WATCH, LISTEN ONCE, AND SYNC

audio-visual synchronization with multi-modal regression CNN

Toshiki Kikuchi and Yuko Ozasa

Keio University, Japan

Background

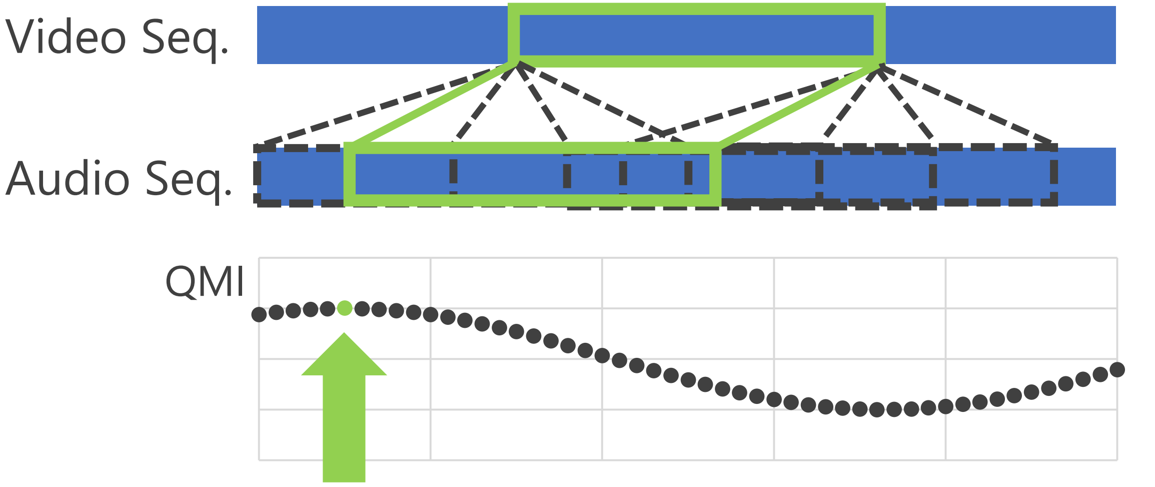
- The audio-visual synchronization drifts in some videos uploaded to video hosting/social networking services
- The viewers evaluate the video contents more negatively when the synchronization drifts [Reeves&Voelker, 1993]
- Recovering audio-visual synchronization is an important task in the field of visual speech processing
- Our work focuses on recovering audio-visual synchronization of single-person speech videos

B. Reeves and D. Voelker, “Effects of AudioVideo Asynchrony on Viewer’s Memory, Evaluation of Content and Detection Ability,” Research Report, Standford University, 1993.

Previous Work

Recovering AV-sync using QMI [Liu&Sato, 2010]

computes QMI between audio and visual features, and uses it as the correlation value to determine whether the audio and video are synchronized correctly or not.



The sliding window approach

Y. Liu and Y. Sato, "Recovery of audio-to-video synchronization through analysis of cross-modality correlation," Pattern Recognition Letters, vol. 31, no. 8, pp.696701, 2010.

Weakness of the Previous Work

a. Feature Extraction
They use the vertical optical flows of speaking lip image sequences as visual feature.
→ Lack of Robustness

b. Searching Approach
They have to shift by all possible synchronized position using the sliding window to find the correct shift position.
→ Worse Computational Complexity

