# [2020]SyncWISE: Window Induced Shift Estimation for Synchronization of Video and Accelerometry from Wearable Sensors

## Abst

ウェアラブルデバイスを用いて日常の人間の行動（例えば、食事、喫煙、歯磨きなど）を検出するための計算モデルの開発と検証には、自然のフィールド環境から収集したラベル付きデータと、微小行動（例えば、喫煙中の手と口のジェスチャーの開始時間と終了時間）と関連するラベルの時間同期を厳密に行う必要があります。ビデオデータは、このようなラベル収集のためにますます使用されるようになってきている。残念ながら、独立した（そしてドリフトする）時計を持つウェアラブルデバイスやビデオカメラでは、厳密な時間同期が困難です。この問題に対処するために、我々は同期のためのWindow Induced Shift Estimation method for Synchronization (SyncWISE)アプローチを提示します。我々は、実際の禁煙研究に登録された21人の参加者の45.2時間分のデータを163本の動画から得た、ウェアラブルカメラとウェアラブル加速度センサのタイムスタンプを同期させることで、我々の手法の実現可能性と有効性を実証しました。我々のアプローチは、高いデータロスがあっても、700ミリ秒の同期許容誤差で90%の同期精度を達成するなど、**最先端の手法**に比べて大幅な改善を示しています。また、本手法はCMU-MMACデータセットにおいても最先端の同期性能を達成しています。

## 1. Introduction

　センサー信号から受動的に行動を感知して推論するモデルを構築するためには、センサーデータの時間的に正確なアノテーションが必要である。喫煙、食事、歯磨きなどの手足の動きを含む行動については、ウェアラブルカメラのビデオ記録を用いて時間精度の高い真値ラベルを取得するケースが増えてきています（図1）。ウェアラブルカメラを装着して、フィールド条件下での動きを撮影します（図1参照）。ビデオはターゲットのモバイルセンサデータと同時に記録され、標準的なビデオコーディングを使用してセンサストリームの真値ラベルを取得します。これらのデータは、既存の手法の精度を検証するためと、新しいモデルを訓練するための両方に使用することができます。このアプローチは、飲食や歯磨きなどの活動[1-4, 9, 45]に使用されており、数秒単位の細かい活動には特に価値がある。 しかし，このアプローチでは，ビデオから得られたアノテーションを自動的に転送してセンサデータにラベルを付けることができるように，**ビデオシーケンスとセンサデータストリームの間の正確な時間同期が必要である．**ビデオとセンサストリーム間の時間的なずれは、ラベルノイズ（すなわち、センサデータの不正確なラベリング）の原因となり、検出器の精度を著しく低下させる可能性があります。

　汎用センサハードウェアでは、独立した非同期クロックを利用するのが一般的であるため、以前のセンサシステムアーキテクチャでは、センサ同期への効果的なアプローチが組み込まれていました[12, 25, 38]。 残念ながら、これらのアプローチをビデオキャプチャに簡単に拡張することはできません。センサーネットワークアプローチ[38]とは対照的に、GoProのような市販のウェアラブルビデオカメラは、他の非カメラセンサーとの同期用に設計されていません(Timecode Systems [30] の SyncBac Pro や ':pulse' などのサードパーティベンダーのアドオン製品は、ウェアラブルカメラの同期化ソリューションを提供することができますが、複数のカメラの同期化に限定されており、私たちのシナリオには対応していません)。さらに、バッテリーの制約により、ビデオデータをワイヤレスで送信することができず、他のセンサーストリーム[25]と同時に中央の収集ポイントでタイムスタンプを付けることができません。**これらの問題の結果として、ビデオカメラ[20, 37, 41]とウェアラブルデバイス[35]の時間同期の問題は、研究の実施における実用的な課題であり、データ精度のサイレントキラーであることは、コミュニティ内でよく知られています[4](**映像制作のために音声と映像を同期させるような基本的なことでも、長い歴史の中で課題や失敗があります。例えば、1972年に行われたアレサ・フランクリンの伝説的なライブパフォーマンスは、録音時に映像と音声を適切に同期させることができなかったこともあり、46年間公開されませんでした[19]**)**

　データが実験室条件下で収集される場合、オーディオビジュアル（AV）同期のためのよく知られたクラッパーボードや、身体に装着した加速度計をカメラ[34, 35]と同期させるための特別なハンドジェスチャーの使用など、同期ポイントを確立するために多くの戦略を使用することができます。これらのアプローチは、参加者に大きな負担をかけ、参加者のアドヒアランスに依存するため、実地研究には非現実的である[10]。代わりに、ELAN [11]やChronoviz [18]のようなツールを使って手動で同期を取ることもできる。このアプローチは手間と時間がかかり、クロックドリフトの結果、長時間の記録に渡って複数の時間点で実行する必要があるかもしれません。

　本研究では、Window Induced Shift Estimation for Synchronization of video and accelerometry (SyncWISE)と呼ばれる全自動手法を紹介する。動画と加速度センサのデータが与えられたとき、同期化のための時刻オフセットを出力する（図1(a)参照）。**我々は、部分観測性と座標登録という2つの重要な技術的課題に取り組んでいる。**部分的な観測可能性とは、同期点を確実に識別できる時間間隔が散発的に分散していることを意味します。例えば、街角に立っている参加者の胸に装着したカメラは、重要なダイナミック・ビデオ・コンテンツをキャプチャしますが、同位置に設置された加速度センサは動きを登録しません。これは、同期化を推定するためにはすべての時間の瞬間が等しく良いと暗黙のうちに仮定していた先行研究[13, 21]とは対照的です。**我々は、重み付けされたセグメントペアを相関させ、その票を集約して最終的なオフセットを得るカーネル密度推定アプローチを用いて、部分的な観測可能性に対処している。**第二の座標登録の課題は、加速度センサなどのモーションベースのセンサが三次元座標系を基準にしてデータを出力している映像を同期させる場合に発生する。この場合、信号を正しく比較するためには、2 つの座標系（カメラとセンサ）を登録し、対応する動きの方向を比較する必要があります。これに対して、自律走行車のセンサ同期に関する先行研究[21]では、センサが固定的に取り付けられ、設置時に校正されていることを利用しています。 **我々は、登録前にモダリティ間の共通の原理的な方向を特定するためにPCA分析を使用して座標登録に取り組んでいます**。CMU-MMACの活動データセット[16]と、Sense2StopSync (S2S-Sync)と呼ばれる新しい実世界のデータセットで、禁煙前3日間の45.2時間の記録で構成された21人の参加者を対象とした禁煙フィールド研究から得たものです。本研究では、以下の3つの貢献をしています。

* 加速度センサや慣性計測ユニット（IMU）などの動きベースのセンサデータとビデオクリップを自動的に同期させるSyncWISE方式を紹介します。私たちは、フィールド環境で発生する部分観測性や座標登録の課題をいち早く発見し、対処していると考えています。
* 喫煙と摂食のジェスチャーをアノテーションした21人の被験者から収集した2台の胸に装着した装置から得られた45.2時間の時間同期されたオプティカルフロービデオと加速度計データからなるSense2StopSync(S2S-Sync)データセットを研究コミュニティに提供する(4)。
* 本研究では、CMU-MMACとS2S-Syncデータセットについて、最先端の自動同期化の結果を発表する。この手法は， ベースライン法の2つのバージョンを有意に凌駕する[21]。このソフトウェアは自由に利用できるようにする。

