion and line-motion well.

Motion Estimation by Differentials

The motion direction of a contour is ambiguous, because the motion component parallel to the line cannot be inferred based on the visual input. This means that a variety of contours of different orientations moving at different speeds can cause identical responses in a motion sensitive neuron in the visual system.

Differential techniques, (Lucas and Kanade (1981)) and (Horn and Schunck (1981)) compute the optical flow from spatial derivatives of the image intensity and the temporal difference between a pair of frames in an image sequence **Fig. 4**. The approach assumes that the studied patch contains points and that the patch undergoes a translational motion, to be precise the image observed a time instant t later is obtainable from the original patch at $t = 0$ by translation, as follows

$$I(\mathbf{x},t) = I(\mathbf{x} - \mathbf{v}t,0) \tag{1}$$

Here I represents the local image with the spatial vector $\mathbf{x}$, and $\mathbf{v} = (v_x,v_y)^{\mathrm{T}}$ is the velocity to be estimated. A differential expression for the brightness change constraint equation can be obtained if the

mathematical concept of total differential is utilized. It amounts to that the gray value change of the same point, i.e. the total differential, is nil as represented by the following equation

$$\frac{dI}{dt} = 0 \tag{2}$$

This, when the chain rule of multivariable functions is utilized,

$$\frac{dI}{dt} = \frac{dI}{dx}\frac{dx}{dt} + \frac{dI}{dy}\frac{dy}{dt} + \frac{dI}{dt}\frac{dt}{dt} = 0 \tag{3}$$

yields the desired differential expression for BCC.

$$\nabla_s I(\mathbf{x},t)\cdot\mathbf{v} + I_t(\mathbf{x},t) = 0, \tag{4}$$

Here, $I_t(\mathbf{x},t)$ denotes the partial time derivative of $I(\mathbf{x}, t)$ and $\nabla_s I(\mathbf{x},t) = (I_x(\mathbf{x},t), I_y(\mathbf{x},t))^T$ is the spatial gradient. The first component of this equation is a projection of the velocity vector on the gradient. If the local patch is a line (violating the underlying assumption of translating points), the velocity components that are parallel to this line will be orthogonal to the gradient (which is orthogonal to the line in the patch) and will produce zero after the projection. This means that any velocity parallel to the line direction will not be recoverable from equation (4), which is another way of telling that there is an aperture problem.

However, as it stands this equation cannot be solved even if the patch contains only points because there is one equation and two unknowns, (v_x, v_y). To obtain the velocity components, the equation is applied to every point in the patch and new equations are obtained for different points, in practice for all points of the patch. Because the patch pattern consists of dots (and not lines) and all dots move with the same translational velocity, the common velocity components can be obtained in the least squares error sense as below.

$$\mathrm{g} = -\mathrm{Gv} \tag{5}$$

where **v** is unknown

$$G = \begin{pmatrix} \frac{\partial(x_1,y_1,t_0)}{\partial x} & \frac{\partial(x_1,y_1,t_0)}{\partial y} \\ \frac{\partial(x_2,y_2,t_0)}{\partial x} & \frac{\partial(x_2,y_2,t_0)}{\partial y} \\ \vdots & \vdots \\ \frac{\partial(x_N,y_N,t_0)}{\partial x} & \frac{\partial(x_N,y_N,t_0)}{\partial y} \end{pmatrix}, \quad \begin{pmatrix} \frac{\partial(x_1,y_1,t_0)}{\partial t} \\ \frac{\partial(x_2,y_2,t_0)}{\partial t} \\ \vdots \\ \frac{\partial(x_N,y_N,t_0)}{\partial t} \end{pmatrix} \tag{6}$$

The equation (6) contains the first order partial derivatives coming from all points in the observed image patch (N in total) and can be estimated by convolutions efficiently. Suggested by (Lucas and Kanade (1981)), this is a linear regression problem for optical flow estimation. The standard solution of such a system of equation is given by mean square estimate, obtained by multiplying the equation with $\mathbf{G}^T$ and solving the 2x2 system of equations for the unknown $\mathbf{v}$

$$G^T g = - G^T G v \tag{7}$$

For a discrete 2D neighborhood $I(x_k, y_k, t_0)$, a unique solution exists if the matrix

$$\mathbf{S} = \mathbf{G}^T \mathbf{G} = \sum_k (\nabla_{s_k} I) \cdot (\nabla_{s_k}^T I) \tag{8}$$

is invertible where

$$(\nabla_{sk} I) = \begin{pmatrix} \dfrac{\partial I(x_k, y_k, t_0)}{\partial x} \\ \dfrac{\partial I(x_k, y_k, t_0)}{\partial y} \end{pmatrix} \tag{9}$$

However, $\mathbf{S}$ is the structure tensor for the 2D discrete image $I(x_k, y_k, t_0)$,

$$\mathbf{S} = K \begin{pmatrix} \left\langle \dfrac{\partial I(x_k, y_k, t_0)}{\partial x} \dfrac{\partial I(x_k, y_k, t_0)}{\partial x} \right\rangle & \left\langle \dfrac{\partial I(x_k, y_k, t_0)}{\partial x} \dfrac{\partial I(x_k, y_k, t_0)}{\partial y} \right\rangle \\ \left\langle \dfrac{\partial I(x_k, y_k, t_0)}{\partial x} \dfrac{\partial I(x_k, y_k, t_0)}{\partial y} \right\rangle & \left\langle \dfrac{\partial I(x_k, y_k, t_0)}{\partial y} \dfrac{\partial I(x_k, y_k, t_0)}{\partial y} \right\rangle \end{pmatrix} \tag{10}$$

with "< >" representing the average over the pixels (K is the number of pixels) in the 2D neighborhood. If the structure tensor $\mathbf{S}$ has en eigenvalue that equals to zero (singular $\mathbf{S}$) then no unique velocity can be estimated from the image measurements. This is explained by the fact that then the pattern in the patch consists of lines. The tensor can be singular no matter how many points participate into the regression of velocities, because the 2D pattern can consist of long (possibly parallel) lines. Accordingly, this situation represents the aperture problem. In this case the structure tensor is not invertible, which the method in (Lucas and Kanade (1981)) chooses to avoid by not calculating it. Alternatively the optical flow estimations for such patches are down weighted (Horn and Schunck (1981)) since they otherwise would cause severe discontinuities.

Motion Estimation by the 3D Structure Tensor

This method can estimate both the velocity both in the translating points, and the translating lines sce-

narios, as it can provide a measure of confidence as to which type of scenario is most likely prevailing in the investigated patch.

Assume that the local intensity function *f* represents the intensity (gray value) of a local image in a 3D spatiotemporal image (continuously stacked patches) and that the local intensity function *f* consists of parallel planes. This corresponds to parallel planes in 3D which is the same as that the energy is concentrated along an axis through the origin in the 3D Fourier transform of *f*. Thus, the problem of finding a representative velocity for the local image corresponds to finding the inclination angle of the parallel planes, which in turn can be solved by fitting an axis through the origin of the local image's Fourier representation (Bigun (2006)). Fitting an axis is classically performed by the minimization problem, in the total *least square error* (LSE) sense. The solution is obtained by an eigenvalue analysis of the 3x3 matrix, also known as the 3D structure tensor of the local intensity function. This 3x3 tensor can, however, be obtained directly in the spatial domain thanks to the conservation of the scalar product between the spatial and Fourier (frequency) domains. It can be written as follows in the spatial domain

$$J = trace(A)I - A \tag{11}$$

with

$$\mathbf{A} = \begin{pmatrix} \iiint \left(\frac{\partial f}{\partial x}\right)^2 & \iiint \left(\frac{\partial f}{\partial x}\cdot\frac{\partial f}{\partial y}\right) & \iiint \left(\frac{\partial f}{\partial x}\cdot\frac{\partial f}{\partial t}\right) \\ \iiint \left(\frac{\partial f}{\partial x}\cdot\frac{\partial f}{\partial y}\right) & \iiint \left(\frac{\partial f}{\partial y}\right)^2 & \iiint \left(\frac{\partial f}{\partial y}\cdot\frac{\partial f}{\partial t}\right) \\ \iiint \left(\frac{\partial f}{\partial x}\cdot\frac{\partial f}{\partial t}\right) & \iiint \left(\frac{\partial f}{\partial y}\cdot\frac{\partial f}{\partial t}\right) & \iiint \left(\frac{\partial f}{\partial t}\right)^2 \end{pmatrix} = \iiint (\nabla f)(\nabla f)^T$$

where

$$\left(\frac{\partial f}{\partial x}\right),$$

$$\left(\frac{\partial f}{\partial y}\right)$$

and

$$\left(\frac{\partial f}{\partial t}\right)$$

correspond to partial derivatives of the image in *x*, *y* and *t* coordinate directions and *trace*(**A**) is the sum of the diagonal elements of **A**, which also equals to the sum of all eigenvalues of **A**. The matrix **A** can be estimated by a discrete approximation:

