修 士 論 文

食材の位置と撮影者の動きに着目したCNNの融合による一人称調理映像の自動要約

平成30年度修了

三重大学大学院　工学研究科 情報工学専攻
ヒューマンインターフェース研究室

島田　尚宜

はじめに

一人称視点の映像は，今日多くの場面で目にする映像コンテンツである．例として，映画，ゲーム，Virtual Reality (VR)，そしてウェアラブルカメラが挙げられる．これらの例からも確認できるように，これまではアミューズメント用途が主な利用場面であった．しかし近年，これを教育目的に用いる試みがなされている [1]．一人称映像は作業従事者の視点で撮影した作業内容を再生できるので，様々な技能の伝達に効果的であると考えられる．

この特性を活かしてインターネットでは一連の調理動作を擬似的な一人称映像で提供するサービスも運営されている．しかし，そのような映像作成のためには膨大な時間と手間が必要である．なぜなら，現時点ではそれらの調理動画は固定カメラによる専用の環境をセットアップして撮影されるからである．しかしウェアラブルカメラで実際の調理場面を撮影すれば，料理の技能 (コツ) や段取りといった現実的な技能の記録，伝達に活用できる．

一方で，作業現場を一人称カメラで撮影する場合，録画の頻繁な一時停止は煩雑であるため，一般的には長時間にわたって連続した動画が撮影される．そのため映像が冗長となり，そのままでは必要な情報を取得しにくいという問題がある．この問題の解決策として，所望の長さに縮めて再生する「映像要約（ダイジェスト化）」が有効であるが，人手で要約を行うことは手間を要する．そこで本研究では，一人称調理映像の自動要約を目的として，撮影された一連の調理動作から，調理内容，技能の伝達に有効な映像を取り出す手法を提案する．

調理動作は，動作の特徴が手元に表れる細かい動作である．そのため各動作が似ており，他の動作との区別も困難である．この特性は，一人称調理映像の要約を困難にする．一人称映像自動要約に関しては，オプティカルフローベクトルの大きさを用いたダイジェスト化の手法 [2] が提案されている．この研究では美術館における鑑賞体験を撮影した一人称映像を対象としていた．しかし本研究で対象とする調理動作映像は，調理動作と類似する皿洗いのような動作を含むため，オプティカルフローベクトルの大きさだけを用いて効果的なダイジェスト化を行うのは困難である．調理動作映像を要約する場合，食材が

写っていて，かつ手の動きがあるフレームを抽出すればよいという事前知識を活用できると考えられる．

そのため，一人称調理映像の各フレームから食材と腕の位置, そして撮影者の動きを表す 2 種類の画像を生成し，それらと一人称調理映像フレームを合わせた 3 種類のフレーム画像を入力とする Convolutional Neural Network (CNN) を提案し，フレーム単位で要約映像に残すべきか否か識別を行った．また提案手法の性能評価実験のため，データセットを作成した．評価基準にはフレーム単位の再現率・適合率の調和平均である F 値を用いた．本データセットを用いて実験を行った結果，提案した CNN によって 65.61%の F 値で要約できることがわかった．また，一人称調理映像フレームを単独で学習した CNN の結果である 7.11%に対して，食材位置を学習させた場合は 9.79%，撮影者の動きを学習させた場合は 25.85%，食材と腕の位置を学習させた場合は 48.93%，食材位置と撮影者の動きを学習させた場合は 51.65%，食材と腕の位置, そして撮影者の動きを学習させた場合は 60.92%，一人称調理映像フレームと撮影者の動きを学習させた場合は 63.03%，提案したネットワークでは 65.61%の F 値で自動要約ができた．提案する自動要約でポイントとなる調理動作を保持しつつ映像の長さを縮めることに成功したため，F 値 65.61%でも調理方法を理解できる十分な F 値といえる．

本論文では，1 章で研究背景と目的，2 章で関連技術，3 章で提案手法，4 章で提案手法の性能評価実験とその結果，5 章でまとめと今後の展望について述べる．

目次

第 1 章

緒言

1.1　研究の背景

近年，頭部に装着したカメラで撮影した映像を解析して，行動支援を実現する First Person Vision [3] 技術が注目されている．さらに最近は，カメラの小型化，高性能化により，一人称視点映像撮影は容易かつ手軽になっている．その結果，我々は日常の様々な場面で一人称映像コンテンツを見ることができる．映像コンテンツの例として，映画，ゲーム，Virtual Reality，そしてウェアラブルカメラが挙げられる．とりわけ，ウェアラブルカメラは手軽に一人称映像を撮影できる特性から，日常的に利用されつつある [3]．GoPro [4] や Google Glass [5]，またマイクロソフトの SenseCam [6] などのウェアラブルカメラが一般に普及すれば，我々は一人称視点の映像を日常的に撮影する多くの機会を得る．最近では，これらのカメラで撮影した一人称映像を SNS 上に投稿する利用者も増えている．また，これを教育目的に用いる試みがなされている [1]．一人称映像は撮影者の視点で映像を保存でき，あらゆる技能の伝達に効果的なためである．

調理は，日常的な行為でありながら，その習得には多くの知識を要する．そのため，以前から調理支援の需要は多く存在する．インターネットには Cookpad (クックパッド) [7] や Tasty [8] のようなレシピ共有サイトが存在し，多くの利用者が日常的に利用している．これらの Web サイトサービスでは，テキストベースの調理レシピに加えて，調理の過程を短く効果的に要約した動画が掲載されている．このような動画は，視覚的に調理を学ぶことができるため，技術習得に効果的である．しかし，そのような要約映像作成のためには膨大な時間と手間が必要である．なぜなら，現時点でそれらの調理動画は専用にセットアップされた環境で撮影されており，撮影した映像を最初から最後まで人が閲覧して内容を確認することで要約を行うためである．

そこで，ウェアラブルカメラで実際の調理場面を撮影すれば，撮影準備に手間をかけることなく，料理の技能 (コツ) や段取りといった現実的な技能の記録，伝達に活用できる．一方で，作業現場を一人称映像で撮影する場合，録画の頻繁な一時停止は煩雑であるため，一般的に長時間にわたって撮影される．そのため映像が冗長になり，そのままでは必要な情報を取得しにくいという問題が残ってしまう．この問題の解決策として，再生時には所望の長さに自動的に映像を縮めて再生する「自動映像要約 (自動ダイジェスト化) 」が有効であると考える．

1.2　一人称映像要約

一人称映像要約とは，一人称視点で撮影された映像から，冗長なシーンや意味のないシーンを排除し，盛り上がったシーンやコンテクストの把握に必要なシーンだけを抽出する作業である．

1.2.1　一人称映像コンテンツにおける映像要約の有用性

1990 年台，Mann 氏がウェアラブルカメラを世界に発表してから多くの変化があった [9]．近年では，ウェアラブルデバイスの利用は研究者にとどまらず，その市場は拡大を続けている．2015 年にはウェアラブルカメラ出荷台数が約 580 万台を記録し，2023 年には 3320 万台に到達すると予想されている [10]．ここで，ウェアラブルカメラの使用例を示す．

- 法執行機関での日常業務：　アメリカでは 63 の主要都市の中で,43 都市の警察署がパトロール中にウェアラブルカメラを使用している [11]．またその映像はパトロール中に事件が発生した場合には，証拠として用いられる．しかしながら，将来的には要約アルゴリズムを用いて，危険な行動パターン検知への応用が期待されている．
- 教育ビデオ：　撮影者の視点で記録を残せるという一人称映像の性質上，他者への技術伝達や指示伝達に効果的なことは明らかである．例えば，新設のビルでの行き先案内，工場での組み立て作業あるいは調理といった幅広い利用先が考えられる．先に述べたように，一人称映像は長時間撮影される．そこで，情報量の多い教育ビデオを制作するためには，映像要約による冗長性の排除が必要不可欠である．

1.2.2　一人称映像の特徴とその要約における課題

Tan 氏の調査 [12] に基づき，以下に一人称映像の特徴を三人称映像と比較しながら示す．

- 撮影方法：　三人称映像は，撮影対象としている道具や行動，人物に対してズーム等を用いて，撮影対象を限定する．それに対して，一人称映像の撮影では撮影のターゲットは決められていないため，撮影者の頭部の動きが反映される．この特性は，一人称映像と三人称映像の大きな違いであり，これを用いた一人称映像要約手法 [13, 14, 15] が存在する．
- 冗長性：　三人称映像では，撮影時間・内容ともに限定的なものとなる．その一方，一人称映像撮影時はハンズフリーになるため，自分の意図しない要素も映像に含まれる．また撮影も比較的長時間にわたるため，映像が冗長になる．
- 映像の質：　三人称映像の撮影では，三脚等を用いて映像を安定させるため，ゆれやぼけを含みにくい．これに対して一人称映像では，頭部や胸部の動きに伴う揺れやぼけを含む傾向がある．

上記のような特徴を持つため，一人称映像要約に三人称映像要約手法を用いることは困難である．それに加えて，三人称映像要約手法では声の抑揚や拍手，ニュース字幕，バックグラウンド音楽など，映像に存在する特徴的な要素を用いて要約する [16] が, 一人称映像にはこれらの要素を常に含むとは限らないことも理由として挙げられる．また一人称映像に生じるゆれやぼけは，特徴抽出を困難にし，たとえ三人称映像要約手法を適応できる対象であったとしてもその性能低下は否めない [18].