## 2. Related Work

複数のセンサストリームの時間同期は、幅広いアプリケーション領域にまたがる長年の課題であり、**映画撮影中にオーディオとビデオを同期させるための1931年のクラッパーボードの発明から**、センサネットワークを同期させるために使用されるプロトコルに至るまで、長い歴史を持っています[12]。このレビューでは、モバイル・センシング・アプリケーションのためのウェアラブル・センサ・ストリームとビデオを同期させる方法に焦点を当てています。 このレビューでは、4 つのカテゴリーのアプローチを特定する。1) 我々の研究がその一例である自然主義的な手法であり、信号のキャプチャに特別な要件を課さない。2) 明示的な方法，それはデータ収集の間ソフトウェアかハードウェアのレベルで強制的に動悸する方法である．3) 参加者ベースの方法，これは同期のための特定行動を参加者に行ってもらうものである．4) 手動，すなわち人間自身がビデオを加速度を観測し，同期点を特定する手法である．

**2.1 Naturalistic Methods**

これらの手法の目的は、特別なハードウェアや特定の参加者の行動を伴わずに、フィールドでキャプチャされたセンサデータを扱うことである。我々の研究に最も近い先行研究は Fridman ら [21] であり、自律運転の研究のためにマルチモーダル信号を同期させるために設計された相互相関ベースの手法を記述している。彼らのアプローチでは、時間内のすべての瞬間が信号の同期に等しく適していると仮定し、最大量のデータを利用するためにグローバルな相互相関を使用しています。これは、彼らのセンサーが剛体に取り付けられており、座標軸が整列して校正されているために効果的である。対照的に、モバイルウェアラブルセンシングでは、センサが不適切に装着されていることによるセンサノイズ、センサ軸間の可変アライメント、部分的な観測性に悩まされています。我々のマッチングアプローチは、重み付きカーネル密度推定フレームワークで窓付き相互相関を使用し、データのどの窓が同期のための信頼できる信号を提供するかを特定することによって、部分観測可能性に対処します。我々のPCA-alignmentアプローチは、複数のセンサ間の座標フレーム軸を自動的に整列させる手段を提供する。**Sec.5の実験的評価では、2つのデータセットについて[21]のベースライン法よりも我々のアプローチの利点を実証している。**

　関連する自然主義的手法のセットは、GPSナビゲーションシステムのための同期化ソリューションを提供しており、[33, 39]はその代表的な例である。Skogら[39]は、GPS受信機とIMUの両方が直接センサシステムの空間的な位置に関連する信号を提供するという事実を悪用したクロックドリフトのためのカルマンフィルタベースのソリューションを提供します。対照的に、我々の設定では、我々はビデオから計算された光学的な流れは、部分的な観測可能性の問題のために加速度測定ストリームに直接関連付けることはできません。

　別の関連する手法として，複数のビデオストリームの自動同期化に対応したものがある[37, 41, 43]． これらのアプローチは、動画が固有の時空間的特性を持つ**単一のモダリティ**であるという事実を利用しています。これに対して、我々の研究では、各センサの信号から適切な特徴表現を抽出する必要がある、センサモダリティ間での同期化の場合に取り組んでいます。Chung and Zisserman [13]による関連研究では、ビデオダビングにおけるリップシンク効果を補正するために、音声とビデオのストリームを整列させるためにディープラーニング表現を使用しています。彼らのソリューションは、部分的な観測可能性が一般的な我々のケースとは異なり、**映像信号と音声信号が常に直接相関しているという事実を利用しています。**最後に、複数の先行研究では、同期された音声-映像信号が利用可能であることを仮定して、音源分離や音の分類などのタスクのために、共同の音声-映像特徴表現を構築しています[6, 23, 44]。これらの研究の中には、データ増強の手段として、音声チャンネルと映像チャンネル間の人工的な時間シフトを使用しているものもあるが、信号の同期化には利用されていない。

**2.2 Explicit Synchronization**

ハードウェアソフトウェアレベルで同期を強制する信号捕捉ソリューションが幅広く設計されています。ここでは、3つのアプローチに焦点を当てます。最初のアプローチは、ボディエリアネットワークを含むセンサネットワーク[12, 28, 38]で使用されます。すべてのセンサが同じネットワーク上にあるため、センサクロックの同期を保つためにプロトコルを使用することができ、アドレスクロックのドリフトやスキューに補正を適用することができます[12]。2 番目のアプローチでは、すべてのセンサ信号をスマートフォンなどの集中収集ノードに無線で送信することができ、そこで共通のクロックにタイムスタンプを付けて同期化を達成することができます（mCerebrum [25]が代表例です）。**これら2つのアプローチは、ネットワークサポートの欠如、ネットワーク帯域幅、およびバッテリーの制限のため、ウェアラブルカメラでは機能しません。**例外は、すべてのデータ収集が同じ場所で行われる場合である。 [1]では、歯磨きが行われる場所にスマートフォンホルダーを設置し、スマートフォンカメラ（集中ノード）自体に映像を記録することで同期を実現している。[16]では、実験室内で調理の様子を撮影し、ウェアラブルカメラと他のセンサーをハードウェアで同期させることが可能である**（例えば、あるデバイスからの参照信号を用いて他のデバイスを同期させるためにgenlockを用いる）。5.4節の実験では、[16]のデータセットを使用していることに注意してください。**3番目のアプローチは、リアルタイム同期を実現するために特別なハードウェアを使用する[7, 14]。[14]では、周期的に点滅するLEDを制御して、異なるモダリティを同期させるための合図を提供する。このようなアプローチは効果的であるが、追加の実装とシステムの複雑さを必要とし、LED信号の自動検出は制御されていない環境では困難な場合がある。 当社のアプローチは、汎用ハードウェアと標準化された研究グレードのmHealthソリューションを活用し、幅広い研究デザインをサポートしています。

**2.3 Participant-Based Methods**

オーディオビジュアル同期へのクラッパーボードアプローチは、モダリティ間で見える明示的な同期ポイントを導入することで、信頼性の高いソリューションを提供します。似たようなアプローチが他のマルチセンサー同期タスクにも存在します。**Plötzら[34]では、特定の手のジェスチャーを参加者に割り当てて、ビデオと加速度計データを整列させるための明示的な同期ポイントを提供しています。同様に、Hanら[22]は、脚の最大後方スイングを検出してマッチングすることで、歩行行動のためのビデオとセンサデータを同期させる方法を提案している。Bannachら[8]は、参加者に割り当てられた特定のジェスチャー（例えば「拍手」）を自動的に検出する方法を開発している。**これらのアプローチは制御された設定で動作することができるが、モバイル設定では参加者の負担が増え、失敗の一点張りとなる。

**2.4 Manual Synchronization**

代替的な同期化アプローチが失敗した場合、代替的な解決策は、ELAN [11]やChronoviz [18]のようなツールを使用することであり、検査を介して同期化ポイントの手動識別を可能にする。このアプローチは、mHealthやモバイルセンシングの先行研究で日常的に使用されており、デフォルトの方法と考えられるべきである[2, 3, 27, 35, 45]。我々の目標は、そのような手動の努力の必要性を取り除き、この重要な実用的な問題に対して完全自動の解決策を提供することである。

## 3. Study Design and Data Collection

ここでは，S2S-Syncデータセットの収集方法について詳細に説明します．まず，Sec.3.1 で研究デザインについて説明する．センサデータの収集についてはSec.3.2で、データアノテーションのアプローチについてはSec.3.3で説明します。