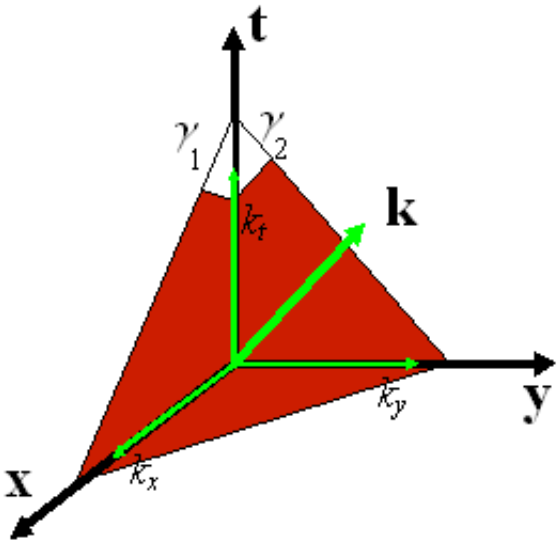
$$\mathbf{A} \cong \sum_{j} (\nabla f_j)(\nabla f_j)^T \tag{12}$$

where (∇f_j) is the gradient at a specific discrete image position *j* wherewith *j* running over all positions (in all 3 directions *x*, *y* and *t* in the three dimensional neighborhood). The least square error corresponds to the least eigenvalue of **J** with its corresponding eigenvector representing the optimal plane fit. Finding the eigenvector corresponding to the least eigenvalue of **J** is the same as finding the eigenvector corresponding to the largest eigenvalue of **A**. By investigating the errors of the fit (the eigenvalues), an approximation of the quality of the fit to the local image can be estimated along with the plane fit, the normal of which encodes the normal velocity in *f*.

The gradient image (∇f_j) can be estimated through convolutions with partial derivative filters of three dimensional Gaussians. After that, the above mentioned outer products and the three dimensional smoothing corresponding to the triple integral equation (12) is carried out.

It turns out that even motion of points case reduces to the same eigenvalue problem as above. The difference is that the multiplicity of the smallest eigenvalue is 1 and this smallest eigenvalue is zero, (close to zero in practice) whereas for motion of lines case the multiplicity of the smallest eigenvalue is 2.

Figure 5. The graph shows the geometry used to derive the 2D velocity vector from the 3D normal vector along with the plane generated by a translating line



However, this method requires multiple image frames although it simultaneously derives the velocity of moving points and lines. Accordingly, the computations can be excessive for applications that only need line motion features. Assuming that line motion is the most relevant motion type in lip images the computations can instead be carried out in 2D subspaces of the 3D spatiotemporal space. This is described next.

Motion Estimation by Line Translation, *Normal Optical Flow*

Assume that $f(x,y,t)$ is generated by a line translated in its normal direction with a certain velocity. The local image containing a moving line in the xy manifold will generate a plane in the (spatiotemporal) xyt space, **Fig. 5**. The normal of the plane, $\mathbf{k} = (k_x, k_y, k_t)^T$ with $\| \mathbf{k} \| = 1$, is directly related to the observable normal velocity. Thus this velocity is encoded by the orientation of the spatiotemporal plane in the xyt space. Let the normal velocity, $\mathbf{v} = (v_x, v_y)^T$, be encoded as $\mathbf{v} = v\mathbf{a}$ with v as the absolute speed and $\mathbf{a}$ as the direction of the velocity which also represents the normal of the line. Being a normal vector, the length of $\mathbf{a}$ is fixed to 1, i.e. $\|\mathbf{a}\| = 1$. Because the local image f is assumed to consist of a moving line, it can then be expressed as

$$g(\mathbf{a}^T\mathbf{s} - vt), \qquad \mathbf{s} = (x, y)^T \tag{13}$$

for some 1D function $g(\tau)$, where $\mathbf{s}$ represents a spatial point in the image plane and t is the time. Defining now $\tilde{\mathbf{k}}$ and $\mathbf{r}$ as

$$\tilde{\mathbf{k}} = (a_x, a_y, -v)^T, \qquad \mathbf{r} = (x, y, t) \tag{14}$$

in equation (13), we have a (linearly symmetric) function f that has iso-curves that are parallel planes i.e.

$$f(x, y, t) = g(\tilde{\mathbf{k}}^T\mathbf{r})$$

Here $\| \tilde{\mathbf{k}} \| \neq 1$ because

$$\sqrt{(a_x^2 + a_y^2)} = 1$$

is required by the definition of $\tilde{\mathbf{k}}$ (equation (14)). Given f, the problem of finding the best $\mathbf{k}$ fitting the hypothesis

$$f(x, y, t) = g(\mathbf{k}^T\mathbf{r}) \text{ with } \| \mathbf{k} \| = 1$$

in the total LSE sense is given by the most significant eigenvector of $\mathbf{A}$. Calling this vector $\mathbf{k}$, and assuming that it is already computed using $\mathbf{A}$, $\tilde{\mathbf{k}}$ is simply obtained by normalizing $\mathbf{k}$ with respect to its first two components as follows

$$\widetilde{\mathbf{k}} = \frac{\mathbf{k}}{\sqrt{\left(k_x^2 + k_y^2\right)}} \tag{15}$$

In agreement with the definition of $\widetilde{\mathbf{k}}$, equation (14), we will have **a** (2D direction of the velocity in the image plane) and *v* (the absolute speed in the image plane) as

$$\mathbf{a} = \left(\frac{k_x}{\sqrt{\left(k_x^2 + k_y^2\right)}}, \frac{k_y}{\sqrt{\left(k_x^2 + k_y^2\right)}} \right)^T \tag{16}$$

$$v = -\frac{k_t}{\sqrt{\left(k_x^2 + k_y^2\right)}} \tag{17}$$

Consequently, the velocity or the *normal optical flow* can be obtained by v**a**

$$\mathbf{v} = v\mathbf{a} = \frac{k_t}{k_x^2 + k_y^2}(k_x, k_y)^T = \frac{1}{\left(\frac{k_x}{k_t}\right)^2 + \left(\frac{k_y}{k_t}\right)^2}\left(\frac{k_x}{k_t}, \frac{k_y}{k_t}\right)^T = (v_x, v_y)^T \tag{18}$$

so that the velocity components are given by

$$v_x = \frac{k_x k_t}{k_x^2 + k_y^2} = -\frac{\left(\frac{k_x}{k_t}\right)}{\left(\frac{k_x}{k_t}\right)^2 + \left(\frac{k_y}{k_t}\right)^2} \tag{19}$$

$$v_y = \frac{k_y k_t}{k_x^2 + k_y^2} = -\frac{\left(\frac{k_y}{k_t}\right)}{\left(\frac{k_x}{k_t}\right)^2 + \left(\frac{k_y}{k_t}\right)^2} \tag{20}$$

As discussed above, **k** can be estimated by the most significant eigenvector of 3D tensor **A**, (Bigun et al. (1991)), if computational resources would not be an issue.

Forming the 3D matrix **A** via triple integrals and solving for its eigenvectors and eigenvalues may be avoided all together if only normal flow is needed for the application at hand. From equations (18)-

(20), the velocity and direction can be estimated by determining the tilts (k_x/k_t) and (k_y/k_t). The tilts can in turn be estimated by local orientation estimation of the intersection of our original motion plane with the *tx* and *ty* planes, (Isaac-Faraj and Bigun (2006)) and (Kollreider et al. (2005)). This 2D orientation estimation can be done by fitting a line to the 2D spectrum in the total least square error sense. This is discussed next.