一人称映像要約における課題は，三人称映像要約とは異なり，事前にコンテクストを把握しておくことが困難な点である．なぜなら，頭部・胸部にカメラを装着して撮影するため，その対象を限定できないからである．したがって、最適な一人称要約手法はコンテクストに依存しない手法である [19, 20]．その一方で，頭部・胸部にカメラを装着して撮影を行うことで，撮影者視点でのログが保存できる．したがって，撮影者の視線パターンやカメラの動きから要約を行うことができる利点がある [13, 14, 15, 21].

1.3 関連研究

一人称映像を用いた研究では，多種多様な要素を含んだ問題を解決するために，ここ数年で様々な取り組みが行われてきた．過去の研究では，物体認識 [22, 23, 24] や行動認識 [25, 26, 27, 28, 29, 30, 31, 32]，さらには映像の要約 [33, 17, 34] や社会的インタラクションの予測 [35] に関する研究等が存在する．

また，一人称映像特有の特徴に関連する興味深い研究分野として，映像のカテゴリー分け [36, 37] や重要なフレームの選択 [38, 39]，ハイパーラプス [40] が挙げられる．ハイパーラプスとは，一定間隔で一人称視点撮影した写真を動画にした映像である．比較的最近のトピックとしては，視線推定 [41] や装着者の識別 [42, 43] などに関する研究も行われている．

1.3.1 一人称映像要約に関する研究

一人称映像要約の研究では，三人称映像要約手法をそのまま採用することは難しい．そこで，一人称映像に対応した要約手法を開発すべく，近年様々な研究が行われてきた．以下に，Ana 氏の調査 [44] に基づき, その主な手法や要約結果の出力方法，そしてデータセットについて示す．

- 手法
  一般的に，低レベルの特徴（色ヒストグラムやオプティカルフローなど）を用いたボトムアップ型の手法が採用されている [19, 38, 45, 46, 47, 48]．色ヒストグラムとは，画像の赤成分 (R) , 緑成分 (G) , 青成分 (B) 各チャンネル毎の輝度及び色合い分布を示し, 各色が画像中に現れる頻度を示している. またオプティカルフローとは，連続するフレーム間での物体やカメラの動きをベクトルで表現したものである．トップダウン型の手法としては，教師あり学習 [13, 14, 17, 18, 20, 34, 49, 50] や脳波（Electroencephalogram ; EEG ）信号 [51, 52]，視線 [15] を用いた手法が提案されている．
- 出力方法
  要約を行うとき，3 種類の出力方法が考えられる．その出力方法を以下に示す．
  1. Story board: この方法は一人称映像から抽出した数枚から数十枚の画像で構成される出力である [17, 18, 46, 53]．主にライフログが入力となる場合に使用され，その入力は映像ではなく，胸部に取り付けたカメラで撮影された一連の写真である [54]．ライフログとは，人間の生活や行動をデジタルデータとして記録したものを示す．

2. Skim: この方法は一人称映像中の重要であると判断できる数個から数十個の連続フレームの塊で構成される出力である [13, 14, 19, 20, 21, 34, 45, 47, 48, 49, 50, 51, 52, 55]．この出力方法は，ヘッドマウンテッドカメラで撮影された休日やホームビデオが入力の場合に使用される．
3. Fast-forward: この方法では，一人称映像の重要でない部分は早送り再生で出力される [38, 40, 56]．この出力方法は，映像がノーカットであることが望ましいエクストリーム・スポーツのような動きの激しい映像が入力である場合に使用される [38, 40, 56]．

- データセット

一人称映像研究に使用できるデータセット数は，顔認識や行動認識などの分野で利用されているデータセット数に比べて少ない．しかし，2010 年以降，その数は増加している．以下に，一人称映像要約研究で利用可能なデータセットを示す．

1. UT Egocentric [17]: 重要な物体や人物に着目した一人称映像要約手法評価のため，撮影されたデータセットである．フレームレートが 15fps と低い点で，他のデータセットとは異なる．
2. Activities of Daily Living [29]: 胸部に装着した GoPro カメラを利用して，事前に用意された一連の動作を行う様子が記録されたデータセットである．物体とそれに対するインタラクション，動作のそれぞれにアノテーションがつけられている．
3. GTEA-gaze+ [26]: このデータセットはキッチンで 7 種類のレシピを調理した様子が収録されている．撮影には SMI eye-tracking カメラを使用しており，100 種類の異なる動作に対してアノテーションがつけられている．
4. Disneyworld [35]: 一日を通して，遊園地で行った社会的インタラクションを評価するために作成されたデータセットである．しかしながら，このデータセットが持つ制約のない映像の性質，ビデオの長さ，そして [57] によって与えられたテキストベースのアノテーションは，一人称映像要約研究のデータセットとして有効である．
5. VideoSet [57]: このデータセットは，UT Egocentric と Disneyworld の一人称映像に対して，要約のアノテーションが付与されている．
6. Huji EgoSet [15]: あらゆる種類のアクティビティ，場所，照明設定で，GoPro カメラを用いて撮影された 40 の映像が収録されている．それに加えて，YouTube 上の日常的アクティビティ（ドライブ，散歩，スキー，乗馬など）を収録した 37 の映像が含まれている．
7. Microsoft's sports dataset [40]:　3 つの要約出力形態の一つ Fast-Forwarding を評価するために，ヘルメットに取り付けた GoPro カメラで山登りやサイクリ

ングを撮影したものである．

8. EgoSum+gaze [21]: 日常的な生活を視線情報とともに記録した初めてのデータセットであり，視線情報に基づいた要約手法を評価するために作成された．アノテーションは撮影者と外部の専門家によって与えられたもので，ビデオごとに 5 個から 15 個の出来事で構成されている．
9. Microsoft's video highlights [50]: YouTube から抽出した合計 100 時間・15 種類のスポーツの映像で構成される．一人称映像のみを抽出するため，すべての映像は YouTube 上で「スポーツの種類（スキー，サイクリングなど）+ GoPro」の検索で発見されたもののみを使用している．それらの映像は，それぞれ 2 分から 15 分の長さで，5 秒毎に映像の興味深さを示す 3 段階ラベルがつけられている．

1.4 本研究の取り組み

本研究の目的は，一人称映像として撮影された一連の調理動作から，調理内容，技能の伝達に有効な映像を自動的に取り出す手法の開発である．

一人称調理映像は，いくつもの調理動作とその類似動作（皿洗いなど）から構成されている．調理動作は，その特徴が細部に表れる細かい動作であり，調理中に行われるその他の動作との判別が困難である．これら特性は，一人称調理映像の要約を難しくする．

ミュージアムの鑑賞体験を記録した一人称映像の要約では, オプティカルフローベクトルを用いた要約手法 [2] が提案されている．しかし先に述べたように, 本研究で扱う映像には調理映像と類似する別の動作を含むため，オプティカルフローベクトルのみを用いた手法では効果的な一人称調理映像要約は期待できない．しかし調理動作映像を要約する場合，食材が写っていて，動きのあるフレームを残せば良いという事前条件を活用できると考えられる．

本研究では，研究目的を実現するため，一人称調理映像の各フレームから食材と腕の位置, そして撮影者の動きを表す 2 種類の画像を生成し，それらと一人称調理映像フレームを合わせた 3 種類のフレーム画像を入力とする Convolutional Neural Network を提案する．

第 2 章

本研究に関連する技術

本章では，提案手法で用いた，一人称調理映像要約手法を構成する要素技術について述べる．

2.1　ディープラーニング（深層学習）

近年，ディープラーニングという言葉を，様々な場所で目にする．ディープラーニング技術の発達によって，画像認識や音声認識の性能が爆発的に向上し，様々な場面で実用化されるようになった．

ディープラーニングは機械学習手法の一つであり，ニューラルネットワークと呼ばれる脳の神経ネットワークを模したアルゴリズムを用いる．ディープラーニングは，ディープニューラルネットとも呼ばれていて，ディープは「深い層で構成される」を意味する．つまり，ディープラーニング（ディープニューラルネット）とは，「深層学習（深い層で構成されるニューラルネットワーク)」という意味になる．

2.1.1 Neural Networks: ニューラルネットワーク

ニューラルネットワークはニューロンモデル（ユニット，ノード）を多層的に結合したモデルである．ニューロンモデルとは，脳のニューロン（神経細胞）をモデル化したものである．

ニューロン（神経細胞）

ニューロンは複数の受信器（樹状突起：dendrite）と一つの送信器（軸索：axon）で構成され，軸索上を伝わる電気パルスによってその他のニューロンへと情報が伝達される．軸索は，シナプスと呼ばれるインターフェイスを介して，パルスの到来をニューロンに伝達する．