**3.1 Study Design**

データは禁煙研究Sense2Stopで収集された。参加者（18～65歳）は、過去1年間に1日1本以上のタバコを吸っていた場合にSense2Stopの対象となりました。S2S-Syncデータセットは、被験者が維持行動（典型的な喫煙パターン）を示した3日間の禁煙前期間中のSense2Stopから生成されます。禁煙前期間は、被験者の喫煙行動と食事行動のベースラインデータを提供した。禁煙前に収集されたビデオは、機械学習モデルの検証と改良のための喫煙行動と摂食行動のアノテーションをサポートしている。

**3.1.1 Study Timeline**

1日目、参加者はラボを訪れ、モバイルデバイスを装着して指示を受けた。4日目、参加者は研究室に戻り、ウェアラブルビデオデータを研究サーバーにアップロードしました。この訪問中に、参加者は共有したくないビデオ映像を削除する機会がありました。その後、参加者はビデオカメラを使わずに終了後の期間に入りました。

**3.1.2 Participant Instruction**

禁煙前の期間中、参加者は、少なくとも1回の喫煙イベントと少なくとも1回の食事と1回の間食を含む少なくとも2つの別々の日に4時間、提供されたGoProカメラを着用するように指示され、合計8時間の野生の記録されたビデオのために。

**3.2 Devices**

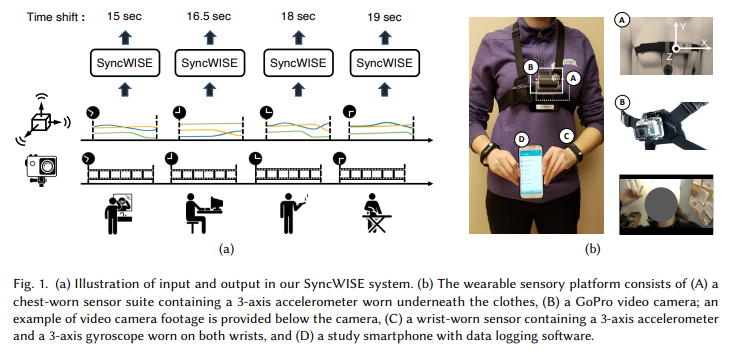
参加者が着用したウェアラブルデバイスには、胸に装着したGoProビデオカメラ、加速度計、心電図（ECG）センサー、呼吸胸膜厚計（RIP）センサーで構成される胸部装着型センサー群、それぞれの手首に3軸加速度計とジャイロスコープを搭載した1組の手首装着型デバイスが含まれていた。 さらに、データ収集ソフトウェアがインストールされた研究専用のスマートフォンが提供された。我々は、胸に装着した加速度センサとGoProビデオカメラとの同期に焦点を当てて分析を行っています。

**3.2.1 Video Camera**

参加者は、30Hzで1080pのビデオを記録するGoPro Hero 4カメラを装着した。GoProは、埃、水、その他の要素からカメラと画質を保護するチェストマウントストラップとケースを使用して胸に取り付けられました。カメラは参加者の顔の方を向いていました。キャプチャされたビデオは、それぞれ4GB、17分43秒の長さの一連のMP4ファイルとしてμSDカードに保存された。展開前に、GoProの時計をPCで国立標準技術研究所(NIST)のタイムサーバーに同期させた。追加の予防策として、カメラをPCに向けてNISTの時刻のウェブページ（time.gov）を簡単に記録し、カメラがフィールドに出る前に追加の同期リファレンスを提供しています。

**3.2.2 Accelerometers and Data Logging**

本研究で使用した加速度計は2セットである。10.66Hzでサンプリングされた胸に装着したデバイス、AutoSense [17]の加速度計を、自動同期実験のすべてで使用しました。このデバイスは、GoProとは別のハーネスに取り付けられていることに注意してください。図1(b)では、加速度センサは(A)の上にあり、カメラは(B)の上にあります。このように、2つのデバイスはほぼ同じ位置にありますが、カメラは加速度センサに対して、その向きの変化を含めて大きな動きをすることができます。手首に1つずつ取り付けられ、16 HzでサンプリングされたMotionSenseリストバンド [32] の追加の加速度計が、手動同期中に注釈者によって使用された（項3.3.1を参照）。その後、データはオープンソースのCerebral Cortex [24]を実行している安全なサーバーに定期的にアップロードされました。mCerebrumアプリは、データの各パケットを共通のクロックにタイムスタンプし、それによって加速度測定信号を互いに同期させた。ワイヤレス伝送では、パケットがドロップしたり、パケットが順番通りに到着しなかったりすることがあります。ソフトウェアはデータのわずかなギャップに対して補間を行いますが、加速度センサー信号には大きなギャップが残ります。同期化に対する我々のアプローチは、mHealthアプリケーションによく見られるデータの欠落と信号品質の低下を明示的に考慮している[26, 36]。



**3.3 Data Screening and Annotation**

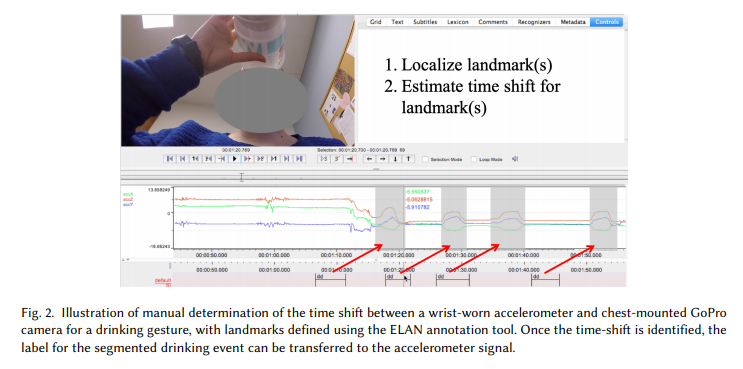
実験のデータ単位は、加速度測定データの関連セグメントを持つビデオクリップです。各クリップは、GoProによってキャプチャされた1つのMP4ファイルに対応し、最大持続時間は17分43秒でした。 30秒より短かったクリップは破棄され、34人の参加者から378個のクリップが得られました。GoProカメラで記録されたタイムスタンプは、クリップと対になっている加速度測定データのセグメントを識別するために使用されます。これは、かなりの誤差を伴う非常に粗雑な対応です。付録A.1に記載されている追加のスクリーニングステップにより、最終的なS2S-Syncデータセットは、合計45.2時間のビデオクリップ163個と、合計21人の参加者からの加速度測定データから構成されています。各参加者は平均2.15時間の解析に使用可能なデータを提供しました。