A local 2D image with ideal local orientation is characterized by the fact that the gray values do not change along one direction. Since the gray values are constant along lines, local orientation is also denoted as linear symmetry orientation. Generally, an image is linearly symmetric if the iso-gray values are represented by parallel hyperplanes. A linearly symmetric 2D image in particular consists of parallel lines in 2D, and has a Fourier transform concentrated along a line through the origin. Detecting linearly symmetric local images is consequently the same as checking the existence of energy concentration along a line in the Fourier domain, which corresponds to the minimization problem of solving the inertia matrix in 2D. By analyzing the local image this time as a 2D image, *f*, the structure tensor can be represented as follows for the *tx* plane:

$$\begin{pmatrix} \iint\left(\frac{\partial f}{\partial t}\right)^2 & \iint\left(\frac{\partial f}{\partial t}\cdot\frac{\partial f}{\partial x}\right)^2 \\ \iint\left(\frac{\partial f}{\partial t}\cdot\frac{\partial f}{\partial x}\right)^2 & \iint\left(\frac{\partial f}{\partial x}\right)^2 \end{pmatrix}$$

Note that this structure tensor has double integrals as opposed to its 3D counter part in equation (11). Eigenvalue analysis in 2D yields a particularly simple form by using complex numbers (Bigun and Granlund (1987))

$$I_{20} = (\lambda_{\max} - \lambda_{\min})e^{i2\varphi} = \iint\left(\frac{\partial f}{\partial t} + i\frac{\partial f}{\partial x}\right)^2 dxdy \tag{21}$$

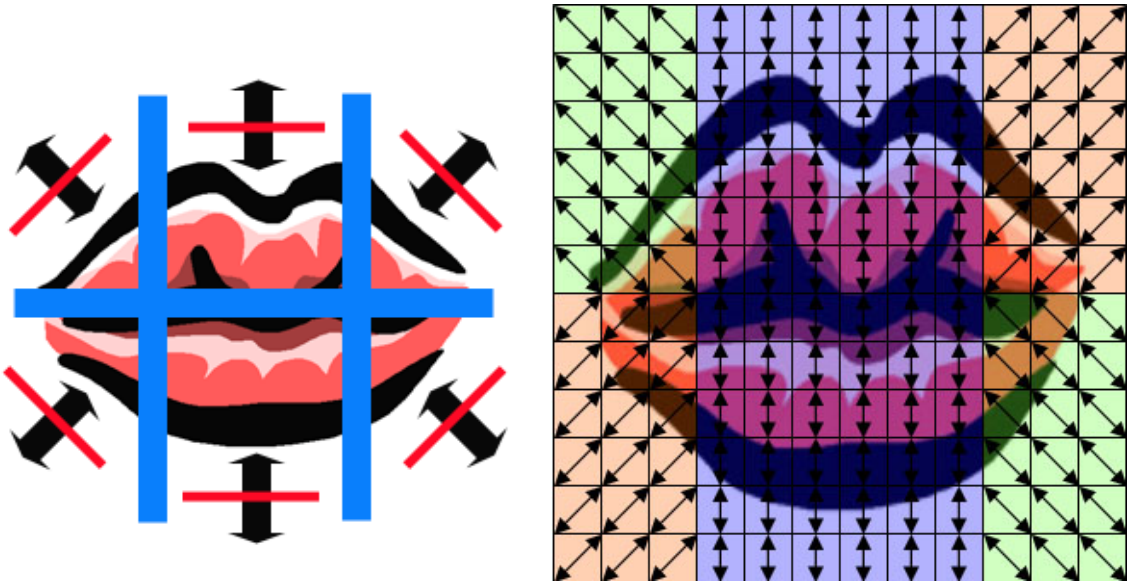
The argument of I_{20}, a complex number in the *t*- and *x*-manifold, represents the double angle of the fitting orientation if linear symmetry exists. In consequence, this provides an approximation of a tilt angle via

$$\frac{k_y}{k_x} = \tan\left(\frac{1}{2}\arg(I_{20})\right) \tag{22}$$

Using this idea both in the *tx* and *ty* manifolds and labeling the corresponding complex moments as I_{20}^{tx}, and I_{20}^{ty} the two tilt estimations and in turn velocity components are obtained as follows:

$$\frac{k_x}{k_t} = \tan\gamma_1 = \tan\left(\frac{1}{2}\arg(I_{20}^{tx})\right) \Rightarrow \widetilde{v}_x = \frac{\tan\gamma_1}{\tan^2\gamma_1 + \tan^2\gamma_2} \tag{23}$$

Figure 6. The figure illustrates the quantization and reduction technique. (Left) Only a limited free degree of the velocity component is allowed, e.g. direction restraints. (Right) The amount of features is reduced by applying 10x10 block wise averaging.



$$\frac{k_y}{k_t} = \tan\gamma_2 = \tan\left(\frac{1}{2}\arg(I_{20}^{ty})\right) \Rightarrow \tilde{v}_y = \frac{\tan\gamma_2}{\tan^2\gamma_1 + \tan^2\gamma_2} \tag{24}$$

The *tx* and *ty* manifolds are shown in **Fig. 5** along with the angles γ_1 and γ_2. The I_{20}^{ty} corresponds to equation (21) but applied to the *ty* manifold.

Feature Quantization and Reduction: Lip Motion Features

We need to perform additional processing to extract lip-motion specific and discriminative information for speech and speaker recognition applications. To this end we proceed as follows.

1. In each pixel of the lip region image we have a motion estimation, given by the horizontal and vertical components of the velocity, (v_x, v_y).

$$v = \| \mathbf{v} \| = \sqrt{(v_x^2 + v_y^2)} \tag{25}$$

2. The image is divided into six regions, which are physically meaningful in the case of lip movements, e.g. mid of upper lip, left of lower lip, etc. The regions are used to quantize the angle of the velocity estimation. The angle α is computed for every pixel and is represented as –1, 0 or 1. These values represent the motion direction relative to the predetermined line directions of each region

$$\alpha = \mathrm{sgn}(\angle \mathbf{v}) = \mathrm{sgn}(\arctan(v_y / v_x)) \tag{26}$$

   We only allow 3 orientations (0˚, 45˚, –45˚) as marked with the 6 solid lines in the 6 regions, (**Fig. 6** – left). The 1D scalars at all pixels take the signs + or – depending on which direction they move relative to their expected spatial orientations (solid lines).
3. Due to the large dimension of the data, using velocities, **v** at each pixel is not a realistic option for applications. We found that direction and speed quantization are significant to reduce the impact of noise on the motion information around the lip area. The quantized speeds are obtained from the data by calculating the mean value in the boxes shown in **Fig. 6** – right as follows,[4]

$$g(l,k) = \sum_{p,q} f(Nl + p, Nk + q), \tag{27}$$

   where

$$f(p,q) = v(p,q)\alpha(p,q) \tag{28}$$

   Here, $(p, q) = 0 \ldots (N - 1)$, and $(l, k) = 0 \ldots (M - 1)$, where $N = 10$ and $M = 12$ represent the window size of the boxes and the number of boxes, respectively.

Implementation

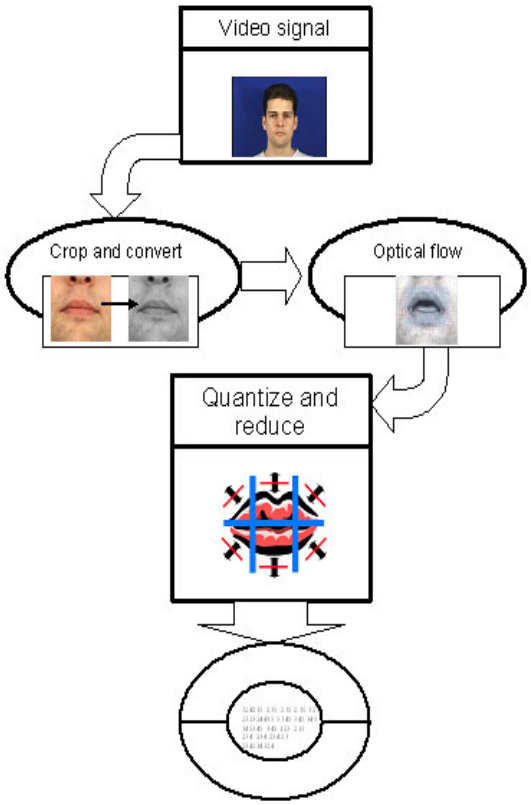
The proposed motion estimation technique is a fast and robust alternative to its more time consuming variant discussed earlier. We will present the implementation steps to determine the optical flow components (v_x, v_y), according to the scheme illustrated in **Fig. 7**. The Gaussian filter derivative, known for its separable filters, are represented here by w_z, where z represents an arbitrary axis x, y and t. Given a spatiotemporal image of four $\{f\}$ frames the following steps are performed:

Step 1. The images of the video sequence are cropped to the lip area with size 128x128 pixels and furthermore converted to a grayscale image by $0.21*R + 0.72*G + 0.07*B$.

