Recovering Audio-to-Video Synchronization by Audiovisual Correlation Analysis

Yuyu Liu [†,‡]
† *Institute of Industrial Science*
The University of Tokyo

Yoichi Sato [†]
‡ *System Technologies Laboratories*
Sony Corporation

{*liuyuyu,ysato*}*@iis.u-tokyo.ac.jp*

Abstract

Audio-to-video synchronization (AV-sync) may drift and is difficult to recover without dedicated human effort. In this work, we develop an interactive method to recover the drifted AV-sync by audiovisual correlation analysis. Given a video segment, a user specifies a rough time span during which a person is speaking. Our system first detects a speaker region using face detection. It then does a two-stage search to find the optimum AV-drift that can maximize the average audiovisual correlation inside the speaker region. The correlation is evaluated using quadratic mutual information with kernel density estimation. AV-sync is finally recovered by the detected optimum AV-drift. Experimental results demonstrate the effectiveness of our method.

1 Introduction

Audio-to-video synchronization (AV-sync) is important for human sensing. Reeves and Voelker [8] discovered that, if the AV-sync drifts, humans evaluate the video contents much more negatively. When audio precedes video by 5 video frames, the satisfaction of the viewers degraded about 84% [8]. Research shows that audio should never lead video by more than 15 ms, and should never lag video by more than 45 ms [1].

AV-sync drift has many causes, such as different processing time between video and audio, different network transfer delay, drift accumulation among different processing stages, and so on. To avoid the drift, researchers have made great efforts, which mainly concentrated on making specifications for both video processing hardware and software to keep AV-sync [1]. Video cameras record a timecode to keep AV-sync for read-out. An MPEG-2 codec prints a Presentation Time Stamp (PTS). However, few attempts have been made to recover AV-sync when video and audio have drifted out of sync. As video processing often includes many stages, the lack of ability to recover means that all stages must be carefully designed to avoid drift. Additionally, even though each stage causes only minor drift, the drift can still be accumulated into an obvious one. As a solution, high quality videos like commercial films always include a final AV-sync adjustment stage, where a dedicated effort of human observers and a special device called an audio synchronizer are used.

In this work, we develop an interactive method to recover the AV-sync drift with only minor human effort. Given a video segment, the user only needs to specify a rough time span during which a person is speaking. The speaker is supposed to be relatively stationary in the time span. Our system automatically detects a speaker region by face detection and analyzes the audiovisual correlation inside this region for recovery. A two-stage search is then performed to find the AV-sync drift by a process called audiovisual correlation analysis.

Audiovisual correlation analysis is a relatively new research topic and has drawn much attention in recent years. In 1999, Hershey and Movellan [2] first introduced a method to analyze the audiovisual correlation for each pixel using *Mutual Information* (MI). Their method assumed that audiovisual signals obey joint normal distribution, which is too strong an assumption. Another correlation measure is tried in [3]. Although it can optimize a projection vector that maximizes the correlation, it cannot give the correlation value. Monaci *et al.* [5] employed the Pearson correlation efficient to compute audiovisual correlation value between object movement parameters and audio energy. Yet, the extraction of movement parameters seems prone to fail if there is a cluttered background.

Thus, to be able to evaluate pixel-based audiovisual correlation in a reasonable time span, we introduce a new measure of audiovisual correlation using *Quadratic Mutual Information* (QMI). It is based on

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

kernel density estimation and can evaluate correlation for arbitrary distributions.

The main contribution of this work lies in: 1) the use of audiovisual correlation analysis to recover AV-sync for video segments with little human effort, and 2) the development of a measure of audiovisual correlation value using kernel density estimation and QMI.

The rest of this paper is organized as follows. In Section 2, we introduce our method to analyze the audiovisual correlation. In Section 3, we introduce the method to detect the AV-drift. In Section 4, we present and discuss our experimental results. In Section 5 we present our conclusions.

2 Audiovisual correlation analysis

In our system, the most important step is to find out the correct audio-to-video drift value (AV-drift) of the AV-sync. Reversely shifting the audio part according to the detected value will set the video back to AV-sync.