ニューロンは電子パルスを受け取ることで細胞内の電気的レベル（膜電位）が上下する．この変動は，入力を受け取るシナプスの状態（シナプス伝達強度）に依存する．そして，膜電位の値がある一定の値を超えると，その電子パルスは発信され，軸索を通して他のニューロンに伝達される．

ニューロンモデル

ニューロンを単純な数理化モデルで表したものをニューロンモデルと呼ぶ．いま，$x_1, x_2, \ldots, x_n$ をニューロンへの入力，$w_1, w_2, \ldots, w_n$ をシナプス伝達強度，b をバイアス，z をニューロンの出力とする．出力 z は次のニューロンへの入力となる．ニューロンの入出力関係は以下のようになる（図 2.1）．

$$y = \sum_{i=1}^{n} w_i x_i + b \tag{2.1}$$

$$z = f(y) \tag{2.2}$$

このとき，$f()$ は非線形関数であり，活性化関数と呼ばれる．活性化関数の役割は，ニューロンの応答に非線形性を与えることである．以下に，いくつかの活性化関数を示す（式 (2.3)，式 (2.4)）．

- ステップ関数

  ステップ関数は閾値を境に入力が切り替わる関数であり，階段関数とも呼ばれる．

$$1[x] = \begin{cases} 1 & (x \geq 0) \\ 0 & (otherwise) \end{cases} \tag{2.3}$$

- シグモイド関数

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{2.4}$$

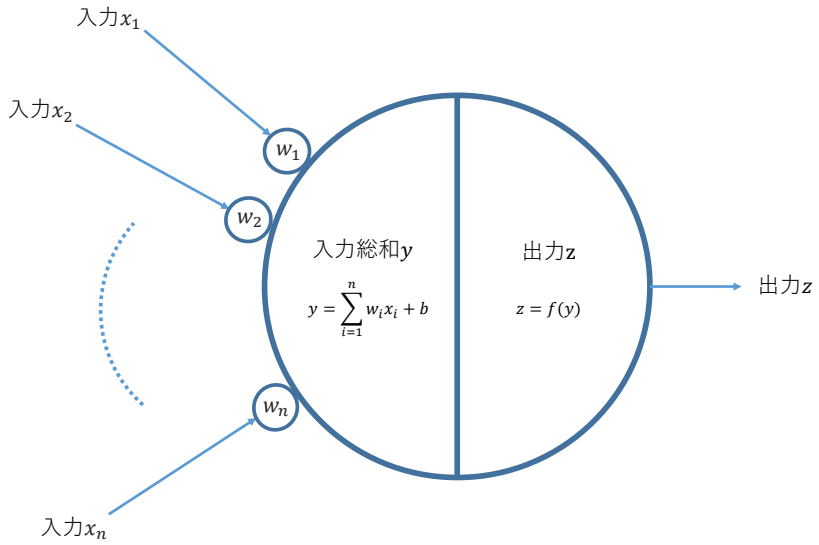


図 2.1: ニューロンモデル

Feedforward Neural Networks: 順伝播型ニューラルネットワーク

順伝播型ニューラルネットワークは典型的な深層学習モデルであり，多層パーセプトロン（multilayer perceptrons, MLPs）とも呼ばれる．このモデルは入力 x から出力 y へと一方向のみに信号が流れるため順伝播（feedforward）と呼ばれる．

このネットワークは，他のネットワークの基礎となっており，写真の中の物体認識に用いられる畳み込みネットワークは多層パーセプトロンの特殊な場合である．畳み込みネットワークについては次節で説明する．

多層パーセプトロンは，3 種類のニューロンモデルから構成される．それぞれ，入力ニューロン，隠れニューロン，出力ニューロンと呼ばれる（図 2.2）．

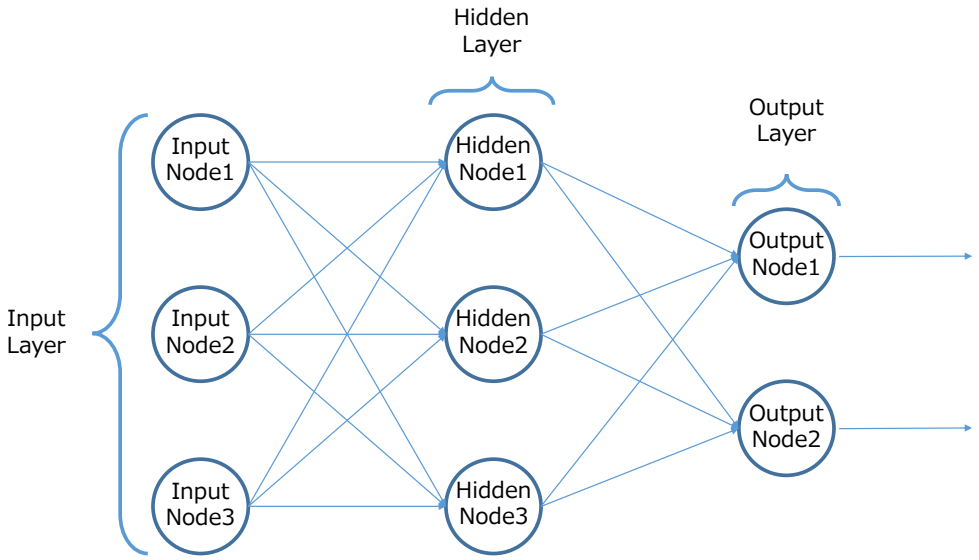


図 2.2: 多層パーセプトロン

2.1.2 Convolutional Neural Networks: CNN

近年，畳み込みネットワーク（convolutional networks）とも呼ばれる畳み込みニューラルネットワーク（convolutional neural networks，CNN）は，画像認識や動画認識の分野で高い性能を発揮し，注目を集めている．この技術は 1990 年代初期から文字認識分野で使用されていた [58]．現在のブームは物体認識技術を競う”ImageNet large-scale visual recognition challenge 2012”において CNN を用いた手法 [59] が従来手法（エラー率 26%）に比べて 17%のエラー率改善を遂げたことがきっかけである．

畳み込みニューラルネットワークの構成

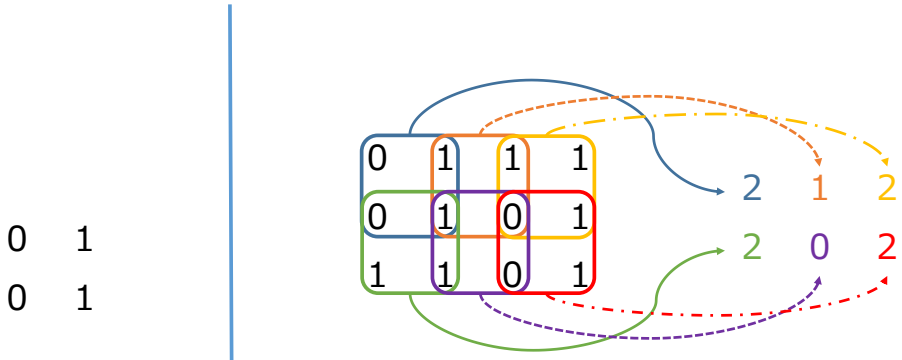
CNN は，いくつかの畳み込み層とプーリング層，全結合層（多層パーセプトロン），出力層で構成される．その入力は通常，3 階のテンソル型である．この入力に対して，畳み込み層での特徴マップ生成，プーリングを交互に行い，前層からの入力情報を次の層へ伝播させる．特徴マップが含む入力から抽出された形状特徴を，プーリング処理によって縮小しつつ，上位の層へ伝播が可能なモデルである．その後，全結合層・出力層によってモデルが推定した事後確率ベクトルを出力する．各層の詳細な機能については以下で示す．

入力層（input layer)

サイズ $W \times H \times D$ の画像を入力する．

畳み込み層（convolution layer)

畳み込み層では，画像の一部とフィルターの要素積の和を，画像をスライドさせながら画像の全領域で求める．例えば 3×4 の入力画像，2×2 のフィルターの場合，図 2.3 のような出力結果になる．



（a）フィルター　　（b）畳み込み処理の入力と出力

図 2.3: 畳み込み処理

3 階のテンソル型が入力の場合も同様に定義できる．l 層目の入力画像サイズ $W^l \times H^l \times D^l$ の入力 x^l に対しての畳み込み処理は式 (2.5) で定義できる．

$$y_{i^{l+1},j^{l+1}} = \sum_{p=1}^{H}\sum_{q=1}^{W} f_{p,q} \times x^l_{i^{l+1}+p,j^{l+1}+q} \tag{2.5}$$

ここで，$0 \leq i < H$，$0 \leq j < W$，$y_{i^{l+1},j^{l+1}}$ は $l+1$ 層目で出力された特徴マップ y の i 行 j 列目の要素，$f_{p,q}$ はフィルタの p 行 q 列目の要素を表す．

