DISS. ETH NO. 22867

Towards Robust Audio-Visual Speech Recognition

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
TOFIGH NAGHIBI
MSc, Sharif University of Technology, Iran
born June 8, 1983
citizen of Iran

accepted on the recommendation of
Prof. Dr. Lothar Thiele, examiner
Prof. Dr. Gerhard Rigoll, co-examiner
Dr. Beat Pfister, co-examiner

2015

Abstract

Human speech production is a multimodal process, by nature. While usually the acoustic speech signals constitute the primary cue in human speech perception, visual signals can also substantially contribute, particularly in noisy environments. Research in automatic audio-visual speech recognition is motivated by the expectation that automatic speech recognition systems can also exploit the multimodal nature of speech. This thesis aims to explore the main challenges faced in realization and development of robust audio-visual automatic speech recognition (AV-ASR), i.e., developing (I) a high-accuracy speaker-independent lipreading system and (II) an optimal audio-visual fusion strategy.

Extracting a set of informative visual features is the first step to build an accurate visual speech recognizer. In this thesis, various visual feature extraction methods are employed. Some of these methods, however, encode visual information into very high-dimensional feature vectors. In order to reduce the computational complexity and the risk of overfitting, a new feature selection algorithm is introduced that selects a subset of informative visual features from the high-dimensional feature vector space. This feature selection algorithm considers mutual information between features and class labels (phonemes) to be the criterion and employs a semi-definite programming based search strategy for subset selection. The performance of the feature selection algorithm is analyzed and it is shown that the difference between the score of the selected feature subset and that of the optimal feature subset is bounded. That is, it guarantees that the score of the selected features is close to that of the optimal solution.

To achieve a speaker-independent visual speech recognizer, this thesis proposes to employ a pool of scale-invariant feature transform (SIFT) coefficients extracted from multiple color spaces. The ensemble of decision

tree classifiers trained with these features yields a high level of robustness against inter-speaker and illumination variations. While for voice activity detection two-dimensional SIFT features are used to build a statistical model, for lipreading we use three-dimensional SIFT features that can capture the time dynamics of video data. It is shown that using three-dimensional SIFT features gives a substantial recognition accuracy improvement in comparison with conventional visual features. The proposed AV-ASR achieves 70.5% utterance classification accuracy which is the highest accuracy reported for the Oulu dataset, in the speaker-independent setting.

While many boosting algorithms have been developed to train an ensemble of classifiers, most of them cannot be naturally generalized to the multiclass classification setting, which is a requirement in our application. In this thesis, we propose a framework to design boosting algorithms. This framework has the advantage that it can be naturally generalized to the multiclass setting. Several properties such as convergence rates and generalization error of this framework are analyzed. Moreover, multiple practically and theoretically interesting algorithms such as SparseBoost are derived. We show that the SparseBoost algorithm only uses a percentage of training samples at each training round (about half of the samples) while still converging to the optimal hypothesis in the sense of probably approximately correct (PAC) learning.

A common practice to fuse audio and visual information is to assign a reliability weight to each modality. It is shown in this thesis that a more suitable criterion to estimate the reliability weights is to maximize the area under a receiver operating characteristic curve (AUC) rather than frequently used criteria such as the recognition accuracy. Moreover, here we estimate a reliability weight for each feature. This generalizes the (conventional) two-dimensional stream weight estimation problem to a fairly high-dimensional problem. In order to efficiently estimate the reliability weights, we use a smoothed AUC function and adopt a variant of the projected gradient descent algorithm to maximize the AUC criterion in an online manner.

Audio-visual voice activity detection (AV-VAD) is an important prerequisite in many audio-visual applications. We propose a robust audio-visual voice activity detector which can be trained in a semi-supervised manner. This interesting property can be achieved by noting the fact that both audio and visual signals represent the same underlying event, namely speech production. In this approach, training data is labeled by iteratively training audio- and visual-based speech detectors and re-labeling the data in order to use it in the next round. The labeled data from the last iteration is then used to train

the final audio-visual voice activity detector. The proposed AV-VAD algorithm results in almost 96% frame-based detection rate (visual-based VAD yields 78% detection rate) on the GRID dataset in the speaker-independent setting.

Kurzfassung

Die Erzeugung menschlicher Sprache ist naturgemäss ein multimodaler Prozess, der akustische und visuelle Information erzeugt. Während sich die menschliche Sprachwahrnehmung üblicherweise primär auf die akustische Information stützt, können visuelle Signale ebenfalls substanziell beitragen, insbesondere in einer lauten Umgebung. Forschung in automatischer audio-visueller Spracherkennung wird motiviert durch die Erwartung, dass Spracherkennungssysteme die multimodale Natur der Sprache auch ausnutzen können. Ziel dieser Arbeit ist, die Hauptherausforderungen zu erkunden, welchen bei der Realisation robuster audio-visueller Spracherkennungssysteme (audio-visual automatic speech recognition, AV-ASR) begegnet wird, d.h. die Entwicklung (I) eines hochpräzisen, sprecherunabhängigen Lippenlesesystems und (II) einer optimalen audio-visuellen Fusionsstrategie.