**3.3.1 Ground truth annotation Process**

我々のデータキャプチャプロセスでは、ビデオ信号と加速度センサー信号に別々のクロック（それぞれGoProのクロックと学習用携帯電話のクロック）を使用しているため、我々の実験のための接地真実を確立するために、各クリップ内のビデオと加速度センサー信号を手動でアライメントを行いました。我々のアプローチでは、基準タイムラインとして加速度計クロックを選択し、ビデオをアライメントにシフトさせる同期オフセットを選択しました。 図2に示す手動同期プロセスは、2つの段階から構成されています：ランドマーク映像の検出段階と、検出された映像のランドマークを2つの加速度計（手首または胸部に装着した加速度計）のうちの1つに整列させる加速度計の整列段階です。ビデオランドマークイベントは、ビデオの中で目に見え、手首または胸部装着型加速度センサのいずれかの信号の中で潜在的に気付くことができる、識別可能な人間の動き（多くの場合、非活動から活動への移行）として定義される。利用可能な同期点の数を増やすために、すべての加速度センサを利用しています。顔の近くでの手の動きは、手首の加速度センサーがビデオと一致することが多く、座っている状態から立っている状態への遷移や、立っている状態から歩いている状態への遷移は、胸部の加速度センサーがビデオと一致することが多いことがわかります。ランドマークとなるイベントがビデオにローカライズされた後、アノテータは、手首または胸部加速度センサ信号のいずれかの突然の変化を、ビデオ内の対応する手または身体の動きに一致させるために、合図の組み合わせを利用します。それらが正常に一致した後、時間オフセットが計算され、ビデオの時間がこのオフセットに基づいて加速度センサに整列するように調整されます。整合された信号を検査して、推定値が十分に正確かどうかを判断します。エラーが残っている場合は、他のランドマークイベントに対してこのプロセスが繰り返され、その結果、追加のオフセット測定が行われます。最終的なオフセットは、これらの測定値を組み合わせて生成されます。図2は、ELAN [11]アノテーションツールを使用して、ビデオフレームに表示されている飲酒イベントを、対応する手首の加速度計の動きに整列させるプロセスを示しています。

**3.3.2 Annotator Agreement**

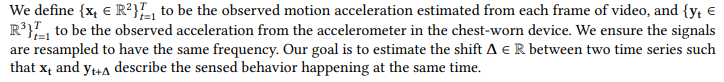
手動ラベリングの一貫性を評価するために、3人の訓練を受けたアノテータが10個のビデオクリップを処理し、基底真実のオフセット時間の独立した推定値を得た。彼らのオフセット時間の平均差は346ミリ秒で、標準偏差は309ミリ秒であった。これは、約346msの微細なラベリングと時間同期の基底真理取得に避けられない誤差があることを示している。また、一方向反復測定ANOVA（各グループの参加者が同じである場合に、3つ以上のグループの平均が異なるかどうかを決定するために使用）を実施しました。その結果、3つの注釈者の間に有意差はないことが示された（F=0.60、P=0.56）。したがって、異なるアノテータによるアノテーションの間には統計的に有意な差はないと結論づけられる。また、アノテータは平均して、1人のアノテータが17分のデータを手作業で同期させるために25分を費やしている。我々の研究は、この時間同期の問題を軽減することを目的としており、研究者が手動で時間同期を行う時間を何千時間も節約できるようにすることを目指している。



## 4 Methodology

本節では、ビデオと加速度センサ間の時間同期の問題を解決するためのアプローチを紹介する。まず、使用する記法、データの前処理、そして時間同期アルゴリズムであるWindow Induced Shift Estimation (SyncWISE)を紹介し、最後に評価指標について議論して終わります。

**4.1 Notation**

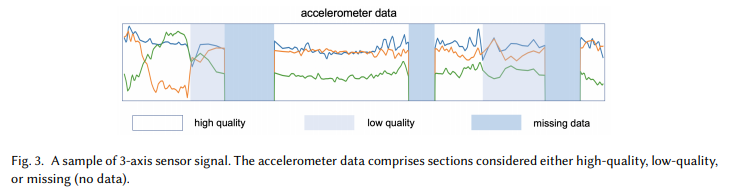


**4.2 Data Preprocessing**

前処理は2つのタスクをかかえる．1つは，マッチングをサポートする加速度データの高品質な窓を特定すること，もう1つは，加速度データと比較可能な動作信号を映像から抽出することである．

**4.2.1 Accelerometer Data Preprocessing**

図3に示すように、加速度センサのデータは、無線通信の問題により、データポイントが欠落している部分があるため、品質に大きなばらつきが生じることがあります。低品質のデータを確実にスクリーニングするために、1秒の長さのセグメントでデータを分析し、1秒間に収集されたデータポイント数を10.66Hzの予想サンプリングレートで割った比率として、セグメントごとの信頼性指標を定義します。そして、信頼性が75%を超えた場合（すなわち、1秒間に10.66サンプルのうち8サンプルが存在する場合）、データの各セグメントを高品質なものとしてラベル付けします。このしきい値を満たさないセグメントは、低品質とラベル付けされます。マッチングを実行するために、高品質のセグメントを10秒の長さのデータのウィンドウに集約します。実行可能なウィンドウは、高品質とラベル付けされた10個の連続したセグメント（すなわち、10秒）で構成されます。我々は、実行可能なマッチングウィンドウをセグメント化するために、1秒のストライドでビデオクリップを一掃します。17分のビデオクリップを同期化するためには、少なくとも10個の実行可能なウィンドウ（各10秒の長さ）を含む必要があります。抽出された各ウィンドウは、カメラのサンプリングレートに合わせて10.66 Hzから30 Hzにアップサンプリングされます。 このアプローチでは、SyncWISE アルゴリズム（4.3 節で説明）のビデオデータのウィンドウにマッチングするための出発点として、高品質の加速度測定データのウィンドウを使用していることに注意してください。このアプローチでは、ビデオフレームのドロップがないことを前提としていますが、これは実験で使用したデータセットに当てはまります。具体的には、S2S-Syncで使用したGoProはフレームをローカルに保存しており（3.2.1節参照）、CMU-MMACのデータはラボで収集したものです。我々のアプローチを拡張して、少数のコマ落ちにも対応できる可能性はありますが、その場合の有効性は主張していません。欠落した加速度計と同時に大量の欠落したフレームに対処することは、今後の研究課題です。



**4.2.2 Motion Estimation from Videos**

ビデオと加速度測定を比較するための重要な操作は、ビデオの動きから加速度の推定値を抽出することです。これには2つのステップがあります。まず、オプティカルフローとして知られる、各フレーム内の各ピクセルの速度の推定値を、ビデオ内の隣接するフレームの各ペアから計算します。動きの特徴は、一人称の動画から日常の活動を認識する際に使用されてきました[29, 31, 46]。 我々は、PWC-net[40]と呼ばれるディープラーニングに基づく密なオプティカルフロー推定フレームワークを使用している。結果として得られる2次元ベクトル場では、各ピクセル位置のベクトルが、𝑥(水平)方向と𝑦(垂直)方向のピクセルの動き推定値を提供する。第２のステップでは、一連の流れ場からスカラー加速度信号を抽出する。これを行うために、各フレームについて光学的流れを空間的に平均化し、次に、現在のフレームと前のフレームとの間の平均光学的流れの差を計算する。これにより、セクション4.1でxtと表記されている各フレーム𝑡の2Dカメラ加速度特徴ベクトルが得られる。