この特徴マップに対して式 (2.6) で示す ReLU（rectified linear unit）関数を通して，Pooling 層に入力する (図 2.4)．

$$z_{i,j} = \max(0, y^l_{i,j}) \tag{2.6}$$

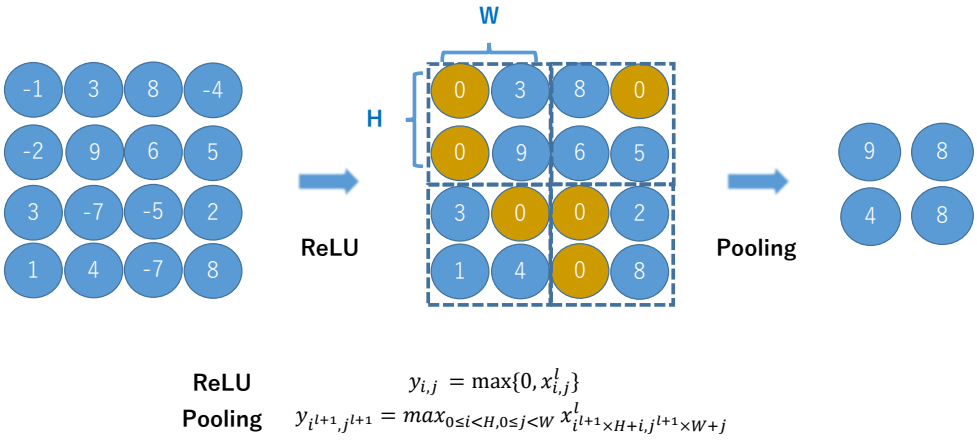


図 2.4: ReLU 関数とプーリング処理

最大プーリング層（max pooling layer）

畳み込み層によって，画像が持つ特徴を利用することができたが，画像の位置ずれに対して頑健でない．そのため，各特徴マップにおいて，局所領域内の最大値を取り出し，1つのノードへ置き換える処理である．例えば，特徴マップサイズが 4×4，局所領域サイズが 2×2 の場合，図 2.4 のようになる．

また l 層目のフィルタサイズ $W^l \times H^l$ の入力 x^l に対してのプーリング処理は式 (2.7) で定義できる．

$$y_{i^{l+1},j^{l+1}} = \max_{0\le i<H,0\le j<W} x^l_{i^{l+1}\times H+i,j^{l+1}\times W+j} \tag{2.7}$$

ここで，$0 \le i < H$，$0 \le j < W$，$y_{i^{l+1},j^{l+1}}$ は $l+1$ 層目の出力 y の i 行 j 列目の要素を表す．

全結合層（fully connected layer）と出力層（output layer）

全結合層とは，多層パーセプトロンを表す．最後のプーリング層の後，全結合層を接続し，出力層へとつなげる．

交差エントロピーによる誤差関数

k クラス分類では，勾配降下法の誤差関数に以下の式 (2.8) の交差エントロピーを用いる．

$$E(w) = -\sum_k t_k \log y_k \tag{2.8}$$
$$y = f(wx)$$

ここで，t は目的とする正解ベクトルで，正解に対応するクラスの値を 1 とし，その他は 0 とする．y はネットワークが出力した事後確率ベクトルである．x はネッワークに対する入力，関数 f は活性化関数を表す．

勾配降下法による学習

構成したネットワークは，式 (2.9) で表される誤差関数を最小化するような重み w を学習し，最適なモデルを生成する．誤差関数を最小化する重み w を求めるため，誤差関数の勾配に基づいた重み w の更新（勾配降下法）をする必要がある．

$$E(w_t) = \frac{1}{N_t}\sum E(w) \tag{2.9}$$

$t+1$ 回目の更新式は t 回目の更新を終えた誤差関数 $E(w_t)$ を用い，式 (2.10) で定義される．

$$w_{t+1} \leftarrow w_t - \eta\frac{dE(w_t)}{dw_t} + \alpha\Delta w_t \tag{2.10}$$

ここで，η は更新率である．また α は慣性項のパラメータであり，前回の更新量に α 倍して加算することでパラメータの更新をより慣性的なものにするという役割がある．

勾配降下法にはミニバッチ学習法を使用する．ミニバッチ学習とは，1 回の重み w を更新するために，学習データの少サンプル（ミニバッチ）の平均誤差を用いる学習法である．学習データからランダムにデータを選択し，N_t 個のミニバッチを作成する．式 (2.9) によって，N_t 個のデータに対する平均誤差を求め，式 (2.10) によって，ネットワークの重み w を更新する．この更新は学習データを使いきるまで行い，この一連の流れを 1 エポックとする．次回のエポックでは，新たにミニバッチをランダムに生成し，同様に重み更新を行う．

2.1.3　Fully Convolutional Networks: FCN

Fully Convolutional Networks（FCN）[60] はセマンティックセグメンテーション分野で使用される特殊な CNN である．セマンティックセグメンテーションとは，画像内の各画素をクラス分類することである．

CNN は主に畳み込み層と全結合層で構成されるが，FCN では，その名の通り，畳み込み層と逆畳み込み層のみで構成される（図 2.5）．

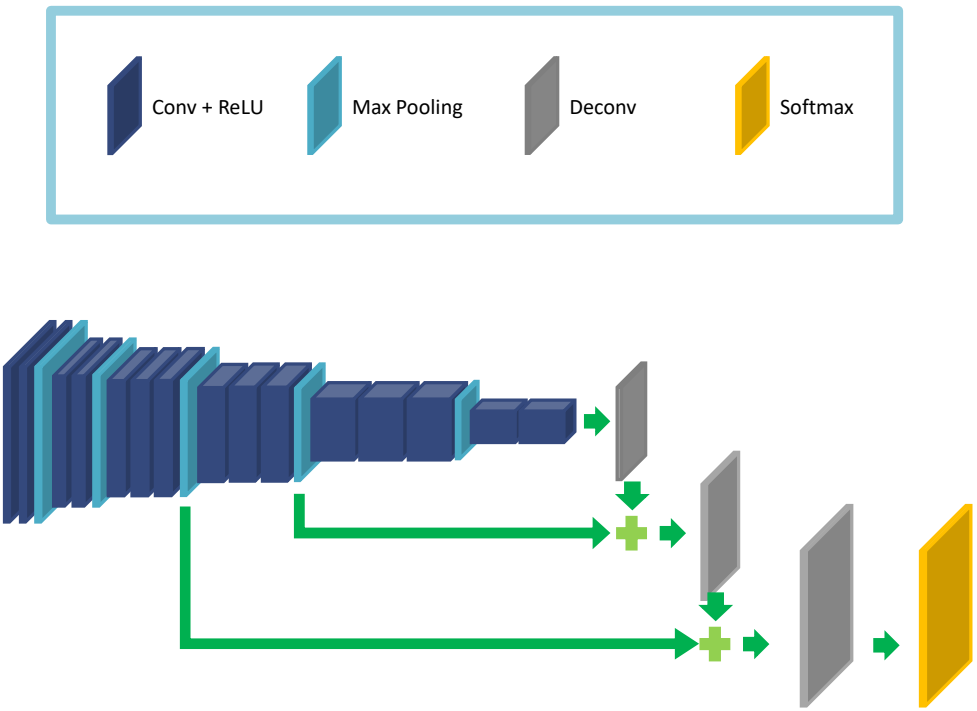


図 2.5: FCN の構成

2.2 オプティカルフロー

オプティカルフローとは，連続する 2 フレーム間におけるカメラや物体の動きをベクトルで表したものである．

時刻 t におけるフレーム内の点 (x, y) における明るさを $f(x, y, t)$ とする．いま，微小時間 Δt の間にその点の座標が微小距離（$\Delta x, \Delta y$）だけ変化したとする．このとき，この点の明度が変化しないと仮定すると，以下の式 (2.11) が成り立つ．

$$f(x, y, t) = f(x + \Delta x, y + \Delta y, t + \Delta t) \tag{2.11}$$

右辺をテイラー展開し，Δx，Δy，Δt の 2 次以上の項を無視すると以下の式 (2.12) が成り立つ．

$$f(x, y, t) \cong f(x, y, t) + \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial t}\Delta t \tag{2.12}$$

この式の両辺を $\Delta\, t$ で割ると以下の式 (2.13) が導出される．

$$\frac{\partial f}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial f}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial f}{\partial t} = 0 \tag{2.13}$$

ここで，フレーム上での空間的な明るさの勾配を f_x，f_y，時間的な明るさの変化率を f_t，x，y それぞれの単位時間あたりの移動量（オプティカルフロー）を $\boldsymbol{v} = (v_x, v_y)$ という形でまとめると，次のように書ける．

$$f_x v_x + f_y v_y + f_t = 0 \tag{2.14}$$

式 (2.14) を，以下の式 (2.15) のように変形すると

$$v_y = -\frac{f_x}{f_y}v_x - \frac{f_t}{f_y} \tag{2.15}$$

となるが，これは (v_x, v_y) が 1 つの直線上に拘束されることを表しているだけである．そのため，実際に得られるフレームの時空間微分からこの式 (2.14) だけを用いてオプティカルフロー $\boldsymbol{v}$ を求めることは不可能である．この問題は Aperture Problem(窓問題) という名前で知られている．