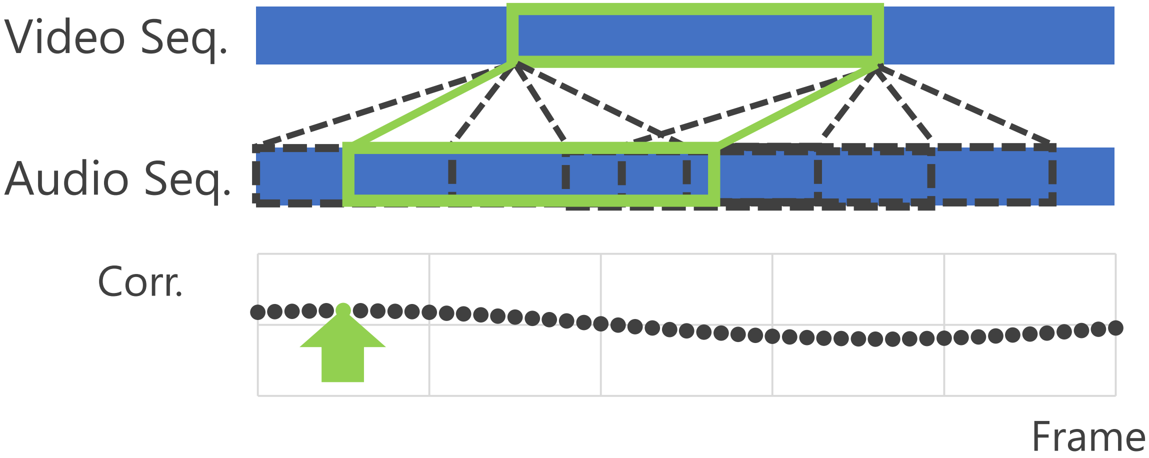
Our Contributions

1. We propose a novel architecture for recovering audio-visual synchronization using a CNN
   → Addressing Feature Extraction problem

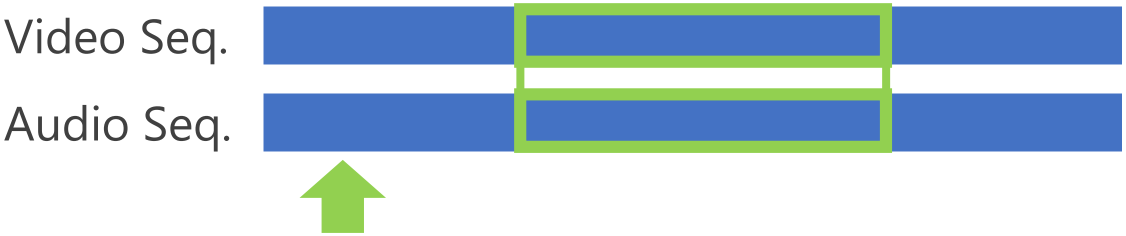
2. We show the benefit of treating audio-visual synchronization as a regression problem
   → Addressing Searching Approach problem

AV-sync Approaches

- The sliding window approach



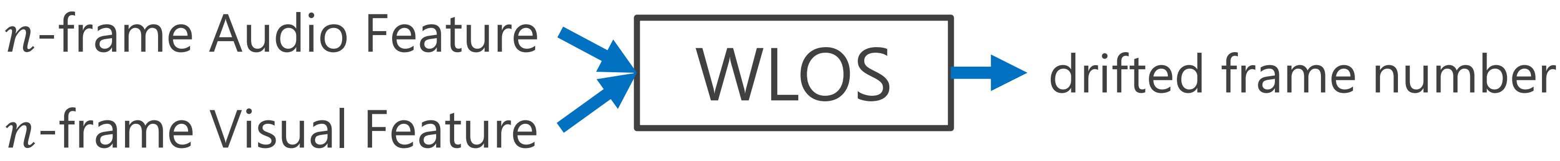
- The regression approach



Predict the correct position and sync!

Summary of Our Work

- We propose a multi-modal regression CNN for audio-visual synchronization for single-person speech videos. We call the proposed model WLOS (Watch, Listen Once, and Sync).



n-frame Audio Feature

n-frame Visual Feature

WLOS

drifted frame number

- We also show experimental results that demonstrate the proposed method outperforms baseline methods.

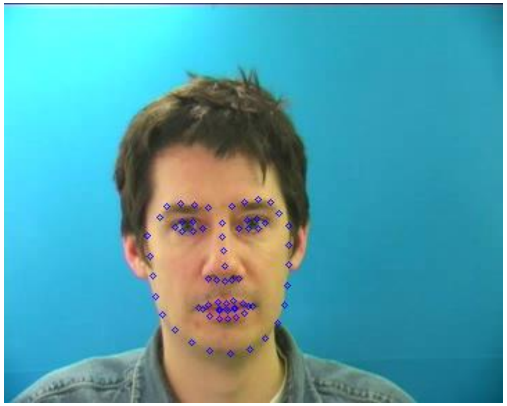
Visual Input Representation

① Detect facial landmarks by Kazemi and Sullivan's method [3] for each video frame

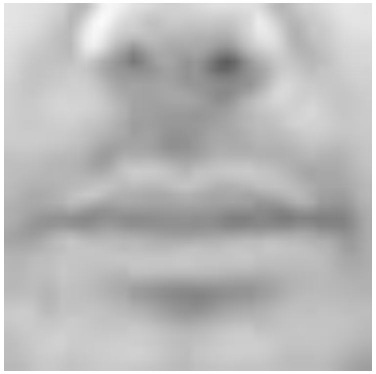
② Facial alignment with Affine Transform

③ Extract lip area of spatial resolution 32 × 32

④ Compute optical flows with Gunnar Farnebäck method [4]



Facial landmarks



Cropped and aligned lip area

[3] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in CVPR, 2014.
[4] G. Farneback, "Two-frame motion estimation based on polynomial expansion," in SCIA, 2003.