**4.2.3 PCA Projection**

ビデオデータの重要な点は、カメラの向きが任意であることであり、これは項4.2.2.2で計算された加速度ベクトルの向きに影響を与えます。参加者はカメラを頭の方向に向けるように指示されていましたが、実際にはデータセットの中で多くの異なる方向を観察しました。カメラハーネスが毎日着脱されるため、カメラの向きは日々変化する可能性があります。 その結果、センサ間の座標変換が発生し、時間の経過とともに被験者内でも被験者間でも変化します。ベースライン・アプローチ[21]では、同期点を決定するために固定された軸のペアからの信号を比較していますが、データセット内で時間を超えて変化する軸を比較する方法が必要です。 主成分分析(PCA)を使用すると、2Dビデオデータと3D加速度計データのそれぞれを、最も支配的な軸に沿った動きをキャプチャする1次元(1D)にマッピングすることができます。1つの軸に沿った動きを最もよく捉える1次元信号を抽出するために、各センシングモダリティ（ビデオ信号と加速度センサ信号の両方を別々に）からのデータを、PCAによって推定された第1原理成分の方向に投影します。第一原理成分は、データの変動が最も大きい方向に対応しており、これは、交差相関（我々が提案するアルゴリズムで使用）が信号の変動を利用するので、我々の目標によくマッチしています。形式的には、映像中に収集された映像データ{xt} 𝑇𝑉𝑘 𝑡=1 と加速度計データ{yt} 𝑇𝑉𝑘 𝑡=1 に対して、{xt} 𝑇𝑉𝑘 𝑡=1 と {yt} 𝑇𝑉𝑘𝑡=1 に分けて計算した PCA からの第一原理成分として、𝑝𝑣𝑑𝑖1 と 𝑝𝑎𝑐𝑐1 を用意した。そして、1 次元投影時系列を {x (p1) t = 𝑝 T 𝑣𝑖𝑑1 xt} 𝑇𝑉𝑘 𝑡=1 とし、{y (p1) t = 𝑝 T 𝑎𝑐𝑐1 yt} 𝑇𝑉𝑘 𝑡=1 とする。本稿では、紙面全体での省略のため、購読 𝑝1 を省略している。

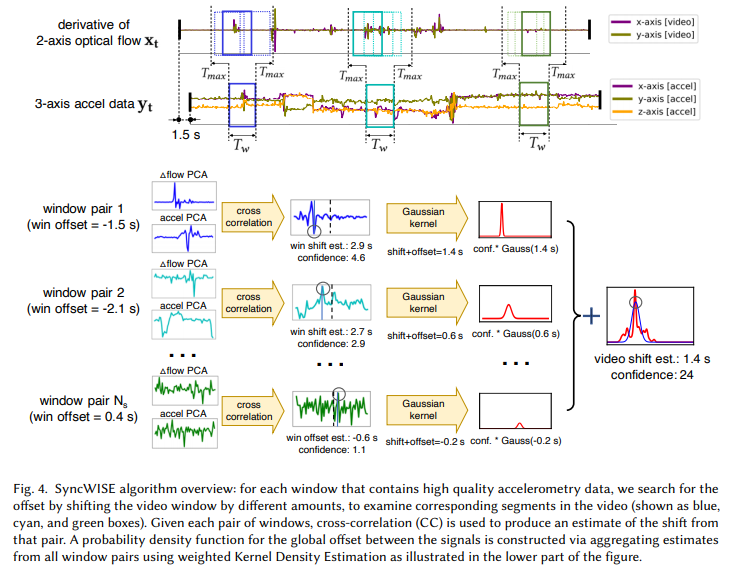
**4.3 SyncWISE Algorithm**

モバイルデバイスからキャプチャされたノイズの多いビデオ信号と加速度センサ信号をマッチングするための我々のアプローチは、2つの重要なコンポーネントを持っています。1つ目は、4.2.1節で説明した手順であり、図3に示すように、この信号のノイズとミッシングネスを克服する方法として、マッチングのために加速度測定データの高品質なウィンドウを選択します。 2番目の重要なコンポーネントは、重み付きカーネル密度推定（wKDE）を使用して、複数の加速度測定データとビデオウィンドウのペアから生成されたシフトのノイズのある推定値を組み合わせて正確な推定値を得るために使用します。これは、各ウィンドウペアからの相互相関応答を利用して重み付きガウスカーネルを取得し、これらのカーネルを組み合わせてオフセットを推定します。 オフセット推定のプロセスを図４に模式的に示す。図の上部には、ビデオクリップからの信号と、それに対応する加速度測定クリップからの信号が、1.5sの基底真実オフセットで区切られていることが示されています。加速度測定データのウィンドウが与えられると、ビデオウィンドウを異なる量だけずらしてオフセットを検索し、ビデオ内の対応するセグメントを調べます。これを図の上部に示します。青、シアン、緑の3つの異なる加速度測定ウィンドウが表示されています。各ウィンドウは、潜在的な一致を検索するために複数回シフトされます。ビデオ信号中の対応するウィンドウの位置の1つが実線で示され、他の位置が点線で示されています。各ウィンドウのペアが与えられると、そのペアからのシフトの推定値を生成するために、相互相関（CC）が使用されます。これは、実線で示された３つのウィンドウ・ペアのそれぞれについて、図４の下部に示されています。CC関数のピークを検出した後、ガウシアン・カーネルがデータにフィットします。すべてのペアが相関されると、信号間のグローバルなオフセットのための確率密度関数（PDF）は、ガウスカーネルの重み付けされた和（すなわち、オフセットPDFのためのwKDE）を介して構築されます。 実質的に、各ウィンドウペアは、オフセットに対する加重投票を行っていることになります。オフセットの単一の推定値は、PDF内のピークを検出することで得られ、その結果、この模式的な例では誤差100msで1.4秒の推定値が得られます。 最終的な推定値の信頼度スコアは、分散（この例では24）を得るために、ガウス分布をPDFにフィットさせることで得られます。この方法の擬似コードは付録A.2にあります。ここでは各ステップを詳細に説明します。

**4.3.1 Window Pair Sampling**

開始点はy  i = [𝑦𝑠𝑖  , 𝑦𝑠𝑖+1, . . . , 𝑦𝑠𝑖+𝑇𝑤  ]と定義される加速度データの長さTwのウィンドウwiであり，ここでsiとはwindowiのスタート時刻のindexである．各オフセット 𝑗 に対して、対応するビデオサンプルを x i,j = [𝑥𝑠𝑖+𝑜 𝑖 𝑗 、 𝑥𝑠𝑖+𝑜 𝑖 𝑗 +1 、 ... 、 𝑥𝑠𝑖+𝑜 𝑖 𝑗 +𝑇𝑤 ]と表記して得る。各 𝑜 𝑖 𝑗 は、Sec.4.3.2 の相互相関を介したマッチングのためのウィンドウペアを定義する（図 4 参照）。

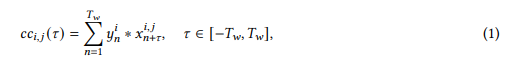
このアプローチには3つの設計パラメータがある：窓幅を制御する𝑇𝑤、探索範囲を制御する𝑇𝑚𝑎𝑥、オフセットの数を制御する𝑁𝑠。これらを図4に示す。これらのパラメータ設定の結果については4.3.4節で、その最適化については付録A.3で議論する。𝑁𝑠 オフセットを生成する方法はたくさんあることに注意してください．最も簡単な方法は，探索範囲を固定ステップで一様にサンプリングすることである．あるいは、シフトの事前推定値が利用可能な場合（例えば、キャプチャのセットアップや手動検査の知識から）、事前分布からサンプリングすることで、より可能性の高いオフセットにウィンドウ比較を集中させることができます（重要度サンプリングのように）。我々の実験では、𝑇𝑚𝑎𝑥 上の一様分布からオフセットをランダムにサンプリングしました。この方法は、固定ステップで探索範囲をカバーする方法と区別がつかないことを実験的に確認した。