今日まで，前述の窓問題を解決するため多くのオプティカルフローを推定手法が提案されてきた．本研究では，ピクセル単位でのオプティカルフロー推定手法としてよく知られており，OpenCV の API として実装されている Farneback 法 [61] を用いた．以下に，その概要について述べる．

Farneback 法では，各画素の明るさを 2 次の多項式で近似し，連続する 2 フレーム間でその多項式の係数比較を行い，オプティカルフローを推定する．時刻 t における座標 x の近傍領域の明るさを 2 次多項式で表し，その係数を重み付き最小 2 乗法で最適化し，以下の式 (2.16) で表す．

$$f_t(x) = x^T A_t x + b_t^T x + c_t \tag{2.16}$$

ここで，A_t は 2 行 2 列の対称行列，b_t は 2 次の列ベクトル，c_t はスカラーである．また時刻 t における座標 x の点の時刻 $t+1$ までの移動量（オプティカルフロー）を d とする．このとき

$$\begin{aligned} f_t(x-d) &= (x-d)^T A_t (x-d) + b_t^T (x-d) + c_t \qquad (2.17)\\ &= x^T A_t x + (b_t - 2A_t d)^T x + d^T A_t d - b_t^T d + c_t \\ f_{t+1}(x) &= x^T A_{t+1} + b_{t+1}^T x + c_{t+1} \qquad (2.18) \end{aligned}$$

となる．また $f_{t+1}(x) = f_t(x-d)$ なので，係数比較すると以下の関係式を得る．

$$A_{t+1} = A_t \tag{2.19}$$
$$b_{t+1} = b_t - 2A_t d \tag{2.20}$$
$$c_{t+1} = d^T A_t d - b_t^T d + c_t \tag{2.21}$$

そのため移動量 d は以下の式 (2.22) で算出される．

$$d = -\frac{1}{2} A_t^{-1} (b_{t+1} - b_t) \tag{2.22}$$

第 3 章

提案手法

本章では，本研究で提案する要約システムの流れについて述べる．提案する要約システムの目的は，撮影された一人称調理映像から調理の内容とその方法理解に必要なフレームを抽出し，元映像より短時間の映像を再構成することである．なぜなら，作業現場を一人称カメラで撮影する場合，録画の頻繁な一時停止は煩雑であるため，一般的に長時間にわたって撮影され，映像が冗長となり，そのままでは必要な情報を取得しにくいからである．そこで，提案手法によって冗長なシーンを排除し，情報密度の高い映像を生成する．調理内容とその方法理解に必要なフレームを調理動作フレームとする．調理動作フレームの定義は，表 3.1 で示される調理動作を含むフレームである．

表 3.1: 調理動作の例

調理動作	具体例
混合	加える，混ぜる
加熱	炒める，焼く
切砕	切る，砕く
装飾	盛る，添える
浸漬	浸す，漬ける
冷却	冷やす，冷ます
ろ過	振る，絞る
その他	包む，溶かす

3.1　要約システム

要約システムの概要を図 3.1 に示す．提案する要約システムは，一人称調理映像から抽出したフレームを（一人称調理映像フレーム，Input0）入力として，そのフレームが調理動作フレームか否かを出力する．

提案手法は主に 2 つの処理で構成されている．

1. 特徴抽出：一人称調理映像要約に有効な 2 つの特徴（食材と腕の領域，そして撮影者の動き）を抽出する．
2. クラス分類：CNN を用いて，入力されたフレームに対する調理動作フレーム判定を行う．

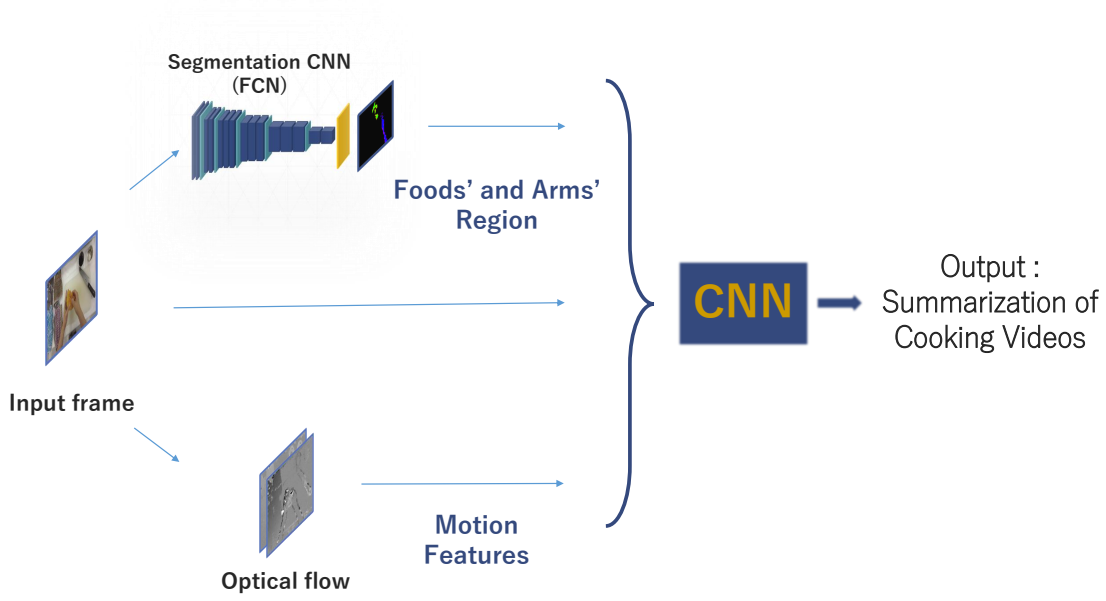


図 3.1: 要約システムの概要

3.1.1 特徴抽出

一人称調理映像から取り出した処理対象となるフレームから特徴量を抽出する．本論文では，以下の 2 つの特徴量を抽出した．

食材と腕の領域特徴

調理動作は任意の食材に対して視線を向けつつ行われることが多い．このことは一人称調理映像のフレームにおける撮影者の腕と食材の有無に関する情報が，要約映像生成の重要な手がかりとなることを示唆する．そのため，提案手法では一人称調理映像の各フレームに対して画素単位で食材と腕の領域判定を行う．

本研究では，FCN を用いて一人称調理映像の各フレームに対して撮影者の腕領域と食材領域のセグメンテーションを行った．まず学習データを作成するため，手動でピクセル単位の腕領域と食材領域のアノテーションを作成した．作成した学習データと FCN から得られた出力結果例を，図 3.2 から図 3.4 に示す．以降，この FCN で作成された画像は，調理動作領域画像（input1）と呼ぶ．

- 学習データ：3 つの異なるレシピ（小松菜のおひたし，豚の生姜焼き，かぼちゃの甘煮）を同時に調理した様子を撮影した一人称映像から食材を含むフレームに対して，アノテーションを作成した．学習で使用した画像は合計 1050 フレームである．撮影には，ウェアラブルカメラ Pivothead [62] を使用した．
- FCN アーキテクチャ：7 層の畳込み層と 3 層の逆畳み込み層で構成されたネットワークモデル（FCN）は，全結合層の部分を 1×1 サイズのフィルタを用いた畳込み層と考え，改良が加えられたセマンティックセグメンテーション専用のネットワークモデルである．本研究では，Long 氏らによって提案された FCN-8s [60] を用いた．学習に用いた勾配法の学習率は，$\Gamma = 10^{-3}$ とする．エポック数は 100，バッチサイズは 10 とする．誤差関数には式 (3.1) で表される交差エントロピーを用いた．ピクセル単位で誤差を計算し，その合計値を用いて，勾配法により学習を行う．

$$D(y,t) = -\sum_{i}^{I} t_i \log y_i \tag{3.1}$$

ここで y はネットワークの最終層の出力，t は正解データ，I はクラス数を表す．

図 3.2: 元画像

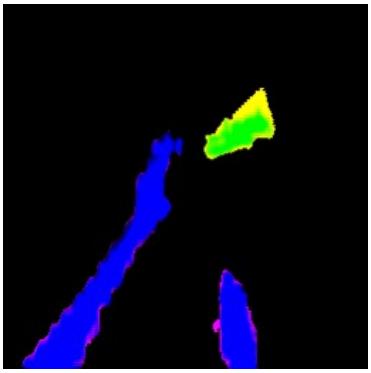


図 3.3: FCN の出力結果

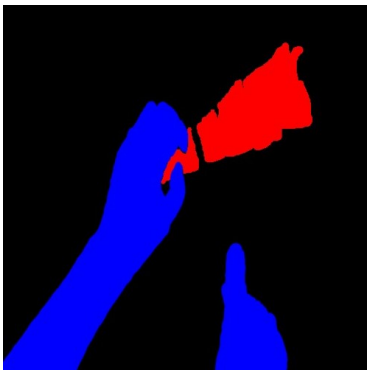


図 3.4: Ground-Truth

撮影者の動き特徴

カメラや映像中の動き，そして撮影者の頭の動きを CNN に学習させることが目的である．CNN に動き特徴を学習させるため，Simonyan 氏らが用いた方法 [63] に基づいて，連続する 2 フレーム間のオプティカルフローベクトルを算出する．そのベクトルを水平成分と垂直成分に分解し，2 枚のグレースケール画像で表す．この処理の詳細手順を以下に示す．