Audio Input Representation

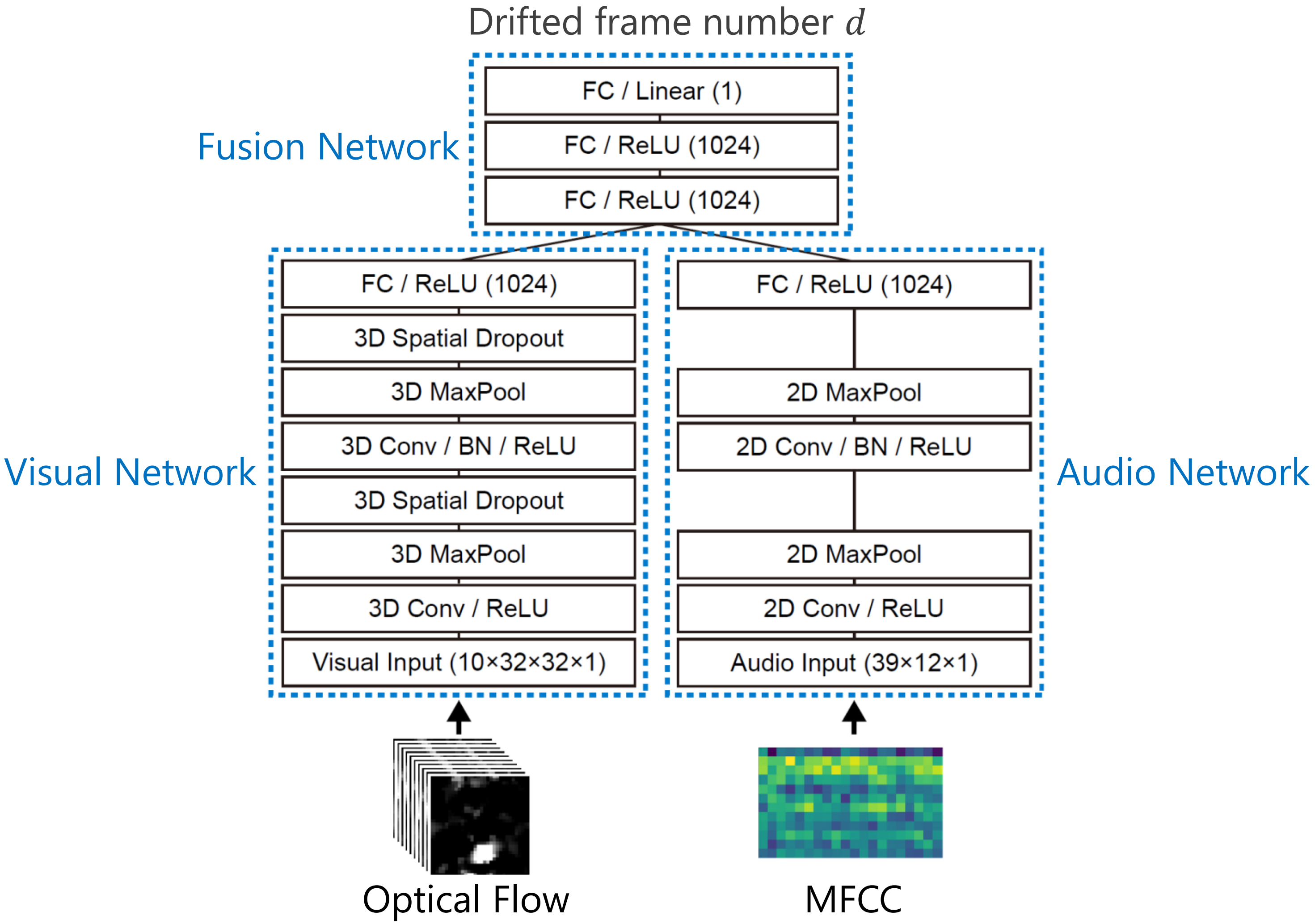
We use Mel-Frequency Cepstrum Coefficients (MFCCs) [Davis& Mermelstein, 1990] as the audio feature