We perform audiovisual correlation analysis to detect the AV-drift. To do this, we compute QMI to analyze the correlation between our extracted audio and visual features. For a group of video frames and their audio parts, correlation analysis will be done for each image coordinate (u, v) inside the speaker region. Below we will first introduce our audiovisual features and then the QMI computation.

2.1 Audio feature

We adopt differential energy as our audio feature. Input audio data are framed first. The frame duration T_a is set to be the same as the visual frame duration T_v. An overlap of the duration of $T_a/2$ between each pair of two successive frames is set. Finally, we multiply a Hamming window to the framed audio samples.

The logarithm energy $a(t)$ of each audio frame is computed by $a(t) = \log\left(\frac{1}{M}\sum_{m=1}^{M} s^2(t,m)\right)$, where $s(t,m)$ refers to the processed audio sample m in frame t. The audio feature is defined to be the differential energy between the current and next frames, i.e., $fa_t = a(t+1) - a(t)$.

We perform a verification of the existence of speech. This verification is done by checking whether or not the audio energy $a(t)$ is beyond a pre-defined threshold or not. Frames failing this test are regarded as silence and dropped, together with their corresponding image frames. Note that, if $a(t+1)$ fails the test, fa_t will not be extracted either. From here on, if we talk about N video frames with speech, this refers to the frames that pass this test.

2.2 Visual feature

We believe that better audiovisual correlation exists not in pixel values but between the movements of the photographed objects and their audio signal. Hence, we adopt the optical flow to be our visual feature. Optical flow has two elements for each pixel: horizontal and vertical movement. We take its vertical element only since most speaking actions move vertically. The visual feature $fv_t(u, v)$ is thus the vertical optical flow at (u, v) computed between frame t and $t+1$. We compute the optical flow on gray images using the Lucas-Kanade method [4], where the window size we used is 9×9.

2.3 Quadratic mutual information

We regard the audio and visual features as two random variables and compute their statistical correlation using their temporal samples. Note that the correlation will be computed independently for each image coordinate (u, v) inside the speaker region.

We adopt kernel density estimation to estimate their joint probability density function (pdf) first. Kernel density estimation [6] (also known as Parzen window estimation) is a method to estimate the arbitrary pdf of a random variable. Given N data points $x_i, i = 1, \cdots, N$ in the d-dimensional space R^d, the multivariate kernel density estimation with kernel $K_H(x)$ and a symmetric positive definite $d \times d$ bandwidth matrix H, computed in the point x is given by

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} K_H(x - x_i), \tag{1}$$

where $K_H(\cdot)$ is the specified kernel function. In practice, Gaussian function $G_H(x)$ with the diagonal bandwidth matrix $H = diag(\sigma_1^2, \cdots, \sigma_d^2)$ is often adopted.

Based on the estimated pdf, we adopt MI to compute the correlation. However, direct computation of the Shannon MI by Eq. (1) is difficult since it does not yield a closed-form solution. Thus we adopt QMI proposed by Xu *et al.* [10], which is based on the quadratic form of Renyi entropy. For two random variables x_1 and x_2, this is defined as $C(x_1, x_2) =$

$$\log \frac{\iint p^2(x_1,x_2)dx_1dx_2 \iint p^2(x_1)p^2(x_2)dx_1dx_2}{(\iint p(x_1,x_2)p(x_1)p(x_1)dx_1dx_2)^2}. \tag{2}$$

It is easy to show that $C(x_1, x_2) \geq 0$ and the equality holds true if and only if $p(x_1) = p(x_2)$ using Cauchy-Schwartz inequality.

For random variables whose pdf's are estimated by Eq. (1) with diagonal Gaussian kernel adopted, QMI can be computed very efficiently since we have $\int G_H(x - x_i)G_H(x - x_j)dx = G_{2H}(x_i - x_j)$ [10].

3 AV-drift detection

Given a video segment, we assume that its AV-drift is a constant. Considering the reasons resulting in the drift, this assumption stands for most situations. For the drifts that are not completely constant, such as drifts caused by sudden network transfer delays, AV-drift is considered to be piecewise constant. The assumption still holds if we divide these videos into segments to process.