1. N フレームの映像 $I = \{I_1, I_2, I_3, \ldots, I_N\}$ に対して，連続する 2 フレーム間における各ピクセルのオプティカルフローベクトルを算出する [61].
2. オプティカルフローベクトルを水平成分 $U = \{U_1, U_2, U_3, \ldots, U_N\}$ と垂直成分 $V = \{V_1, V_2, V_3, \ldots, V_N\}$ に分解する．
3. 分解した成分において，[-20, 20] の範囲に存在するオプティカルフローベクトルの大きさを [0, 255] の範囲でピクセル単位の正規化を行うことで，2 枚のグレースケール画像で表す．

以降，この画像をオプティカルフロー画像（Input2）と呼ぶ．作成したオプティカルフロー画像例を，図 3.5 から図 3.8 に示す．

3.1.2　クラス分類

抽出した特徴を用いてそのフレームが調理動作フレームか否かを分類する．提案手法では，この処理を 2 クラス分類として扱う．2 クラス分類器として，5 層の畳み込み層と 3 層の全結合層で構成される CNN（図 3.9）を提案する．このネットワークで各クラス（調理動作フレームクラスとそれ以外）の事後確率を推定し，事後確率が最大となるクラスを

図 3.5: フレーム 1

図 3.6: フレーム 2

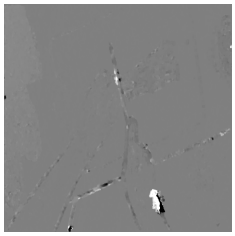


図 3.7: 水平成分

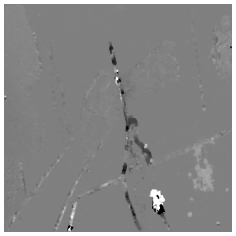


図 3.8: 垂直成分

入力されたフレームのクラスとしている．ネットワークの詳細な構成は次項で述べる．

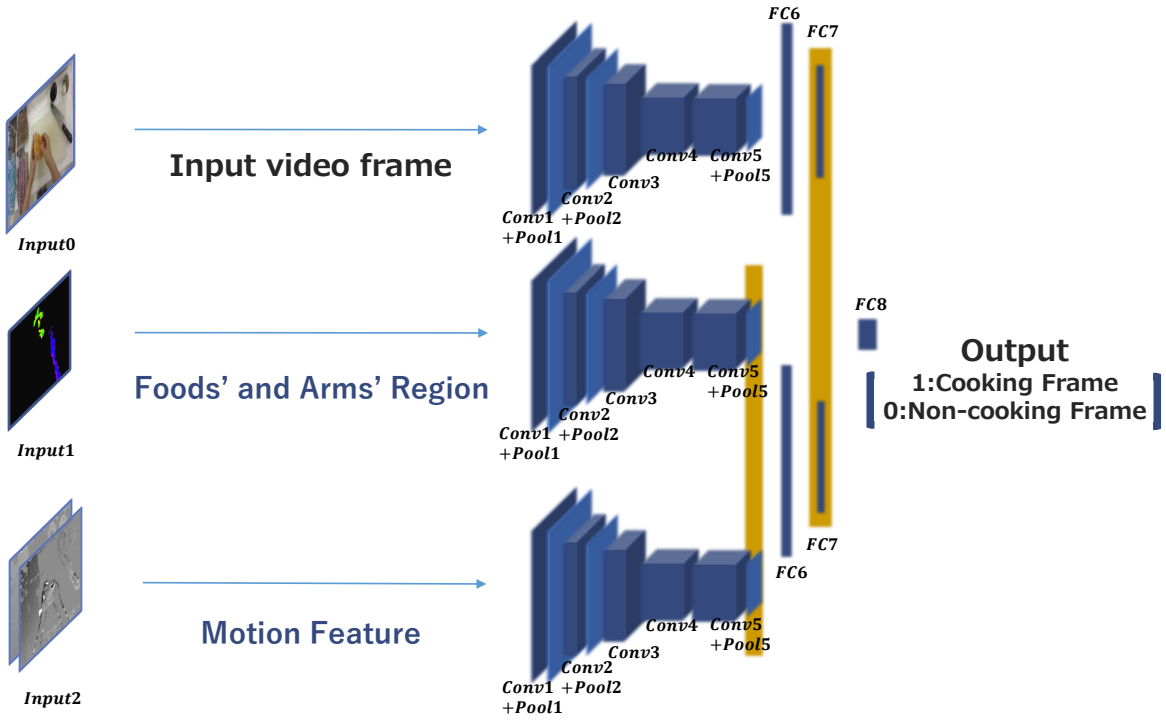


図 3.9: 提案手法の CNN アーキテクチャ

3.2 ネットワークアーキテクチャ

提案手法で用いる CNN の詳細な構成を表 3.2 に示す．ここで Input は入力画像であり，チャンネル数 × 縦 224× 横 224 で表す．また，Conv は畳み込み層，Pool はプーリング層，FC は全結合層をそれぞれ表す．

CNN は，フィルターを用いたニューラルネットワークである．フィルターに当てはめる数値によって，様々な画像処理ができるが，CNN では，そのフィルター自体を学習させることで，分類に必要な特徴を抽出できる．本研究では，活性化関数は Leaky Rectified Linear Unit (Leaky ReLU) を用いる．Leaky ReLU は式 (3.2) のように表される関数であり，畳み込み層の各フィルタに対する適合度合いが低いピクセルに対して 0.01 を乗算した値となる．そのため Leaky ReLU の出力に対して Max Pooling を行うことで，調理動作フレーム分類の特徴を保持したまま次元を削減できる．

$$f(x) = \begin{cases} x & (x > 0) \\ 0.01x & (x \leq 0) \end{cases} \tag{3.2}$$

その後，得られた特徴マップを次の畳み込み層に入力する．

提案するネットワークは，Input0，Input1 と Input2 をそれぞれ Conv1-Conv5 層で畳み込んだ後，Pool5 でプーリングを行う．Input1 と Input2 に対応する Pool5 が出力する 256 次元の特徴を連結して 512 次元の特徴として FC6 層に入力し，その出力を FC7 層へ入力する．Input0 に対応する Pool5 が出力する 256 次元の特徴はそのまま FC6 層に入力し，その出力を FC7 層へ入力する．その後，Input0 に対応する FC7 層が出力する 128 次元の特徴と Input1 と 2 に対応する FC7 層が出力する 128 次元の特徴を連結して 256 次元の特徴とし，FC8 層に入力する．

3.3 要約映像

提案手法で分類された調理動作フレームのみで再構成された映像を要約映像と定義する．要約システムの実践的な利用を考えた場合，その目的によって許容される映像の長さが異なる．そのため，ユーザが指定する時間長の要約映像作成は重要である．提案手法を使用した場合，以下のような手順で所望の長さに縮めた要約映像を生成できる．

1. 提案するネットワークを用いて，各フレームに対して要約映像に含めるか否かの 2 クラス分類を行う．
2. 指定する時間長に合わせて，その調理動作フレームクラスの事後確率推定値が高い

表 3.2: 提案手法で使用する CNN の詳細

Layer	Structure
Input0 (一人称調理映像フレーム)	$3 \times 224 \times 224$
Input1 (調理動作領域画像)	$3 \times 224 \times 224$
Input2 (オプティカルフロー画像)	$2 \times 224 \times 224$
Conv1	filter: $96 \times 3 \times 3$ 活性化関数: Leaky ReLU stride: 2, padding: 1
Pool1	種類: Max Pooling kernel: 2, stride: 2
Conv2	filter: $128 \times 3 \times 3$ 活性化関数: Leaky ReLU stride: 2, padding: 1
Pool2	種類: Max Pooling kernel: 2, stride: 2
Conv3	filter: $256 \times 3 \times 3$ 活性化関数: Leaky ReLU stride: 2, padding: 1
Conv4	filter: $256 \times 3 \times 3$ 活性化関数: Leaky ReLU stride: 2, padding: 1
Conv5	filter: $256 \times 3 \times 3$ 活性化関数: Leaky ReLU stride: 2, padding: 1
Pool5	種類: Max Pooling kernel: 2, stride: 2
FC6	ユニット数: 256，dropout: 0.9 活性化関数: Leaky ReLU
FC7	ユニット数: 128，dropout: 0.8 活性化関数: Leaky ReLU
FC8	ユニット数: 2, dropout: 0.7 活性化関数: Leaky ReLU