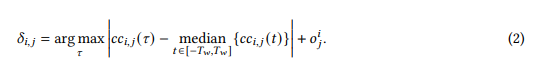
**4.3.2 Window Pair Matching**

与えられた加速度計-ビデオウィンドウのペアyiとxi、jは、セクション2で定義されています。 4.3.1、相互相関[21、34]を計算してそれらを照合し、クリップ間のシフトの推定値𝛿𝑖、𝑗を取得します。

特に，我々は相互相関関数を次のように，FFTベースの手法で計算します．



この加速度計とビデオデータのペアに関する最適なシフト量は，メディアンとシフト量で中心化された相互相関関数の絶対値を最大化するように推定されます．



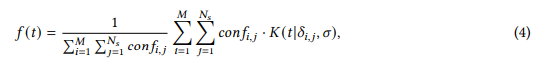
ここで，各δijは加速度窓wiとオフセットojから得られた推定シフト量を表します．

そして我々は，各シフト量に対する信頼値を次のように計算します．信頼地は，すべての窓ペアが心のオフセットに対して等しく有益ではない，という事実を捉えています．

直感的に，高くて鋭いピークほど信頼のおける推定に関連します

**4.3.3 Synchronization offset Estimation**

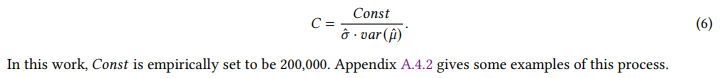
4.3.1項と4.3.2項で説明した手順を、入力クリップから選択した加速度測定データのウィンドウに対して、𝑀(𝑀)の回数だけ繰り返す。これにより、𝑀𝑁𝑠シフト推定値と関連する信頼値{𝛿𝑖,𝑗,𝑐𝑜𝑛 𝑓𝑖,𝑗 }のセットが得られる。これらの入力は、2つのクリップ間の可能なオフセット時間の分布のための重み付きカーネル密度推定値（wKDE）を得るために使用されます。これは、以下のように信頼度重み付き投票法によって構築されます。



ここで、 𝐾(𝑡|𝛿、𝜎) は、平均 𝛿 と標準偏差 𝜎 を持つガウスカーネル関数である。関数 𝑓 (𝑡) は、入力クリップのオフセットの範囲にわたる pdf であり、𝑀窓と𝑁𝑠 シフトから重み付けされたガウスカーネルを合計することによって構築される。最終的に推定されたビデオシフトΔは、次のように、𝑓 (𝑡) が最大となるところで達成される。



推定されたオフセットΔに関連付けられた信頼度スコアを持つことは貴重なことです。例えば、一致させるべきクリップのペアが与えられた場合、信頼度スコアは、ウィンドウサンプルの数 𝑀 が十分であるかどうかを判断するために使用することができます。我々は、非線形最小二乗法[15]を用いて、ガウシアン曲線 𝑔(𝑡) = N (𝜇、𝜎) を𝑓(𝑡) にフィットさせ、ガウシアンの分散 𝜎ˆと平均値の分散 𝑣𝑎𝑟(𝜇ˆ) を用いて、オフセット推定値の信頼度を推定します。ここでの直感は、ガウシアンの分散が高ければデルタオフセットの信頼度は低く、推定平均の分散が高ければ信頼度も低くなるということです。 推定分散 𝜎ˆと推定平均の分散 𝑣𝑎𝑟(𝜇ˆ)は、𝑓(𝑡)がガウス分布にどれだけフィットするかを予測します。信頼度スコアを次のように定義します。



**4.3.4 Discussion of Impact of Parameter Choices**

この方法は、ウィンドウサイズ𝑇𝑤、最大オフセット𝑇𝑚𝑎𝑥 、ウィンドウペアの数（シフト）𝑁𝑠 の選択に依存する。基底真理オフセット付近でサンプリングされたウィンドウのみが、応答曲線 𝑓 (𝑡) に正票を貢献できる。正票の割合は、せいぜい 𝑇𝑤𝑇/𝑚𝑎𝑥 . そのため、𝑇𝑤 の値は大きく、𝑇𝑚𝑎𝑥 の値は小さい方が好ましい。しかし、𝑇𝑤 の効果はデータの質にも依存し、𝑇𝑤 が大きすぎると、質の高いサンプルで十分な数の窓を見つけるのが難しい場合があります。 同様に、𝑇𝑚𝑎𝑥 は、期待される真のオフセットの最大値よりも大きいことが望ましい。𝑁𝑠は、候補となる同期点で探索範囲をカバーするのに十分な大きさであるべきであるが、その値の唯一の上限は計算資源である。 一般的に、これらのパラメータの最適な選択はデータセットに依存する。

## 5. Experiments and Results

提案したSyncWISEアルゴリズムを評価するために、広範な実験を行いました。我々は2つのデータセットで4つの異なる同期方法を評価しました：我々の新しいS2S-SyncデータセットとCMU-MMACのキッチンアクティビティデータセット[16]です。これらのアルゴリズムのバリエーションは5.1節で詳しく説明し，5.2節ではエラーメトリクスについて説明します．S2S-SyncとCMU-MMACの実験については，それぞれ5.3節と5.4節で詳しく説明します．

**5.1 Algorithm Variations**

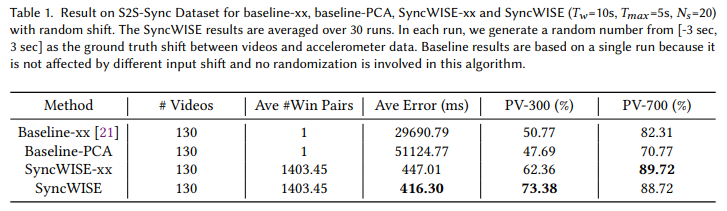
我々は、部分観測可能性と座標登録問題の扱いによって異なる4つの異なる同期化手法を特定した。我々のベースラインとなる手法は、[21]のアプローチであり、単一のグローバル相互相関(CC)を実行します。S2S-Syncデータセットでは、最後の観測を繰り越すことで、不足している加速度センサデータを埋めます。最初の Baseline-xx では、4.2.1 節で定義した加速度特性とカメラ加速度特性の両方の𝑥 成分を選択しています。この選択は、センサーの向き（図 1(b)参照）に動機づけられています。 𝑥 軸は横から横への動きに対応しており、実世界の設定では動きの間に強い動き信号が発生する可能性が高いからです7。SyncWISEファミリーの手法は、項4.3で説明したwKDEとのペアウィンドウマッチングアプローチを使用しています。変種の SyncWISE-xx は上述の 𝑥 軸座標の選択を使用しますが、SyncWISE には PCA 表現が組み込まれています。SyncWISEは両方のデータセットにおいて最も性能の高いバリアントであり、この問題の新しい最先端を確立しました。公正な比較をするために、Baselineに相互相関を適用する前に、各クリップから低品質のウィンドウをすべて削除し、ビデオと加速度計の両方で高品質のウィンドウを連結しています。