Step 2. The sequence is extracted and the orthogonal cross sections $\{tx\}$ and $\{ty\}$ are generated from the permuted xyt space.

Step 3. The $\{tx\}$ and $\{ty\}$ space time manifolds are determined by computation of the orientation according to equation (22) and its analogue for $\{ty\}$. For each $\{tx\}$ plane and $\{ty\}$ plane calculate its gradient by filtering with Gaussian derivative filters w_x, w_y and w_t, yielding a set of images with complex values representing the linear symmetry.

Figure 7. The figure illustrates the processing steps to obtain quantized and dimension reduced lip-motion features



Step 4. Every calculated linear symmetry slice enables us to estimate the normal image velocities in the lip images from equation (23)-(24). In that, only the processing along two planes embedded in the 3D spatiotemporal images is needed.

Step 5. The motion features are quantized and reduced by the mean method.

Using the quantization and reduction the feature vector is represented by 144 dimensions (free variables) instead of the original 128x128x2 = 32768 (free variables) dimensions that describe the motion at all pixels of the mouth region. It is worth noting that the local lip motions are not completely free but must follow physical constraints. It is possible to conclude from related studies and our early published work that the articulation of the lips progresses in a constrained manner during lip movement, i.e. motion in lip image sequences is very symmetrical.

IDENTITY RECOGNITION AND LIVENESS DETECTION BY UTTERED DIGITS

Acoustic Feature Vector

In our experiments, the data stream comes from the audio part of the XM2VTS database and the Mel Frequency Cepstral Coefficient (MFCC) vectors are generated by the Hidden Markov Model Toolkit (HTK) (Young et al. (2000)), where the vectors originate from 25 ms frames (overlapping time periods), streaming out every 10 ms. For each frame, the audio feature vector contains a total of 39 real scalars–12 cepstral coefficients plus normalized log energy, 13 delta (velocity) coefficients, and 13 delta-delta (acceleration) coefficients.

Visual Feature Vector

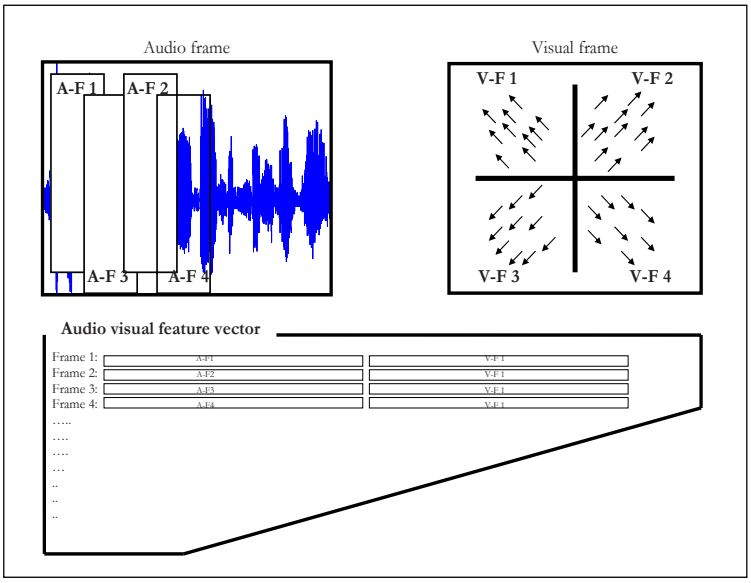
The image sequences used for our experiments are based on the video part of the XM2VTS database. The video was captured with 720x576 pixel frames. In order to reduce the computational complexity, the image frames, before computing the lip-motion features, were automatically cropped to the lip area (128x128) by the technique presented in (Kollreider et al. (2007)). This method suggest quantized angle features ("quangles") designed to reduce the impact of illumination variation. This is achieved by using both the gradient direction and the double angle direction (the angle provided by the 2D structure tensor, (Bigun and Granlund (1988))), and by ignoring the magnitude of the gradient. Boosting techniques are applied in a quantized feature space to detect the mouth. However, by visual inspection, we verified that the cropping functioned as intended to eliminate the impact of localization errors on the errors that can be attributed to the suggested lip-motion features. Furthermore, the color images are transformed to grayscale. After the motion estimation, the features were quantized and the dimension is reduced to represent the relevant mouth movement information automatically.

Feature Fusion: Association and Concatenation

Images come at 4 times slower pace than the audio features. If a classifier is to model the audio and video information at a certain time, somehow the rates of audio and video features must be equalized while keeping as much information as possible from both. For simplicity, this rate equalization problem is called synchronization, which is a term also used by several audio-video compression studies, here. The vectors will be merged to a single vector because we wanted to develop synergetic (joint) modeling of the data as opposed to merging decisions in a late stage of the classification process.

Synchronization is carried out by first extracting the reduced motion features discussed earlier. This amounts to a 144 feature vector. The lip image is divided into 4 sub quarters so that each sub quarter is represented by 36 scalars of the total 144, **Fig. 8**. Second, each one of the obtained sub quarter feature vectors is concatenated with one of the four audio vectors available at the time support of the lip image (the audio features come at 4 times faster rate than image frames). There are different possibilities to do this concatenation given that the four visual vectors can be associated with 4 audio vectors in different combinatorics. However, it turns out that, the particular order does not impact the recognition performance significantly, (Isaac-Faraj and Bigun (2007)). Even using only one of the lip sub quarter motion features in the mentioned concatenation (i.e. repeating it 4 times) will yield almost as good recognition results. Experimental results on this will be presented in Section Experimental setup and tests.

Figure 8. The figure illustrates the joint audio-video information utilizing only 1 sub quarter of the visual information



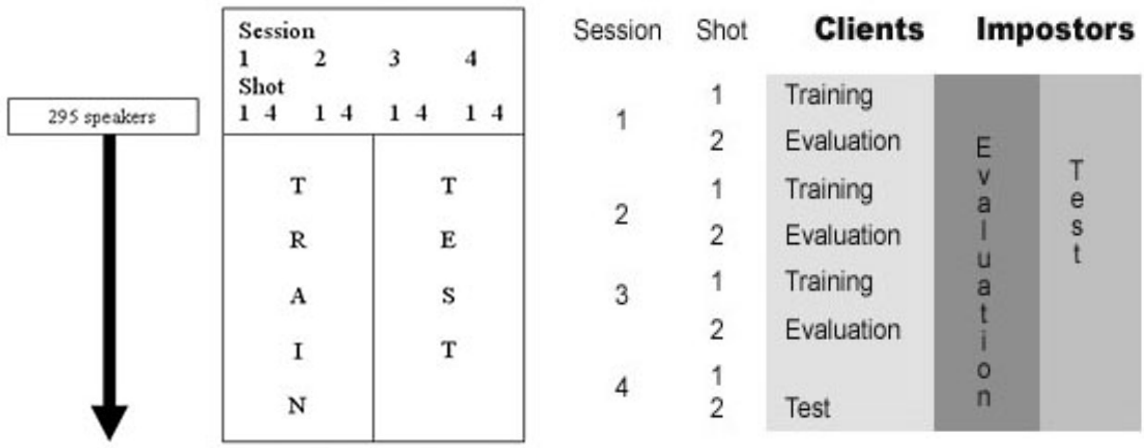
Throughout this work, we assumed that motion in lip images is reasonably symmetric. Because of this and that the motion vectors are both quantized and compactly represented as 1D scalars, we assumed that the order of associating the four visual sub quarters of lips with speech has not a significant impact on the performance. In next section we present experiment where this assumption is tried out. In the experiment we have merged audio visual feature vectors by only using 1 sub quarter of the visual frame and repeating it 4 times, **Fig. 8**.

Experimental Setup and Tests

The XM2VTS Database

In all the experiments, one sequence ("0 1 2 3 4 5 6 7 8 9") was used from the XM2VTS database for every speaker. The database contains 295 speakers (speech with faces) (Messer et al. (1999)). In each

Figure 9. The figure illustrates protocol 1 (left) used for identity recognition and protocol 2 (right) used for digit recognition



session, the subject is asked to pronounce three sentences when recording the video sequence. Because of the different purposes, one extra protocol is presented for digit recognition in addition to the well known *Lausanne protocol* used for identity recognition. The database needed to be segmented for the digit recognition experiment, which was performed nearly 100% automatically. The "continuous" pronunciation of "0 1 2 3 4 5 6 7 8 9" was divided into single digit subsequences 0 to 9 using the methods presented in (Teferi and Bigun (2007)). Furthermore, the segmentation was manually verified and corrected so as to eliminate the impact of database segmentation errors.