ものから順に要約映像を構成するフレームとする．

第 4 章

実験

本章では，まず性能評価実験で用いたデータセットと性能評価方法について述べる．続いて，提案する要約システムの性能評価実験について述べる．

実験では，提案する要約システムの有効性検証のため，一人称調理映像フレーム，調理動作領域画像，オプティカルフロー画像それぞれを単独で学習したネットワーク，あるいはそれらを学習したネットワークを融合させた場合の性能を比較検証する．実験で使用したネットワークの詳細な構成は後述する．

4.1　実験条件

4.1.1　データセット

本実験では，調理する様子を一人称カメラ Pivothead[62] で撮影した．このカメラはフレームレート 30fps，画角 77 度で撮影できる．調理者は 2 名，調理される料理とその食材，そして撮影環境は学習用データと評価用データで異なる．学習用データでは，豚の生姜焼き，かぼちゃの甘煮，小松菜のおひたしの 3 メニュである．評価用データでは，トマト，レタス，ブロッコリーを用いたサラダである．学習用データと評価用データ，それぞれにおけるフレーム数の内訳を表 4.1 に示す．また図 4.1 にデータセットのフレーム例を示す．

アノテーションは，各動画のフレームに 0 から順にフレーム番号を割り当て，調理動作フレームの開始フレーム番号，終了フレーム番号を目視で定め，列挙したテキストファイルを作成した．

表 4.1: 各データにおけるフレーム数の内訳

データ	全体のフレーム数	調理動作フレーム
学習用	10,300	5,150
評価用	14,890	8,100

(a) frames in training dataset

(b) frames in evaluation dataset

図 4.1: 学習用データセットと評価用データセットのフレーム例

4.1.2 評価方法

評価の指標を，以下の式 (4.1)〜(4.3) のように定義する．

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

$$\mathrm{F-measure} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{4.3}$$

まず，式 (4.1)，(4.2) により，精度 (Precision) と再現率 (Recall) を求め，この 2 つの値の調和平均である F 値を求める (式 (4.3))．F 値を評価値とする．式 (4.1)，(4.2) 中の TP，FP，FN は以下に示す値である．

- TP
  調理動作フレームを，提案したネットワークが正しく検出したフレーム数
- FP
  調理動作フレーム以外を，提案したネットワークが誤検出したフレーム数
- FN
  調理動作フレームを，提案したネットワークが検出しなかったフレーム数

4.2　提案する要約システムの性能評価実験

4.2.1　実験概要

提案手法では，調理動作領域画像（Input1）とオプティカルフロー画像（Input2），そして処理対象となるフレーム（Input0）を入力とする CNN を用いた一人称調理映像要約を行った．一人称調理映像要約に対する各特徴量の有効性について確認するため，各特徴を単独で CNN に学習させ，精度を検証する．実験条件を表 4.2 に示す．

表 4.2: 実験条件

学習用データ	フレーム数：10,300，調理動作フレーム数：5,150
評価用データ	フレーム数：14,890，調理動作フレーム数：8,100
画像サイズ	224×224 [pixels] 前処理として，画像を 256×256 [pixels] に収縮する． 学習時には，その画像から 224×224 [pixels] の領域を ランダムに cliping した画像を入力とする．
フレームレート	30 fps
分類器	CNN
分類器への入力画像の組み合わせ	後述
ネットワークの構成	後述
学習率	10^{-4}
バッチサイズ	103
エポック数	500

実験で使用したネットワークの構成とその入力画像の組み合わせを 3 章で示した表 3.2 を用いて，以下に示す．

実験 1：一人称調理映像フレームに着目したネットワーク

Input0 を Conv1-Conv5 で畳み込んだ後，全結合層へと入力する．全結合層では，まず FC6 層の出力を FC7 層へ，そしてその出力を FC8 層へ入力する．このネットワークの性能評価実験を実験 1 とする．

実験 2：食材領域に着目したネットワーク

Input1 は食材領域と腕領域を表した画像である．このネットワークでは，Input1 の代わりに食材領域のみを表した画像（以降，Input1′）を入力とする．このネットワークの性能評価実験を実験 2 とする．

実験 3：オプティカルフロー画像に着目したネットワーク

実験 1 で用いたネットワークの入力を Input2 に変更する．このネットワークの性能評価実験を実験 3 とする．

実験 4：調理動作領域画像に着目したネットワーク

実験 1 で用いたネットワークの入力を Input1 に変更する．このネットワークの性能評価実験を実験 4 とする．

実験 5：オプティカルフロー画像と食材領域画像に着目したネットワーク

Input1′ と Input2 をそれぞれ Conv1-Conv5 層で畳み込んだ後，全結合層に入力する．その後，FC7 層が出力する 128 次元の特徴を連結して 256 次元の特徴とし，FC8 層に入力する．このネットワークの性能評価実験を実験 5 とする．

実験 6：オプティカルフロー画像 + 調理動作領域画像に着目したネットワーク

実験 5 で用いたネットワークの入力を Input1′ から Input1 に変更する．このネットワークの性能評価実験を実験 6 とする．

実験 7：オプティカルフロー画像 + 一人称調理映像フレームに着目したネットワーク

実験 4 で用いたネットワークの入力を Input1′ から Input0 に変更する．このネットワークの性能評価実験を実験 7 とする．

実験 8：オプティカルフロー画像 + 一人称調理映像フレーム + 調理動作領域フレームに着目したネットワーク（提案手法）

3 章で述べた提案手法のネットワークを用いる．このネットワークの性能評価実験を実験 8 とする．

4.2.2　実験結果

前項で述べた実験 1 から実験 8 の Precision，Recall，F 値を表 4.3 に示す．ネットワークの性能は一人称調理映像フレーム（Input0）を学習させた場合に比べて，食材領域画像（Input1′）を学習させた場合は 7.11%（実験 1）→ 9.79%（実験 2），オプティカルフロー画像（Input2）を学習させた場合は 7.11%（実験 1）→ 25.85%（実験 3），調理動作領域画像（Input1）を学習させた場合は 7.11%（実験 1）→ 48.93%（実験 4），食材領域画像とオプティカルフロー画像（Input1′&2）を学習させた場合は 7.11%（実験 1）→ 51.65%（実験 5），調理動作領域とオプティカルフロー画像（Input1&2）を学習させた場合は 7.11%（実験 1）→ 60.92%（実験 6），一人称調理映像フレームとオプティカルフロー画像（Input0&2）を学習させた場合は 7.11%（実験 1）→ 63.03%（実験 7），3 種類の画像（Input0&1&2：提案手法）を学習させた場合は 7.11%（実験 1）→ 65.61%（実験 8）の F 値向上があった．

一人称調理映像要約性能における調理動作領域画像の寄与度に対する考察

食材領域画像に着目した場合の 2.68%の F 値向上に対し，調理動作領域に着目した場合は 41.82%の F 値向上があったことから，一人称調理映像要約には食材領域だけでなく，食材と調理者の腕の位置関係が重要であることが分かる．調理動作が行われる際，食材と調理者の腕の位置関係はパターン化される．一人称映像であれば，撮影者が注目している対象（調理であれば，調理中の食材）はフレーム中央に位置する場合が多い．また，その対象に対してインタラクションが発生する場合には，撮影者の腕はフレーム下部，あるいは左右から映り込む形となる．以上の理由から，図 4.2 のように食材領域はフレーム中央部，腕領域はフレーム下部（あるいは左右）から映り込むパターンになる．そのため，食材領域に着目した場合では単に食材をフレーム内に捉えている場合と調理動作フレームを区別できなかったのに対し，調理動作領域画像に着目することでそれを実現できたため 40%以上 F 値が向上したと考えている．

一人称調理映像フレーム（Input0）を学習させた場合に比べて，調理動作領域画像（Input1）を学習させた場合は 41.82%の F 値向上となった一方で，調理動作領域画像とオ

プティカルフロー画像（Input1&2）を学習させた場合に比べて，一人称調理映像フレームとオプティカルフロー画像（Input0&2）を学習させた場合は 2.11%の F 値向上となった．

一人称調理映像フレームに着目した場合より，調理動作領域画像に着目した CNN の F 値は 41.82%高い．そのため，オプティカルフロー画像と一人称調理映像フレームに着目した場合より，オプティカルフロー画像と調理動作領域画像に着目した CNN のほうが F 値が高くなると予測される．しかし実験結果は予測と異なった．その理由は，図 4.3 のように調理動作領域を正しく判定できないフレームの存在が原因と考えられる．

一人称調理映像要約性能における調理動作オプティカルフロー画像の寄与度に対する考察

ネットワークの性能は一人称調理映像フレーム（Input0）を学習させた場合に比べて，オプティカルフロー画像（Input2）を学習させた場合は 18.74%の F 値向上があった．