To begin detection, the user first specifies a rough time span where a relatively stationary speaker is speaking. The face of the speaker must be photographed in video with the speech recorded in audio. Only data inside the time span will be used. For a video segments having no part satisfying this condition, our current system cannot detect its AV-drift. Yet, if it has adjacent segments whose AV-drift is detected, it is possible to use this value to recover the current segment.

Our system then detects a speaker region automatically using a face detection method in [9]. To reduce noisy movements, we take the lower half part of the face as the speaker region, which mainly represents the mouth. When multiple faces are detected, our system selects the one that has the highest audiovisual correlation. If face detection fails, the user can also specify the speaker region manually. As the speaker is supposed to be relatively stationary, the decision of the speaker region is made using the first image frame in the time span only.

The optimum AV-drift is searched by first a coarse and then a fine stage. In the coarse search, drift value is quantized into integer times of the video frame duration T_v, i.e., $d_1 = \{-L, \cdots, -1, 0, 1, \cdots M\}$. $[-L, M]$ represents a pre-defined search range. Positive AV-drift values represent how much audio lags video. Negative ones show how much audio precedes video. For each drift value d_1, we shift the audio data temporally and compute the average audiovisual correlation inside the speaker region. The optimum value d_1^* which has maximum average correlation is regarded as the found AV-drift in stage one.

The second stage fine search refines d_1^* to a better resolution than T_v. We fit a parabola to the analyzed correlation values around the d_1^* and take its maximum d_2^* as the refined result, which is computed by

$$d_2^* = d_1^* + \frac{0.5 \cdot (C(d_1^* - 1) - C(d_1^* + 1))}{C(d_1^* - 1) - 2C(d_1^*) + C(d_1^* + 1)}, \quad (3)$$

where $C(\cdot)$ gives the average audiovisual correlation for different AV-drift values. The final detected AV-drift will be $d_{av}^* = d_2^* \cdot T_v$.

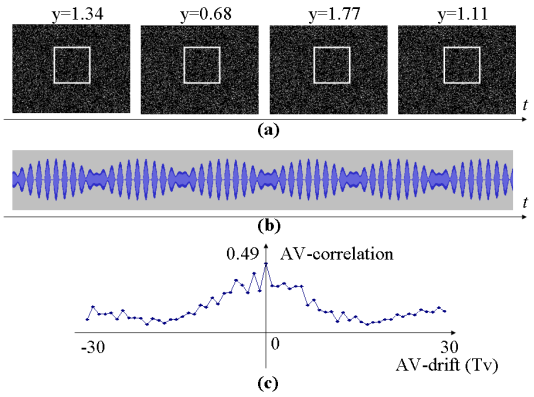

Figure 1. Experimental result of the ground truth. (a) shows video frames of the moving random dotted pattern with the vertical shift value y denoted above them. The white rectangle shows the speaker region adopted. (b) shows the audio signals. (c) plots the correlation values w.r.t. different AV-drifts.

4 Experimental results

As most off-the-shelf video cameras can supply 30 frames per second (fps) or higher video data, our experiments mainly concentrated on this kind of data. The search range was set as $[-1, 1]$ second, i.e., $L = M = 30$. The kernel bandwidths are set to $(\sigma_1, \sigma_2) = (0.4, 0.3)$ for all our experiments. We also supposed that the user specified the time span to be 1–4s, within which we adopted 40 video frames to compute the experimental results.

To test the correctness of our method, we synthesized a ground truth clip that simulates a speaking action by shaking a random dot image (320×240) vertically to represent the video and modulating a 2KHz sine wave to represent audio, which is shown in Figure. 1 (a) and (b). The movement is computed by multiplying two sine curve with a long and a short duration. We first let the AV-drift be zero and tested our method. The AV-drift d_{av}^* was closely detected as 1.63ms. Changes to the average audiovisual correlation are shown in Figure. 1 (c). We also produced ground truth data whose AV-sync drifted and applied our method. The detection results are shown in Table. 1.

We tested our method on real data. We used the clips from the CUAVE audiovisual database [7], in which people spoke English numbers in front of a green background singly or in groups. Images were converted into gray and down-sampled from 720 × 480 to 240 × 160 to speed up processing.

We tested the situation when a single person exists.