Protocol 1: **Fig. 9**(left), this is the Lausanne protocol (Configuration I) defined by the M2VTS consortium standardizing person recognition experiments conducted on XM2VTS. It splits the database into training, evaluation, and test groups (Luettin and Maitre (1998)). This protocol is used in person verification and person identification experiments below. For the XM2VTS database, the Lausanne protocol is commonly used as a standard protocol for speaker identity experiments. However, no protocol is proposed for speech recognition by the M2VTS consortium which conceived the database.

Protocol 2: **Fig. 9**(right), illustrates this protocol wherein 10 words (digits from zero to nine) are spoken by 295 speakers, each with 8 pronunciations. For the training group, sessions 1 and 2 are used and sessions 3 and 4 are used for the test set. The training samples we used were completely disjoint from the test samples. We used a total of 4 pronunciations for training and another 4 for testing.

Classification: Speaker Verification

Speaker verification is carried out in the following steps and is implemented in the HTK software environment (Young et al. (2000)) and (Veeravalli et al. (2005)).

Step 1. Partition the database for training, evaluation, and testing according to protocol 1.
Step 2. Use single left-to-right state constellation using 5 HMM states and a GMM comprised of 3 Gaussians at each state
Step 3. Perform training process by using Baum-Welch re-estimation. In the training, a model for each client is built. Additionally a world model (average impostor), is also built, λ_W. This world model is common for all clients and is built by aggregating the entire training set specified by the Lausanne protocol.
Step 4. Verify using the Viterbi decoding giving a score L which is obtained as the difference between the client probability and the world probabilities $log(L) = log(P(\mathbf{O}|\ \lambda_i) - log(P(\mathbf{O}|\ \lambda_W)\ )$ given a word sequence $\mathbf{O}$.[5] Here, the score L is compared to a threshold T obtained from the FAR and FRR curves.[6] Using the threshold T, the decision L is made according to the rule: if $L > T$ accept the speaker else reject her/him. The reported verification rates are 1 – (FAR+FRR).

Classification: Speaker Identification

The following steps are conducted in our speaker identification system and are implemented by using an SVM:

Step 1. Partition the database for training, evaluation, and testing according to protocol 1.
Step 2. Train the SVM for an utterance so that the classification score, L, is positive for the user and negative for impostors.
  a. Identify the speaker from a group of speakers
    i. We construct a classifier for each person in the group to separate the user from other users in the training data. The training data is defined by protocol 1 (Lausanne Protocol).
    ii. The speaker identity is determined by the classifier that yields the largest score.

Classification: Digit Recognition for Liveness Detection

The following steps are conducted in our digit recognition system for the purpose of liveness detection and are implemented by using an SVM (Chang and Lin (2001)):

Step 1. Partition the database for training, evaluation, and testing according to protocol. Audio-visual feature vectors of dimension 75 (39 audio and 36 video) for each digit utterance were extracted. Each digit had thus several feature vectors coming from the same digit. Furthermore the feature vectors of all speakers of uttering the same digit were given the same digit label to obtain a person independent digit recognizer.
Step 2. We constructed simple SVM classifiers to separate the feature vectors of one digit (75 dimensional each) from those of every other digit pair wise, i.e. we solved a two-class problem 45 times (10 choose 2 combinations). The responses of these classifiers, L_{ij} were binary, i, or j. After the training there were thus a fixed hyperplane associated with each classifier such that one could classify an unknown feature vector (of dimension 75) into one of the two digit labels i or j.
Step 3. The feature vectors of the unknown digit were extracted. We note that the utterance of a digit normally has many feature vectors because the duration of utterances of digits are completely free,

i.e. they vary with the digit, the person, as well as the mood of the person, whereas each feature vector has a fixed time support of 25 ms. For each vector we obtained 45 decisions from the SVM classifiers L_{ij}, obtained via training. These responses were digit labels i or j such that they could be used as a vote for one of the 10 digit labels, "Zero", … , "Nine". A voting was thus carried out involving each feature vector (each casting 45 votes "Zero", … , "Nine"). The digit label receiving most votes was output as the recognized digit label.

Experimental Results and Discussion

The experiments were performed for speaker verification using GMM (states in an HMM setup within HTK), speaker identification using SVM and digit recognition for liveness detection using SVM. The systems were tested using joint audio-visual and single modalities, respectively. The tests used the XM-2VTS with all 295 subjects uttering the sentence 0 to 9. The protocol 2 setup was introduced for digit recognition because the XM2VTS Lausanne protocol is mainly proposed for identity recognition.

Table 1 shows the results utilizing protocol 1 for the experiments. The verification performance is approximately 77% for a speaker verification system based on only visual information. Speaker verification based on a bimodal system gives approximately 98% correct verification, which is better than the single modality system based on the audio or the visual information.

In this experiment the merged audio-visual feature vectors are presented according to **Fig. 10**. The features are put into HMM system (with GMM at each state) for audio-visual speaker verification. The ROC curves in **Fig. 10** illustrate the audio-visual speaker verification performance for the evaluation set. System 1(red) represents the feature fusion technique by associating four audio features with four different sub quarters of the video (as in the experiments above). The system 2 (blue) uses by contrast the motion features of one of the image sub quarters and repeats it for 4 consecutive audio frames during the concatenation. The blue line across the figure represents the EER line, i.e. its intersections with the ROC curves yield the EER of the corresponding systems.

We can see that using the EER threshold computed on the evaluation set, the verification rates (1 – TER, i.e. equation (2.5)) are 98% and 97% for System 1 and System 2, respectively. The results support our hypothesis, that the lip-motion of an individual is highly symmetric. Accordingly, it would be possible to reduce the video computations with a factor of 4, with little degradation of recognition performance, **Fig. 10**. For demonstration purposes, we have chosen to use all the estimated velocities rather than repeating one sub quarter. However, if done in a real system, this extra computation can be

Table 1. The table presents the results for acoustic, visual, and merged bimodal audio-visual speaker verification systems using protocol 1 in a GMM model

Set / System	Evaluation	Test
Audio	96%	94%
Visual	81%	77%
Audio-Visual	99%	98%

viewed as a way to increase the robustness against noise (including asymmetric lighting, imperfect detection of lip area, etc.) as the feature vectors of the same (repeated) sub quarter contain noise that is more dependent on each other than those using estimations from 4 different sub quarters. It can also occur that for certain individuals the prevailing symmetry is less pronounced and can be discriminative information in identity recognition.

We also used SVM classifiers with a Radial Basis Function (RBF) kernel to perform speaker identification using a single word. The reason for using a single word is that SVM has a tendency to become computationally exhaustive for large feature vectors. The performance obtained using bimodal recognition (100%) compares favorably with the classical single modality recognition system based only on the speech signal (92%) or only on the visual signal (80%).

Figure 10. The figure illustrates the ROC curve of verification performance for audio visual speaker verification systems using the evaluation set. System 1 represents the feature fusion technique obtained by association of four audio frames with one visual frame(4 sub quarters) and System 2 represents the feature fusion obtained by repeating one sub quarter of lips, ***Fig. 8****. The straight line represents the threshold for FA=FR.*

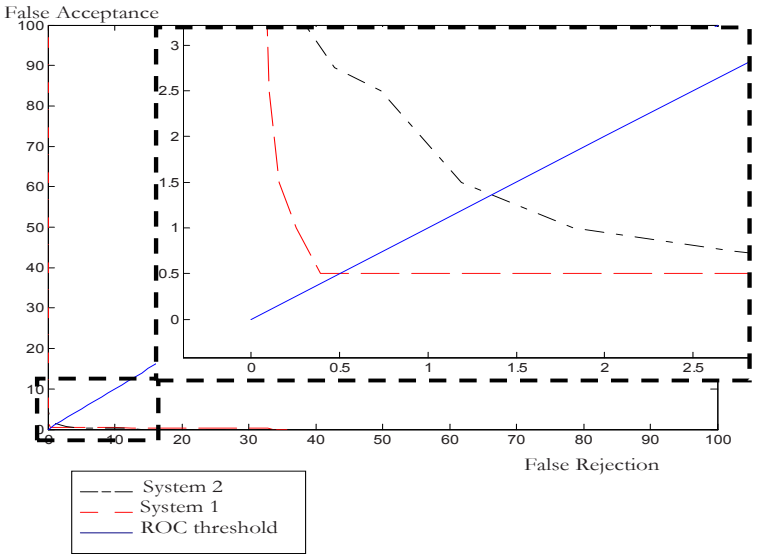


Figure 11. The graph presents multimodal and single modal digit recognition system rates for digits 0 to 9 using protocol 2 in an SVM classifier.