**5.2 Evaluation Metric**

我々はアルゴリズムの性能を評価するために2つの尺度を用いている: 1) 同期誤差の絶対値の平均、𝐸𝑎𝑣𝑔、および2) オフセット誤差が𝑛 ms 未満のオフセット誤差に同期しているクリップの割合、PV-𝑛である。 𝑛 の選択は、実際に役立つためには、時間的同期がどれだけ正確でなければならないかという問題につながります。一般的に、この質問に対する答えはアプリケーションに依存します。例えば、喫煙のパフは500ミリ秒と短いが、喫煙セッションは5-7分 [5] であり、歯磨きは2分 [1] である。S2S-Syncでのアノテータの平均不一致時間は346msであり、[8]などの先行研究では300msを精度目標としていましたが、S2S-Syncでのアノテータの平均不一致時間は346msでした。これらを考慮して、今回の実験ではPV-300と**PV-700**を精度の指標として報告することにしました。PV-𝑛 指標は、𝐸𝑎𝑣 指標を補完するものであり、いくつかの困難な動画が非常に大きな同期誤差を発生させた場合には、外れ値に敏感になる可能性があります。

**5.3 Experimental Results for S2S-Sync Dataset**

5.3.1節で述べたパラメータ設定を用いて、我々の手法の性能を検証するために2つの実験を行った。5.3.2 節の実験では、ランダムに生成されたオフセットを用いてシミュレートしたデータセットを用いて Baseline と SyncWISE の性能を比較し、大規模な評価を行った。5.3.3.3 節の実験では、S2S-Sync の元のクリップを同期させるために、我々の基本的なアルゴリズムを反復的に拡張した手法を用いて、この手法の能力を実証しています。さらに、付録A.3では、パラメータの変更がパフォーマンスに与える影響を示す、ホールドアウトされたデータセットを用いた包括的な感度分析を記述しています。すべての結果については，5.3.4節で説明します．



**5.3.1 Parameter Specification**

4.3.4節で述べたように、この手法のパラメータの選択は一般的にデータセットに依存します。ここでは、このセクションのすべての実験で使用されたパラメータを詳述する。加速度センサ信号の欠落データが比較的多いことから、𝑇𝑤 = 10s を選択した。𝑁𝑠 と𝑇𝑚𝑎𝑥 の性能への影響を特徴づけるために、付録 A.3 で詳述した感度分析を行った。𝑁𝑠 = 20、𝑇𝑚𝑎𝑥 = 5s を選択した。式4のwKDEには𝜎 = 500を用い、式6のコンフィデンスを得る際の𝜇と𝜎の探索境界をそれぞれ[-20, 000𝑚𝑠, 20, 000𝑚𝑠]と[0,𝐼𝑛 𝑓 ]とした。

**5.3.2 SyncWISE Compared to Baseline**

ここでは、20%(33本の動画)がアルゴリズムのパラメータを最適化するために使用され、80%(130本の動画)がアルゴリズムをテストして最終結果を得るために使用されます。130 個のテストビデオクリップの基底真理同期オフセットを使用して（セクション 3.3 を参照）、以下のように合成テストデータセットを生成します：[-3 秒、3 秒]の範囲のランダムオフセットを各クリップごとに 30 回サンプリングし、同期をシフトするために使用し、結果として 3,900 個のテストクリップを生成しました。これらの合成されたテストクリップを用いて、我々のアルゴリズムの性能を評価した8 。𝐸𝑎𝑣𝑣𝑔 は、Baseline と比較して SyncWISE の方が 2 桁小さいことがわかります。これは驚くべきことではないが、Baseline は品質指標に基づいて票を集約しないため、流行のシグナルノイズの影響を受けやすいからである。同様に、PV-300とPV-700は、差はそれほど大きくはないものの、SyncWISEの方が高くなっています。 SyncWISEでPCAを使用することで、特にPV-300については、SyncWISE-xxと比較して適度な利益が得られるが、Baseline-PCAはBaseline-xxよりもパフォーマンスが悪いことがわかる。これはPCAステップが主にBaselineでは使用されない信頼性推定値の質を向上させるためであると仮定している。付録 A.4 に両手法の応答曲線の例をいくつか挙げる。

**5.3.3 Synchronizing with the Original Offsets**

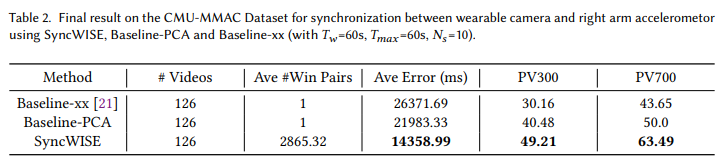
5.3.2 節のシミュレーション実験では、6 秒間のオフセット範囲で大規模な評価を行うことができました。しかし、S2S-Sync クリップの基底真実オフセットはかなり大きく、平均オフセットが 21 秒、最大オフセットが 180 秒、最小オフセットが 387ms と複雑な分布をしています。これらのクリップにSyncWISEを直接適用しても、ノイズが多く複雑な特性を持つ信号を正確に登録するには検索範囲が大きすぎるため、うまくいきませんでした。 そこで私たちは、検索範囲を任意に拡張できるSyncWISE法の対話型バージョンを開発しましたが、検証のために人間が介入しなければならないという欠点がありました。この問題を画像検索に例えて説明します。クエリが簡単な場合、最初に返された結果は信頼性が高く、すぐに受け入れられる。これは、6秒のオフセット範囲を持つ同期の場合に当てはまります。しかし、クエリが難しい場合は、信頼度を使って結果の順位付けをしなければならず、正しい一致を確認するためには手動での検査が必要になります。これは、信頼度の推定値が非常に大きな検索範囲では十分に正確ではないためです。 ここで、Extended SyncWISEのアプローチを以下のように説明する。我々は、ステップサイズ 𝑉𝑡𝑠𝑒𝑝 3s、最大シフト 𝑉𝑚𝑎𝑥 3分を定義する。これらは、S2S-Syncクリップのすべてのオフセットをカバーするのに十分な±3分の範囲で探索するために選択されている。我々は、-𝑉𝑚𝑎𝑥 から始まり、+𝑉𝑚𝑎𝑥 にインクリメントすることで、𝑉𝑠𝑡𝑒𝑝 , 我々の場合は120個のシフトクリップのセットを生成しています。シフトされた各クリップはSyncWISEを用いて同期され、これらのクリップからのオフセットは、式6で計算された信頼度に応じてランク付けされる。次に、ユーザに提案された同期ポイントをランク付けされた順番で調べてもらい、希望の精度レベルを確認します。 この実験の目的のために、我々はすでに真実の同期点を持っているので、ユーザの役割は自動化することができます。結果を図 5 に示す。ｘ軸は、調査しなければならないトップＫランクのプロポーザル（候補）同期ポイントの数Ｋを示し、ｙ軸は、トップＫ内の最良のプロポーザルが選択された場合に、ＰＶ-３００測定に基づいて正しく同期されたクリップの割合を示している。テストデータセットの130本のビデオの中には、理想的な録画条件ではないものがいくつかありました。5つのビデオは薄暗い照明条件で記録され、1つのビデオはレンズ上の水滴でぼやけていました。図5では、130本の動画と92本の動画（これらの問題のある動画を除去した後）について報告しています。問題のある動画を削除した後では、（標準的なSyncWISEのように）上位の同期案を選択しただけで、PV-300は54%程度になっていることがわかります。 しかし、上位5つのプロポーザルを検討すると70%程度に改善され、上位10のプロポーザルを検討すると74%のPV-300が得られます。この結果は完全に自動化されたアプローチには及ばないが、私たちの方法は幅広い応用が可能であり、困難で現実的な録画条件の下でビデオクリップを同期させる際に、人間の労力を大幅に節約できる可能性があるという利点がある。