Table 1. Detected AV-drift *vs.* ground truth

Ground truth (ms)	d^*_{av} (ms)
-540	-536.58
-170	-166.28
230	233.32

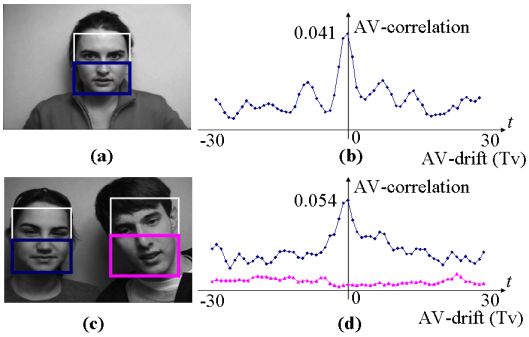

Figure 2. Speaker regions and their average audiovisual correlation. (a) and (c) show the detected faces and the speaker regions. (b) and (d) show the average correlation values w.r.t. different AV-drifts.

The speaker region and the average audiovisual correlation with regard to different AV-drift are shown in Figure. 2 (a) and (b). Our method found d^*_{av} as -11.29ms. Although the accurate real value is unknown, a minus value close to zero seemed to fit the situation of CUAVE data, where taken video were compressed by a heavier MPEG-2 codec than audio. We also intentionally added a drift to the real data and applied our method. The detected results were shown in Table. 2. Our method accurately detected the plused drifts for real data also.

When multiple persons exist, face detection gave multiple results. An example is shown in Figure. 2 (c) and (d). The average audiovisual correlations in the two face regions show that our method can correctly locate the speaker. Our method found d^*_{av} as -12.81ms, which was close to the previous one. Since both clips were from CUAVE database and should have similar AV-drift values, the consistency of the two found d^*_{av} demonstrated the correctness of our method.

Table 2. Detected AV-drift *vs.* drift added

Drift plused (ms)	d^*_{av} (ms)
-540	-543.89
-170	-173.35
230	226.84

5 Conclusions and future work

In this work, we have developed an interactive method to recover the AV-sync drift by audiovisual correlation analysis. In a time span specified by a user, our method can find the AV-drift with the highest audiovisual correlation by a two stage search. We have also introduced a new pixel-based measure of audiovisual correlation using kernel density estimation and QMI.

In the future work, we plan to investigate other noise-robust audiovisual features. Additionally, since our method requires the speaker to be relatively stationary in the time span, we are planning to extend our method to relax this constraint.

References

[1] "Relative Timing of Sound and Vision for Broadcast Operations". *Advanced Television Systems Committee report*, IS-191, 2003.

[2] J. Hershey and J. Movellan. "Audio vision: Using audiovisual synchrony to locate sounds". In *Proc. NIPS*, pp. 813–819, 1999.

[3] E. Kidron, Y. Schechner, and M. Elad. "Pixels that sound". In *Proc. CVPR*, pp. 88–95, 2005.

[4] B. Lucas, and T. Kanade. "An Iterative Image Registration Technique with an Application to Stereo Vision". In *Proc. IJCAI*, pp. 674–679, 1981.

[5] G. Monaci, O. Escoda and P. Vandergheynst. "Analysis of multimodal signals using redundant representations". In *Proc. ICIP*, pp. 145–148, 2005.

[6] E. Parzen. "On the estimation of probability density function and the mode". *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[7] E. Patterson, S. Gurbuz, Z. Tufekci and J. Gowdy. "Moving-talker, speaker-independent feature study and baseline results using the cuave multimodal speech corpus". *EURASIP J. on Applied Signal Processing*, 2002(11):1189–1201, 2002.

[8] B. Reeves and D. Voelker. "Effects of Audio-Video Asynchrony on Viewer's Memory, Evaluation of Content and Detection Ability". *Research report*, Standford University, 1993.

[9] P. Viola and M. Jones. "Robust Real-Time Face Detection". *Int'l J. of Computer Vision*, 57(2):137–154, 2004.

[10] D. Xu, J. Principe and J. Fisher. "A Novel Measure for Independent Component Analysis (ICA)". In *Proc. ICASSP*, 2:1161–164, 1998.