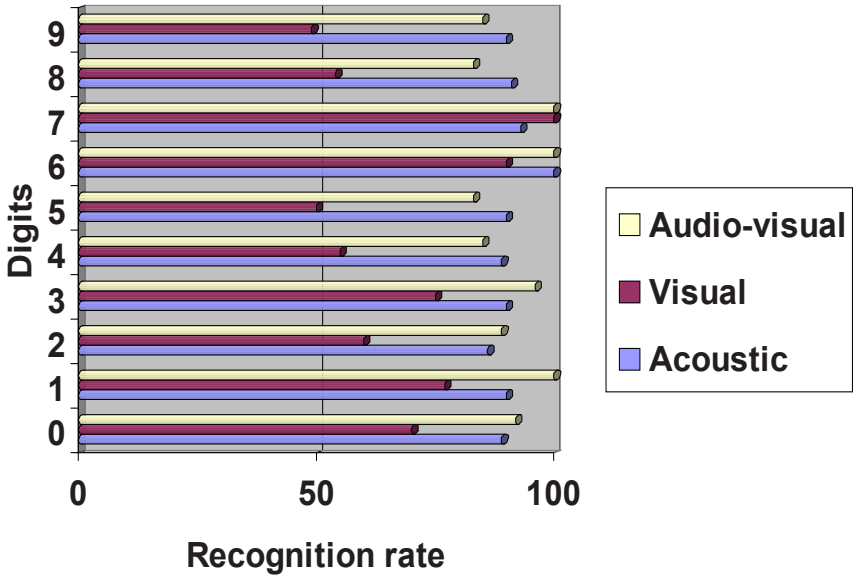


Fig. 11 presents the results for SVM digit recognition for the purpose of liveness detection using protocol 2. Varying between 50-100% for individual digits, the system performed best for digits 1, 6 and 7. The average of the digit recognition over all digits was 68% and 90% for the visual and audio systems in isolation, respectively. Digit recognition using combined audio and video varied between 83% and 100%. The amount of visual information for some of the digits is very little for many people and digit utterances, which is not surprising because the XM2VTS database was collected for identity recognition. As a consequence the speech and the video were recordings of continuous speech without specific emphasis on utterance length or quality. The lack of sufficient visual data for certain digits has negatively influenced the results. However, the uneven results are attributable to the simple SVM classifier we used. The classifier is fast but it does not model the time relationships of the features. A classifier achieving a better and more even digit recognition performance is possible by employing time modeling (at the cost of making the classifier more complex), e.g. HMM or SVM with time modeling. However, it is worth noting that for our main purpose, which is to show that our features are informative in an application targeting digit recognition for liveness assessment, the performance of this classifier is sufficient. This is because we can demand from the user to utter the digits at which the recognizer is good, e.g. any combination of 1, 6 and 7 at any length, when the identity of the person is completely unknown. If the liveness detection is implemented after the identity recognition module, we have then even a possibility to pull out the digits at which the digit recognizer performance is good for the pre-

tended identity and increase the arsenal of digits to be uttered and decrease the length of the sequence to be uttered.

To evaluate our suggested lip features we had to implement audio-visual speech and speaker recognition systems demanding purposive setups of classifier constellations, protocols and databases in addition to implementing compact features, fusing these with appropriate rate conciliation. Each implemented system has a unique difference, in the hope that it will help to evaluate the feature extraction.

Experimental Comparison

We present in this section various audio-visual bi-modal and single modal systems and provide some comparisons related to our work, when possible.

- (Luettin et al. (1996)) developed a visual only speaker identification system using only the lip contour information by extracting *model* and *pixel* based features. These features were extracted by calculating the lip contours and then shape deformations of the contours were modeled temporally by a classifier (HMM). They used the Tulips database, consisting of 12 speakers evaluated by the identification on the speaker models (48 models for each speaker) of the spoken word. The identification system based on HMMs, achieved approximately 73%, 90% and 92% recognition rates when using *shape* based, *pixel* based and joint features. This experiment was extended by (Jourlin et al. (1997)) using the M2VTS database consisting of 37 speakers, utilizing audio-visual information in an identity verification system. The acoustic features were based on the Linear Prediction cepstral coefficients with first and second order derivatives. The visual feature representation was based on the shape and intensity information according to (Luettin et al. (1996)) technique. They utilized HMMs to perform audio only, visual only and audio-visual experiments. The audio-visual score is computed as a weighted sum from the audio and visual classifier. They achieved approximately 97%, approximately72% and approximately 100% verification for audio only, visual only and audio-visual information. The discussed system is comparable to our system except that i) the video features modeled by the classifier were intra frame, and ii) it is a decision fusion system. Comparing the video only experimental results confirm that our features perform better (approximately 6 percentage points) and yet they are complementary because our video features are inter frame based. Their decision fusion has improved the recognition performance by 2 percentage points over the best performing expert (audio) in identification mode and 3% in the verification mode. Our feature fusion has improved the recognition performance by 8 percentage points and 4 percentage points over the best expert (audio) in the (significantly larger) tests corresponding to identification and verification. This indicates that even our feature fusion contributes to performance improvement.
- (Wark and Sridharan (1998)) and (Wark et al. (1999)) presented a system using multi-stream HMMs to develop audio-visual speaker verification and identification systems tested on the M2VTS database. By utilizing decision fusion on the acoustic (based on MFCC) and visual information (based on lip contours and applying PCA and LDA on them), they outperformed the system using only acoustic information. The verification experiments were performed using GMMs on the M2VTS database.
- (Fox et al. (2007)) developed an audio-video person identification system using HMMs as classifier. The experimental tests were performed on the XM2VTS database, using the experts consisting of

acoustic information, dynamic mouth information and static face information. The audio features used are MFCC and their first derivatives. The visual mouth features were derived from pixels to represent the visual information based on the discrete cosine transform. The face features were derived by the PCA technique using the FaceIt software. The 3 experts were fused by a cascade of fusion where expert 1 and 2, decisions were merged in parallel with expert 2 and 3. The resulting two decisions were then merged to yield the final decision. The results highlight the complementary nature of the mouth and face experts under clean and noisy tests.

- (Dieckmann et al. (1997)) and (Frischholz and Dieckmann (2000)) developed a system using three experts acoustic, facial information and lip movements. The acoustic information is based on cepstral coefficients, and facial information is derived to be invariant to rotation and illumination. The optical flow of the lip movement is determined using the traditional method by Horn and Schunk. These three experts are combined by opinion threshold to perform person identification. The experiments were performed on a staff recording of 66 members, achieving best performance approximately 93% when all three modalities were used.
- (Nefian et al. (2002)) demonstrated accurate improvements of speech recognition using audio-visual information. The extracted visual features based on Discrete cosine transform (DCT) and then LDA are combined with acoustic features (MFCC). These features are combined in an intermediate fusion sense, using coupled HMM. The tests were extended by (Liang et al. (2002)), using the XM2VTS database for speech recognition. An extension of visual extraction was also presented based on PCA and LDA and DCT and LDA. They performed in clean SNR environment approximately 99% correct digit recognition using XM2VTS database.
- (Dupont and Luettin (2000)) present a speech recognition system based on bi-modal (audio-visual) information. The visual information is extracted according to (Luettin et al. (1996)) technique by shape and intensity information that generates models for a specific person. The audio information is based on perceptual linear predictive coefficients plus the first derivative and the energy. The features are combined in an intermediate fusion sense, by multi-stream HMMs which were possible by up sampling the visual features, performing approximately 99% correct word recognition.

Other related work using static information are (Brunelli and Falavigna (1995)), (Ben-Yacoub et al. (1999))}, (Sanderson and Paliwal (2004)) and (Hazen et al. (2003)). They developed speaker recognition systems based on audio-visual static information. The visual information was based on static images from the face in combination or without combination of acoustic information. Vector quantization, HMM, GMM and SVM were used to perform different identification and verification tests. In all of these reports decision fusion are used to reconcile the individual decisions.