**5.3.4 Discussion**

我々の結果は、ベースラインのグローバル相互相関法を上回るSyncWISEアプローチの利点を実証しています。ベースラインアプローチでは、ビデオ信号と加速度センサー信号のすべての部分にオフセットに関連する情報が含まれていると仮定し、それらを一つのグローバルな推定値に結合するため、これは理にかなっていると考えます。これは、部分的な観測可能性のため、我々のケースでは真ではないと思われます。我々の手法では、[8, 34]の制御された実験室内研究からのデータを同期する際に手動で定義されたものと同様に、「マーカージェスチャー」として動作するように、高い相関関係を持つ最も顕著な窓を自動的に見つけようとしています。センサーノイズの存在下で非常に長いオフセットに渡ってクリップを同期させる問題はまだ解決されていませんが、Extended SyncWISEアプローチはその場しのぎのインタラクティブな方法を提供します。SyncWISEを使用するにはユーザーがシステムパラメータを設定する必要がありますが、私たちの実験では既知の基準を利用した感度解析を行いました。一般的に、すべての同期問題に対して正しいシステムパラメータを事前に指定することは難しいでしょう。実際には、少数のクリップの同期と解析を繰り返すことで、初期のパラメータセットを改良していく反復的なアプローチが必要になると考えています。効果的なパラメータが見つかれば、データセット全体を自動的に同期させたり、Extended SyncWISEを使ってインタラクティブに同期させたりすることができます。

**5.4 Experimental Results for CMU-MMAC Dataset**

ここでは、よく知られたマルチモーダル活動データセット[16]であるCMU-MMACを用いた実験について説明する。この実験で使用した異なるモダリティのデータについては、5.4.1節で説明します。パラメータの設定については 5.4.2 節を参照のこと。Sec.5.4.3の実験では、データセットに提供された基底真理同期を用いてBaselineとSyncWISEの性能を比較している。さらに、付録A.5は追加の結果と例を提供する。その結果については、Sec.5.4.4で議論する。



**5.4.1 CMU-MMAC Dataset**

このデータセットでは、電球を取り外したヘッドランプにウェアラブルカメラを取り付け、被験者の頭に装着し、被験者の背中、足、腕に加速度センサーを5つ配置しています。データは有線接続を介してラップトップで記録され、信号間の同期はネットワーク時間同期プロトコルを介して達成される。同期を達成するための基底真理オフセットは、平均26.4秒で[1.8秒、178.0秒]の範囲内にあります。ビデオは30fpsで収集され、加速度計データは125Hzで収集されています。有効なビデオとIMUデータを持つ30人の被験者からの126セッションがあります。この実験では、同期化されていないビデオと加速度計を我々の方法への入力として使用しています。

**5.4.2 Parameter Specification**

加速度測定データをビデオサンプリングレートに合わせて補間します。0から60、90、120の3つの範囲でランダムオフセット 𝑜 𝑖𝑗 を選択し、同期化のための初期オフセットを生成し、窓付き相互相関を適用する。我々の方法で回収できる最大の時間的オフセットは、最大のランダムオフセット量𝑇𝑚𝑎𝑥と窓サイズ𝑇𝑤の和であることに注意してください。このデータセットでは、𝑁𝑠 = 10とする。ガウスカーネル 𝐾(𝛿、𝑡)を選択し、𝜎 = 3000𝑚𝑠 とする。𝑔(𝑡)の𝜇と𝜎を推定するための探索境界をそれぞれ[-600𝑠, 600𝑠]と[0,𝐼 𝑛 𝑓 ]とした。

**5.4.3 SyncWISE Compared to Baseline**

表2は、Baseline法とSyncWISEアルゴリズムの結果を示しています。入力クリップはウェアラブルカメラの信号と右腕からの加速度センサのデータを組み合わせたものです。他の場所からの加速度センサとビデオをペアリングした結果を、異なるパラメータの組み合わせと両方の方法からの応答曲線の例とともに付録A.5に掲載している。このデータセットは、加速度センサのどれもがカメラと同位置にないため、挑戦的なものである。センサとカメラの間の頻繁な相対的な動きは、ベースライン法に課題をもたらします。表S2から、PCAがベースラインの性能を向上させることがわかります。これは、おそらくセンサの相対的な動きに直面してモダリティ間で対応する軸を特定することによってです。我々のウィンドウ重み付けカーネル密度推定アプローチは、異なる加速度センサ位置のすべてにおいて一貫して性能をさらに向上させます。これは、「信号集約的」な時間的窓に自動的に焦点を当て、同期の効用が少ない窓を無視する我々の方法の能力を実証しています。

**5.4.4 Discussion**

CMU-MMAC データセットの場合、一定のサンプリング レートでデータを収集し、キャプチャ中にデータフレームを落とさないため、 データ品質の低下による問題を心配することなく、大きなウィンドウサイズ 𝑇𝑤 を選択することができ、𝑇𝑚𝑎𝑥 には経験的に合理的な値を設定している。同時に、このデータセットに対するすべての手法の性能は、S2S-Syncデータセットよりも低く、これは、カメラと加速度センサの体上での位置に起因する、カメラと加速度センサの間の大きな相対運動が、同期タスクを非常に困難なものにする役割を果たしている可能性を示唆している。付録A.5の実験は、パラメータの選択がデータセットに依存していることをさらに示している。同じデータセットであっても、異なる位置にある加速度ピックアップでは、異なるパラメータの選択によって最高の性能が得られます。

## 6. Discussion, Conclusion and Future Work

複数のデータストリームを正確に同期させる能力は、マルチモーダルなウェアラブルセンサースイートから日常の人間の行動を検出するための計算モデルを開発し、検証しようとする研究者にとって、長年の課題であり、その有用性はますます高まっています。具体的には、視覚的確認はターゲットイベントをラベル付けするために広く使われている方法であるため、隣接するセンサストリームとビデオを同期させるための正確で実行可能で負担の少ない方法が緊急に必要とされている。 今回の研究では、ウェアラブルビデオカメラからのデータをウェアラブル加速度センサからのデータと同期させる新しいアプローチ（SyncWISE）の実現可能性と有効性を実証しました。SyncWISEは、ウェアラブルセンサーから収集されたデータを含む実世界の同期タスクを特徴づける、部分的な観測性（センサーが同じイベントを捉えていない）と座標変換（センサーの軸が時間的に空間的に整列していない）の問題を解決します。我々は2つのデータセットでSyncWISEを評価しています：新しい禁煙データセットS2S-SyncとCMU-MMACデータセット[16]です。最近のベースライン手法と比較して、最先端の性能を示した。 我々の研究は、ウェアラブルカメラと胸に装着した加速度センサ間の時間同期問題の解決に焦点を当てているが、アルゴリズムの設計は、人体のさまざまな部分の他のセンシングモダリティにさらに適応させることができる。そうすることで、ビデオから得られたラベルを多様なウェアラブルセンサ信号に時間的に整列させることが可能になります。野生環境から得られるこのような時間的に正確なラベルは、飲食、歯磨き、歯磨き、歯磨き、歯磨き、喫煙など、日常のさまざまな行動の根底にある微細なマイクロビヘイビア（例えば、手と口のジェスチャーのダイナミクス）を検出する精度を向上させることができます。日常の行動を検出するための計算モデルの発展に加えて、自然界での微細な行動の詳細な観察は、人間の日常行動の理解を深めることができます。最後に、我々の新しいS2S-Syncデータセットを研究コミュニティに提供し、野生での時間同期のベンチマークとして利用していただきます。