① Apply the hamming window whose size is 256

② Compute the 13 MFCCs and use 12 MFCCs except the very first MFCC which is not informative about the actual spectral content

S. B. Davis and P. Mermelstein, “Readings in speech recognition,” chapter Comparison of Parametric Representations for MonosyllabicWord Recognition in Continuously Spoken Sentences, pp. 6574. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

Network Architecture of WLOS



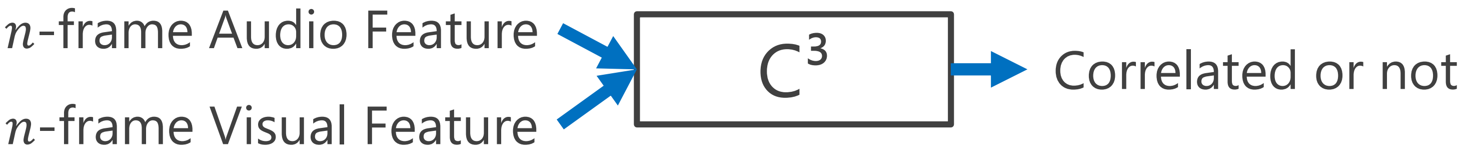
Fine-tuning

We use pre-trained weights as the initial weights of the visual network and the audio network.

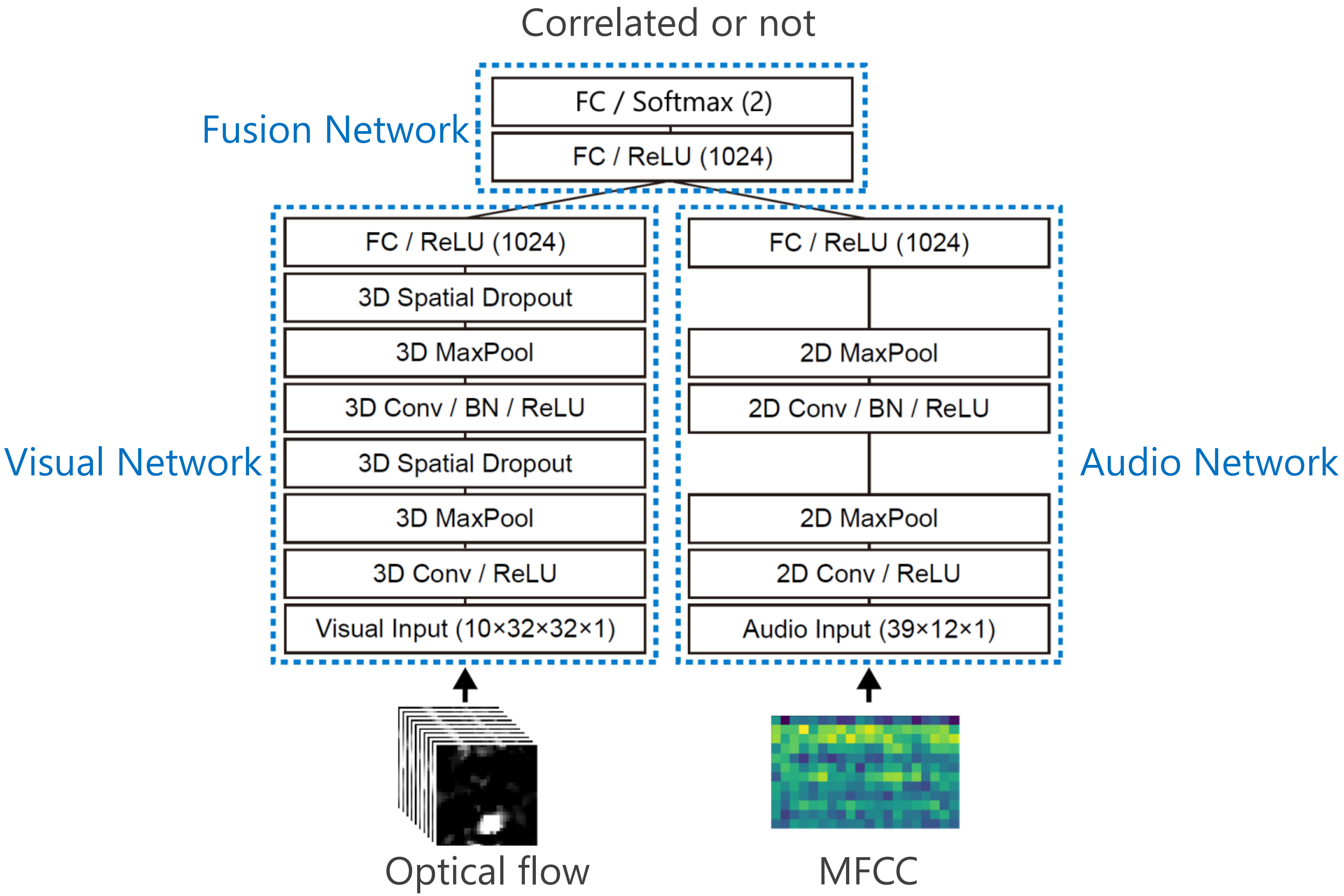
The network for pre-training

Train to predict whether the audio and visual information are correlated or not (binary classification problem)

We call this “classification correlation CNN (C^3)”



Network Architecture of C^3



Experiment – Dataset

Overview

- Intentionally drifted audio-visual dataset from GRID Corpus [Cooke *et al.*, 2006]
- The audio-visual pairs including −9- to +9-frame drifts

Details

- 63739 pairs for *S1*, 64000 pairs for *S2*
  63992 pairs for *S4*, 63997 pairs for *S7*
- 80%:20% random split for train and test

M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 24212424, 2006.

Experiment – Baseline Methods

a. QMI
Sliding Window Search with QMI
(based on [Liu&Sato, 2010])

Y. Liu and Y. Sato, “Recovery of audio-to-video synchronization through analysis of cross-modality correlation,” Pattern Recognition Letters, vol. 31, no. 8, pp.696701, 2010.

Experiment – Baseline Methods

a. QMI
   Sliding Window Search with QMI
   (based on [Liu&Sato, 2010])

b. C^3
   Sliding Window Search with C^3

Y. Liu and Y. Sato, "Recovery of audio-to-video synchronization through analysis of cross-modality correlation," Pattern Recognition Letters, vol. 31, no. 8, pp.696701, 2010.

Experiment – Baseline Methods

a. QMI
Sliding Window Search with QMI
(based on [Liu&Sato, 2010])

b. C^3
Sliding Window Search with C^3

c. WLOS (scratch)
WLOS without pre-training.