Die Extraktion informativer visueller Merkmale ist der erste Schritt beim Design einer exakten visuellen Spracherkennung. In dieser Arbeit werden diverse Methoden zur Extraktion visueller Merkmale angewendet. Für einige dieser Methoden ist jedoch die Menge der resultierenden visuellen Merkmale sehr gross. Zur Reduktion des Rechenaufwandes und zur Vermeidung von Überanpassung wird ein neuer Algorithmus zur Auswahl von Merkmalen eingeführt, welcher eine Untermenge aus der sehr grossen Menge visueller Merkmale auswählt. Dieser Merkmalsauswahl-Algorithmus berücksichtigt die Transinformation (mutual information) zwischen Merkmalen und Klassennamen (Phoneme) als Kriterium und wendet zur Auswahl der Untermengen eine Suchstrategie an, die auf semidefiniter Programmierung basiert. Die Leistung des Merkmalsauswahl-Algorithmus wird analysiert und es wird gezeigt, dass die Differenz zwischen der Bewertung der ausgewählten Untermenge und der optimalen Merkmalsmenge begrenzt ist. Das heisst, es ist

garantiert, dass sich die Bewertung der gewählten Merkmale nahe bei derjenigen der optimalen Lösung befindet.

Zur Verwirklichung einer sprecherunabhängigen visuellen Spracherkennung schlägt diese Arbeit die Verwendung einer Vielzahl skalierungsinvarianter Merkmalstransformations-Koeffizienten (scale-invariant feature transform, SIFT) vor, welche aus mehreren Farbräumen extrahiert werden. Klassifikatoren aus Ensembles von Entscheidungsbäumen, die mit diesen Merkmalen trainiert werden, haben einen hohen Grad an Robustheit gegen Sprecher- und Beleuchtungsvariationen. Während zur Detektion von Sprechaktivität zweidimensionale SIFT-Merkmale genutzt werden, um statistische Modelle zu erstellen, setzen wir für das Lippenlesen dreidimensionale SIFT-Merkmale ein, welche die zeitliche Dynamik von Videodaten festhalten können. Es wird gezeigt, dass die Verwendung dreidimensionaler SIFT-Merkmale im Vergleich zu herkömmlichen visuellen Merkmalen eine wesentliche Verbesserung der Erkennungsgenauigkeit bewirkt. Das vorgeschlagene AV-ASR erreicht eine Erkennungsrate von 70.5%, welches der höchste bisher publizierte Wert ist für die Oulu-Datenbank bei sprecherunabhängiger Erkennung.

Es wurden bisher viele Boosting-Algorithmen entwickelt um Ensembles von Klassifikatoren zu trainieren. Die meisten lassen sich jedoch nicht auf Mehrklassen-Klassifikatoren verallgemeinern, was bei unserer Anwendung notwendig ist. In dieser Arbeit schlagen wir eine Systematik für das Design von Boosting-Algorithmen vor. Diese Systematik hat den Vorteil, dass sie sich natürlich auf den Mehrklassenfall verallgemeinern lässt. Diverse Eigenschaften wie Konvergenzraten und Generalisierungsfehler sind analysiert worden. Ausserdem sind mehrere praktisch und theoretisch interessante Algorithmen wie beispielsweise SparseBoost abgeleitet worden. Es wird gezeigt, dass der SparseBoost-Algorithmus in jeder Trainingsrunde nur einen Teil (ungefähr die Hälfte) der Trainingsmuster nutzt und dennoch auf die optimale Hypothese konvergiert, im Sinne des PAC-Lernens (probably approximately correct).

Ein übliches Vorgehen zur Vereinigung von Audio und visueller Information ist die Gewichtung jeder Modalität entsprechend ihrer Zuverlässigkeit. Es wird in dieser Arbeit gezeigt, dass die Maximierung von AUC (area under the receiver operating characteristic curve) ein passenderes Kriterium zur Schätzung der Zuverlässigkeit ist, als häufig genutzte Kriterien wie die Erkennungsrate. Ausserdem wird hier die Zuverlässigkeit jedes Merkmals geschätzt. Dadurch wird die (konventionelle) zweidimensionale Schätzung

der Gewichte der Modalitäten zu einem ziemlich hochdimensionalen Problem generalisiert. Zwecks effizienter Schätzung der Zuverlässigkeit wird eine geglättete AUC-Funktion verwendet und eine Variante des projizierten Gradientenverfahrens adaptiert für die Maximierung des AUC-Kriteriums im Online-Modus.

Die audio-visuelle Detektion der Stimmaktivität (audio-visual voice activity detection, AV-VAD) ist eine wichtige Voraussetzung für viele audio-visuelle Anwendungen. Es wird eine robuste audio-visuelle Stimmaktivitätsdetektion vorgeschlagen, welche in halbüberwachter Weise trainiert werden kann. Diese interessante Eigenschaft ergibt sich aus dem Umstand, dass sowohl Audio als auch visuelle Signale dasselbe zugrundeliegende Ereignis repräsentieren, nämlich die Sprachproduktion. In diesem Ansatz erfolgt die Etikettierung der Daten durch iteratives Trainieren von audio-basiertem und visuellem Detektor und anschliessender Neuetikettierung, welche in der nächsten Trainingsrunde benutzt wird. Die Etiketten aus der letzten Iteration werden benutzt, um die endgültige audio-visuelle Stimmaktivitätsdetektion zu trainieren. Der vorgeschlagene AV-VAD-Algorithmus erreicht eine Detektionsrate von nahezu 96% pro Analyseabschnitt (der visuelle Stimmaktivitätsdetektor erreicht 78%) für den GRID- Datensatz bei sprecherunabhängigem Einsatz.