Applying SVMs to speech recognition can perform better than HMMs (Wan and Carmichael (2005)) in some cases by the use of an appropriate kernel function that can encode temporal information. An SVM will be more accurate than HMMs if the quantity of training data is limited, (Wan and Carmichael (2005)). The latter work exploited the fact that SVMs generalize well to sparse datasets and SVMs were applied on isolated word identification tasks. However, results on small vocabulary tests with sufficient data the accuracy of HMMs and SVMs will asymptote. In this case the HMM is favored because it is more efficient in terms of speed (Wan and Carmichael (2005)). In our case we exploited SVM classifier to limit the scope of the study while obtaining a quick indication on the usefulness of the features, without introducing time variation models for features vectors.

CONCLUSION

Biometric recognition is a popular subject in today's research and has been shown to be an important tool for identity establishment. Using visual information as an adjunct to speech information improves accuracy for not only identity recognition but also speech recognition.

In this chapter, we have described lip-motion features for dynamic lip image sequences to be used in different recognition systems. The technique exploits information from a set of 2D space time signals in the 3D space time image that yields the normal of an optimal motion plane and allows the estimation of velocities. The visual lip features are extracted without iterative algorithms or assuming successful lip contour tracking, which is a computationally efficient alternative to the available lip dynamics estimations. The experimental tests were performed on the XM2VTS database. The database, representing hundreds of thousands of images and hours of speech, was segmented with respect to digit boundaries automatically and verified manually to be able to test the digit recognition systems for the purpose of liveness detection. The addition of visual data to the systems confirms that it is possible to do feature level fusion in such massive data and obtain benefits for biometric recognition systems.

The experimental performance of the proposed biometric systems, yielding approximately 80% video only identification and 100% audio visual identification (for the word 7) of person identities, supports the conclusion that the proposed lip-motion features contain significant information for person authentication. Furthermore, the average digit recognition rate is approximately 70% using only visual information, is suitable for liveness detection system using simple classifier. Our technique for early audio-video feature integration results in an improved speaker verification performance (approximately 98%) and speaker identification performance (approximately 100%), on top of the already high verification rate achievable by speech only.

The results of the digit recognition system are different for the 10 digits. Our examination of the results indicates that once the visual feature extraction is performed on a sufficient amount of visual speech data, the available modeling for recognition tasks is highly successful. Accordingly, the validation in the digit recognition with respect to digits can be explained with the lack of visual data. One obvious consequence is to use lip reading word selectively or by using weights in future designs.

In all cases, no attempt was made to improve the recognition performance by optimizing the classifier. We used what was available to us and in largest diversity of classification constellation, as a significant goal has been to show that the same basic features contained sufficiently rich information for purposes of identity as well as message recognition, regardless what classification method has been used.

Discussion

The system proposed by (Liang et al. (2002)) showing experimental results on XM2VTS database, can not be studied in conjunction with our results in depth primarily because the experimental details of their tests were not mentioned in the publication. The classification/fusion technique they used is an advanced HMM permitting synchronization of the different strands (audio and video) of the data. However, temporal up sampling of video is used to achieve synchronization between the two sampling rates of the audio and video. Their experimental tests report word error rates for a single word/number i.e. 0123456789, since this is the only word that was tested. Because the temporal segmentation of the XM2VTS audio-video was not undertaken in the study, it is not clear how the performance will be affected if the digits are uttered in a different order. Part of this criticism is valid even to our system

despite segmentation because a digit's utterance is influenced by the pre and post digit utterances in the continuous speech. However, we argue that this influence is more difficult to "abuse" by a simple classifier (our system) as compared to giving the full content of the pre and post digits to an advanced classifier, (the system of (Liang et al. (2002))). Even though a digit sequence is recognized from always the same 0-9 sequence uttered by different and same speakers at different times, no information on where in the uttered digit sequence the recognized digits are to be found is available in the reported experiments. Accordingly, with which digits there is greatest confusion can not be evaluated.

The work of (Jourlin et al. (1997)) presented a speaker verification system using features originating from the work of (Luettin et al. (1996)). The main difference of this feature set and our features is that their model based features demand more from the pre processing step which both need. Whereas we need initially an approximately correct localization of the mouth region only, they would additional need a correct detection of the lip boundaries. Our results (undertaken on the XM2VTS database which is 9 times larger than what was available to them at the time) suggest that one can obtain the same descriptive information even without precise boundary tracking while achieving as good as, or better recognition performance. Furthermore, we think that using a crude pre processing step has a significant importance for robustness since non satisfaction of higher pre processing demands will manifest a higher risk of system failure in practice.

We reached favorable results for person recognition by feature fusion strategy. In the future one could quantify how much feature fusion has brought as compared to decision fusion. This would require to implement 3 classifiers (1 audio, 1 audio, and 1 for decision fusion) with their corresponding training. We refrained from doing this, because i) the qualitative comparisons with the studies of (Luettin et al. (1996)) and (Jourlin et al. (1997)) indicated that there was a gain in feature fusion, ii) this would be a sub optimal solution from computational, implementation, and maintenance view point (3 classifiers must be trained and maintained as opposed to 1), and iii) we had to limit the scope of our study.

REFERENCES

Roth, K., & Neidig, J. (2007). Title of chapter. In C. Coulson (Ed.), *Title of book that the chapter appears in* (pp. 19-29). Hershey, PA: Information Science Reference.

Aleksic, P., & Katsaggelos, A. (2006). Audio-visual biometrics. *Proceedings of the IEEE 94(11)*, (pp. 2025-2044).

Aleksic, P., Williams, J., Wu, Z., & Katsaggelos, A., (2002). Audio-visual speech recognition using mpeg-4 compliant visual features. *EURASIP J. Appl. Signal Process*, 2002 (1), 1213-1227.

Bailly, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariethoz, J., Matas, J., Messer, K., Popovici, V., Poree, F., Ruiz, B., & Thiran, J. P. (2003). *The banca database and evaluation protocol. In: AVBPA*. (pp. 625-638).

Barron, J., Fleet, D., Beauchemin, S., & Burkitt, T. (1992). Performance of optical ow techniques. *IEEE Comp. S. conference on Computer Vision and Pattern Recognition*, (pp. 236-242).

Ben-Yacoub, S., Abdeljaoued, Y., & Mayoraz, E. (1999). Fusion of face and speech data for person identity verification. *IEEE Trans. on Neural Networks, 10*(5), 1065-1074.

Bengio, S. (2003). Multimodal authentication using asynchronous HMMs. *In: AVBPA03. Springer Berlin Heidelberg*, (pp. 770-777).

Bigun, E., Bigun, J., Duc, B., & Fischer, S., (1997a). Expert conciliation for multi modal person authentication systems by bayesian statistics. In J. Bigun, G. Chollet, and G. Borgefors, (Eds.), *Audio and Video based Person Authentication - AVBPA97 1206*, (pp. 291-300).

Bigun, J. (2006). *Vision with Direction*. Halmstad, Springer, Heidelberg.

Bigun, J., Duc, B., Fischer, S., Makarov, A., & Smeraldi, F. (1997b). Multi modal person authentication. *In H. Wechsler et al., & editor, Nato-Asi advanced study on face recognition*, (pp. 26-50).

Bigun, J., & Granlund, G., (1987). Optimal orientation detection of linear symmetry. *In First International Conference on Computer Vision, ICCV. IEEE Computer Society*, (pp. 433-438).

Bigun, J., Granlund, G., & Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis of optical flow. *IEEE Trans. Pattern Analysis and Machine Intelligence, 13*(8), 775-790.

Brunelli, K., & Falavigna, D., (1995). Person identification using multiple cues. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 17*(10), 955-966.

Chan, M., (2001). Hmm-based audio-visual speech recognition integrating geometric and appearance-based visual features. *IEEE Fourth Workshop on Multimedia Signal Processing*, (pp. 9-14).

Chang, C., & Lin, C., (2001). *Libsvm-a library for support vector machines. software available at www. csie.ntu.edu.tw/ cjlin/libsvm.*

Chaudhari, U., & Ramaswamy, G., & Potamianos, G., & Neti, C., (2003). Information fusion and decision cascading for audio-visual speaker recognition based on time-varying stream reliability prediction. *International Conference on Multimedia and Expo, 2003. ICME '03.* (pp. 9-12).

Chen, T., (2001). Audiovisual speech processing. *IEEE Signal Processing Magazine, 18*(1), 9-21.

Chibelushi, C., Deravi, F., & Mason, J. (2002). A review of speech-based bimodal recognition. *IEEE Trans. on Multimedia, 4*(1), 23-37.

Chiou, G., & Hwang, J.-N., (1997). Lipreading from color video. *IEEE Transactions on Image Processing, 6*(8), 1192-1195.