Y. Liu and Y. Sato, "Recovery of audio-to-video synchronization through analysis of cross-modality correlation," Pattern Recognition Letters, vol. 31, no. 8, pp.696701, 2010.

Experiment – Baseline Methods

a. QMI
Sliding Window Search with QMI
(based on [Liu&Sato, 2010])

b. C^3
Sliding Window Search with C^3

Sliding Window Approach

c. WLOS (scratch)
WLOS without pre-training.

Regression Approach

Y. Liu and Y. Sato, "Recovery of audio-to-video synchronization through analysis of cross-modality correlation,"
Pattern Recognition Letters, vol. 31, no. 8, pp.696701, 2010.

Result – Accuracy Evaluation

- CNN-based method outperform QMI baseline
- Positive effect of pre-training

Table 1. Mean absolute error on Test Data (frame).

Method	S1	S2	S4	S7
QMI (based on [Liu&Sato, 2010])	6.305	6.494	6.196	6.479
C^3	1.352	2.134	2.983	1.019
WLOS (scratch)	0.937	1.003	1.116	0.848
WLOS (fine-tune)	**0.907**	**0.916**	**1.038**	**0.799**

Y. Liu and Y. Sato, "Recovery of audio-to-video synchronization through analysis of cross-modality correlation," Pattern Recognition Letters, vol. 31, no. 8, pp.696701, 2010.

Result – Computational Time

WLOS is approximately 19 times faster than C^3 because WLOS does not use a sliding window.

Table 2. Processing time to synchronize a frame

Method	Approach	Time [ms]
QMI (based on [Liu&Sato, 2010])	Sliding Window	46.07
C^3	Sliding Window	34.00
WLOS	Regression	**1.80**

Y. Liu and Y. Sato, "Recovery of audio-to-video synchronization through analysis of cross-modality correlation," Pattern Recognition Letters, vol. 31, no. 8, pp.696701, 2010.

Conclusion

- We proposed a multi-modal regression CNN for recovering audio-visual synchronization
- The proposed approach enables us to recover errors without searching with a sliding window which would increase computational cost
- Experimental results show that the proposed method performs better than the baseline methods
- WLOS is more accurate and faster!
- In future work, we will make it possible to correct audiovisual synchronization errors of general videos instead of speech videos