Dieckmann, U., Plankensteiner, P., & Wagner, T. (1997). Sesam: A biometric person identification system using sensor fusion. *Pattern Recognition Letters, 18*(9), 827-833.

Duc, B., Fischer, S., & Bigun, J. (1997). Face authentication with sparse grid gabor information. *IEEE International Conference Acoustics, Speech, and Signal Processing, 4*(21), 3053-3056.

Dupont, S., & Luettin, J. (2000). Audio-visual speech modelling for continuous speech recognition. *IEEE Trans. on Multimedia, 2*(3), 141-151.

Faraj, M., & Bigun, J. (2007). Audio-visual person authentication using lipmotion from orientation maps. *Pattern Recognition Letters - Advances on Pattern recognition for speech and audio processing, 28*(11), 1368-1382.

Faraj, M. I., & Bigun, J. (2006). Person verification by lip-motion. *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW),* (pp. 37-45).

Fox, N., & Gross, R., & Cohn, J., & Reilly, R. (2007). Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts. *IEEE Transactions on Multimedia, 9*(4), 701-714.

Frischholz, R., & Dieckmann, U. (2000). Biold: a multimodal biometric identification system. *IEEE Trans. On Computer, 33*(2), 64-68.

Hazen, T. J., & Weinstein, E., & Kabir, R., & Park, A., & Heisele, B. (2003). Multimodal face and speaker identification on a handheld device. *In Proc. Workshop Multimodal User Authentication,* (pp. 113-120).

Horn, B., & Schunck, B. (1981). Determining optical flow. *The journal of Artificial Intelligence, 17*(1), 185-203.

Iyengar, G., & Neti, C. (2001). Detection of faces under shadows and lighting variations. *IEEE FourthWorkshop on Multimedia Signal Processing*, (pp. 15-20).

Jourlin, P., Luettin, J., Genoud, D., & Wassner, H. (1997). Acoustic-labial speaker verification. *Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 1206*, (pp. 319-326).

Kittler, J., Li, Y., Matas, J., & Sanchez, M. (1997). Combining evidence in multimodal personal identity recognition systems. *Proceedings of the First 48 International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 1206*, (pp. 327-334).

Kollreider, K., Fronthaler, H., & Bigun, J. (2005). Evaluating liveness by face images and the structure tensor. *In AutoID 2005: FourthWorkshop on Automatic Identification Advanced Technologies - IEEE Computer Society*, (pp. 75-80).

Kollreider, K., H. Fronthaler, M. I. F., & Bigun, J. (2007). Real-time face detection and motion analysis with application in liveness assessment. *IEEE Trans on Information Forensics and Security, 2*(3), 548-558.

Liang, L., Liu, X., Zhao, Y., Pi, X., & Nefian, A. (2002). Speaker independent audio-visual continuous speech recognition. *IEEE International Conference on Multimedia and Expo, 2002. ICME '02. Proceedings*, (pp. 26-29).

Lucas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence*, (pp. 674-679).

Luettin, J. (1997). *Visual speech and speaker recognition*. Unpublished doctoral dissertation, University of Shefield, U.K.

Luettin, J., & Maitre, G. (1998). *Evaluation protocol for the extended m2vts database (xm2vtsdb).* In: IDIAP Communication 98-054 Technical report R R-21.

Luettin, J., & Thacker, N. (1997). Speechreading using probabilistic models. *Computer Vision and Image Understanding*, *65*(2), 163-178.

Luettin, J., Thacker, N., & Beet, S. (1996). Speaker identification by lipreading. *Proceedings of the 4th International Conference on Spoken Language Processing ICSLP 96*, (pp. 62-65).

Mase, K., & Pentland, A. (1991). Automatic lip-reading by opticalflow analysis. *Systems and Computers in Japan, 22*(6), 67-76.

Matthews, I., Potamianos, G., Neti, C., & Luettin, J. (2001). A comparison of model and transform-based visual features for audio-visual lvcsr. *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001*, (pp. 825-828).

Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. (1999). XM2VTSDB: The extended m2vts database. *In: Audio and Video based Person Authentication- AVBPA99*. (pp. 72-77).

Nakamura, S. (2001). Fusion of audio-visual information for integrated speech processing. *Proceedings Third International Conference on Audio- and Video-Based Biometric Person Authentication: AVBPA 2001 2091*, (pp. 127-149).

Nefian, A., Liang, L., Pi, X., Liu, X., & Murphy, K. (2002). Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP J. Appl. Signal Process*, (1), 1274-1288.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Mashari, J. S. A., & Zhou, J. (2000). *Audio-visual speech recognition*. Final Workshop 2000 Report. Baltimore, MD: Center for Language and Speech Processing. The Johns Hopkins University.

Ortega-Garcia, J., Bigun, J., Reynolds, D., & Gonzalez-Rodriguez, J. (2004). Authentication gets personal with biometrics. *IEEE Signal Processing Magazine*, (pp. 50-62).

Petajan, E. (1984). Automatic lipreading to enhance speech recognition. *Global Telecommunications Conference*. (pp. 265-272).

Pigeon, S., & Vandendorpe, L. (1997). The m2vts multimodal face database (release 1.00). *In: AVBPA '97: Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication*. (pp. 403-409).

Potamianos, G., Graf, H., & Cosatto, E. (1998). An image transform approach for hmm based automatic lipreading. . *Proceedings International Conference on Image Processing, 1998. ICIP 9*. (pp. 173-177).

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE,* 91(9), 1306-1326.

Sanderson, C. (2002). The vidTIMIT database (IDIAP communication). *IDIAP Communication 02-06, Martigny, Switzerland*.

Sanderson, C., & Paliwal, K. (2004). Identity verification using speech and face information. *Digital Signal Processing, 14*(5), 449-480.

Tang, X., & Li, X. (2001). Fusion of audio-visual information integrated speech processing. *Third International Conference on Audio- and Video-Based Biometric Person Authentication AVBPA2001, LNCS 2091*. (pp. 127-143).

Teferi, D., & Bigun, J. (2007). Damascening video databases for evaluation of face tracking and recognition - the dxm2vts database. *Pattern Recognition Letters, 28*(15), 2143-2156.

Teferi, D., Faraj, M., & Bigun, J. (2007). Text driven face-video synthesis using gmm and spatial correlation. *The 15th Scandinavian Conference on Image Analysis (SCIA 2007) LNCS 4522*, (pp. 572-580).

Tistarelli, M., & Grosso, E. (2000). Active vision-based face authentication. *IEEE International Conference on Multimedia and Expo, ICME 2001*, *4*(18), 299-314.

Veeravalli, A., Pan, W., Adhami, R., & Cox, P. (2005). *A tutorial on using hidden markov models for phoneme recognition*. Proceedings of the Thirty- Seventh Southeastern Symposium on System Theory, SSST 2005.

Wan, V., & Carmichael, J. (2005). Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data. *Interspeech'2005 – Eurospeech*. (pp. 3321-3324).

Wark, T., & Sridharan, S. (1998). A syntactic approach to automatic lip feature extraction for speaker identification. *IEEE International Conference on Acoustics, Speech and Signal Processing*. (pp. 3693-3696).

Wark, T., Sridharan, S., & Chandran, V. (1999). Robust speaker verification via fusion of speech and lip modalities. *IEEE International Conference on Acoustics, Speech and Signal Processing 1999. ICASSP 99*. (pp. 3061-3064).

Yamamoto, E., Nakamura, S., & Shikano, K. (1998). Lip movement synthesis from speech based on hidden markov models. *Journal of Speech Communication, 26*(1), 105-115.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2000). *The htk book (for htk version 3.0) Http://htk.eng.cam.ac.uk/docs/docs.shtml.*

Zhang, X., Broun, C., Mersereau, R., & Clements, M. (2002). Automatic speechreading with applications to human-computer interfaces. *EURASIP Journal on Applied Signal Processing, 2002*(11), 1128-1247.

ENDNOTES

1. Several phonemes can correspond to the same visual configuration. In fact, in most cases, *visemes* are not uniquely associated with a single phoneme.
2. We assume that the direction of a line is either of the two normals of the line
3. It is worth noting that there is a patch type whose motion can not be observed at all, the patches consisting of a constant gray value.
4. 4 pixels width boundary are removed in the lip region.
5. The output from the Viterbi decoder is logarithmic, and we used this for convenience.
6. According to Lausanne protocol, the evaluation set is selected to produce client and impostor access scores, thereby to produce FAR and FRR curves. From these certain operation points (thresholds) are selected to be used latter on as thresholds on the test set for recognition.