一人称調理映像フレームは単一フレームに対し，オプティカルフロー画像は連続する 2 フレーム間からオプティカルフローベクトルを算出し，画像を生成するため，分類器に撮影者の動きを学習させることができる．一人称調理映像要約において調理動作の動きを学習せずに要約を行うことは困難である．そのため，分類器が撮影者の動きを学習することを可能にするオプティカルフロー画像を用いることで 18%の F 値向上が実現できた．

表 4.3: 各実験で使用したネットワークの要約精度，Precision，Recall，F 値

Input	Precision	Recall	F-measure
実験 1: Input0	34.87%	3.91%	7.11%
実験 2: Input1′	40.79%	5.56%	9.79%
実験 3: Input2	54.08%	17.96%	25.85%
実験 4: Input1	54.66%	44.29%	48.93%
実験 5: Input1′&2	62.82%	43.85%	51.65%
実験 6: Input1&2	57.34%	64.98%	60.92%
実験 7: Input0&2	57.45%	69.93%	63.03%
実験 8: Input0&1&2	57.45%	76.47%	65.61%

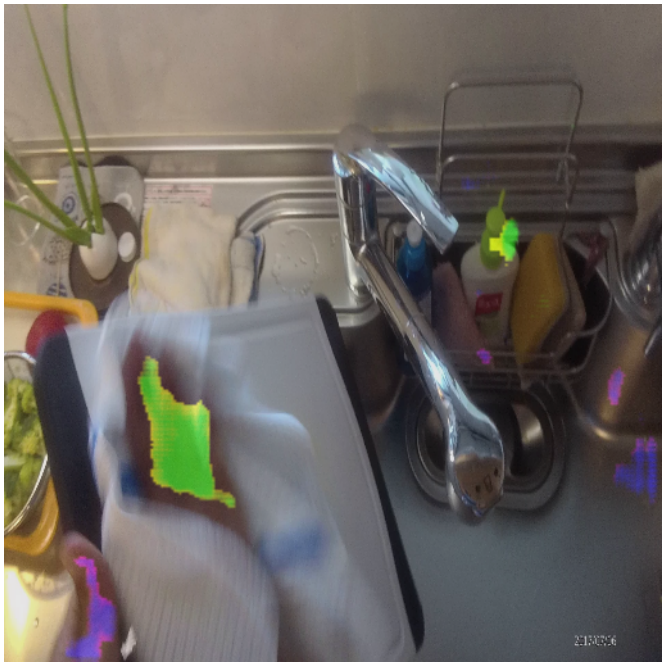
(a) 学習データセットの例 1　(b) 学習データセットの例 2

(c) 学習データセットの例 3　(d) 学習データセットの例 4

(e) 評価用データセットの例 1　(f) 評価用データセットの例 2

(g) 評価用データセットの例 3　(h) 評価用データセットの例 4

図 4.2: 調理動作フレームの画像例



(a) 例 1：手の甲と液体石鹸のボトル部分を食材と誤分類

(b) 例 2：スポンジと液体石鹸のボトル部分を食材と誤分類

図 4.3: 調理動作領域の判定失敗画像例（緑色ピクセルが食材領域，青色ピクセルが腕領域）

第 5 章

結言

5.1　まとめ

本研究では，一人称映像として撮影された一連の調理動作から，調理内容，技能の伝達に有効な映像を自動的に取り出す手法を提案し，学習用と評価用で異なる環境・食材で構成されるデータセットを用いて，提案手法の精度評価実験を行った．研究目的である調理内容・調理技能の伝達に有効な映像を取り出すことは，人間でも難しい試みである．なぜなら，要約映像に含むべき情報は多種多様であり，一意に決めることはできないからである．

実験的に一人称調理映像要約に食材と腕の位置，そして撮影者の動きが有効であると仮定し，一人称調理映像フレーム，調理動作領域画像，そしてオプティカルフロー画像の 3 種類を入力とする CNN を用いた要約システムを提案した．その結果，提案した CNN の F 値は 65.61%となった．提案する自動要約でポイントとなる調理動作を保持しつつ映像の長さを縮めることに成功したため，F 値 65.61%でも調理方法を理解できる十分な F 値といえる．

本研究で提案する要約システムは一人称調理映像に対して有効な手法であるが，着目する対象を変えれば他カテゴリーの映像にも応用できると考えられる．

5.2　今後の展望

本研究では，CNN を用いた一人称調理映像自動要約手法を提案したが，ネットワーク構成の十分な検討は行われていない．今後，ネットワークの深さ，活性化関数の種類，各畳み込み層のフィルター数，全結合層のノード数などの検討が必要である．

本研究の性能評価実験で用いたデータセットのアノテーションをつけた人数は一人である．しかし，最適な映像要約は要約を行う人間によって異なる．そのため，より公平な評

価実験のために同じデータに対して，複数人によるアノテーション付けが必要である．それに伴い，評価方法に関しても，F 値による定量的評価だけでなく, アンケートなどによる定性的な評価も行う必要がある.

付録 A

ソースプログラム

本研究で使用したプログラムを以下のディレクトリに置く．

```
xserver/user/shimada/graduation_thesis/programs/
```

以下のディレクトリの構造と概略を示す．プログラムのコンパイル方法，使用方法については README に記述する．

```
shimada/master_thesis/
|
|- threeStreamCNN/
|          #実験プログラム
|
```

付録 B

実験用データセット

本研究で使用したデータセットの所在を以下に示す．

```
xserver/user/shimada/graduation_thesis/EgoCooking-data/
```

以下のディレクトリの構造と概略を示す．プログラムのコンパイル方法，使用方法については README に記述する．

```
shimada/master_thesis/data/
|
|- test/
|                    #テスト実験映像
|
```

付録 C

発表資料

修士論文発表会で用いたプレゼンテーション資料を本論文の末尾に示す.

謝辞

本論文の執筆を終えるにあたり,関係諸氏の多大な御協力をいただきましたことに対し，深く感謝の意を表します.

研究を進めるにあたり，様々な提案と助言，ご指導を下さった若林哲史教授に深く感謝致します．毎回のディスカッションに参加し，様々なアドバイスを下さった三宅康二名誉教授，大山航准教授，白井伸宙助教，日頃からお世話になりました吉永みゆき事務官，中塚沙智子事務補佐官に深く感謝致します.

さらに，この研究室で 2 年間を共に過ごし，また多くのことを教えていただきました先輩・後輩の方々，同期の皆様に感謝いたします.

最後に，2 年間の大学院生活を長く支えてくれた両親，家族に今一度の感謝を表しまして，本論文の結びとさせて頂きます.

参考文献

[1] L.Rosenberg and J.Gentilucci. The use of first-person video to support teacher education. In EdMedia, 2017.

[2] 長徳 将希, 小泉 直也, 苗村 健. ミュージアムにおける一人称動画短縮のための場面抽出 －自身での振り返りと他社との共有. 情報処理学会論文誌, Vol.57, No12, pp.2516-2525, 2016.

[3] Y. Kameda, H. Fujiyoshi, T. Hagi, K. Yamaguchi and Y. Nakagawa. Challenge to ”First Person Vision”. IEICE, 110(28), 37-38, 2010.

[4] Gopro. http://gopro.com/.

[5] Google glass. https://www.google.com/glass/start/.

[6] Microsoft sensecam. http://research.microsoft.com/en-us/um/cambridge/projects/sensecam/.

[7] レシピ検索 No.1 ／料理レシピ載せるなら クックパッド http://cookpad.com/.

[8] Tasty https://www.buzzfeed.com/tasty.

[9] S. Mann. ’Wear Cam’ (the wearable camera): Personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis. In Proc. 2nd Int. Symp. Wearable Comput, 1998, pp. 124-131.

[10] Global Wearable Camera Market Research Report 2018, QYResearch Group.

[11] Police Body Worn Cameras: A Policy Scorecard. https://www.bwcscorecard.org/

[12] C. Tan, H.Goh, V. Chandrasekhar, L. Li, and J.-H. Lim. Understanding the nature of first person videos: Characterrization and classification using low-level features. In Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2014, pp. 549-556.

[13] P.Varini, G.Serra, and R. Cucchiara. Egocentric video summarization of cultural tour based on user preferences, In Proc. 23rd Annu. ACM Int. Conf. Multimedia Conf., 2015, pp. 539-542.

[14] P.Varini, G.Serra, and R. Cucchiara. Personalized egocentric video summarization for cultural experience. In Proc. 5th ACM Int. Conf. Multimedia Conf., 2015, pp. 2513-2520.

[15] Y.Poleg, C.Arora, and S.Peleg. Temporal segmentation of egocentric videos. In Proc. IEEE Conf. Comput. Vision Pattarn Recognit., 2014, pp. 2537-2544.

[16] A. G. Money and H. Agius,“Video summarisation: A conceptual framework and survey of the state of the art,”Journal of Visual Communication and Image Representation, vol. 19, no. 2, pp. 121143, 2008.

[17] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In CVPR, 2012.

[18] Y. J. Lee and K. Grauman, “Predicting important objects for egocentric video summarization,” International Journal of Computer Vision, pp. 118, 2015.

[19] M. Gygli, H. Grabner, and L. Van Gool,“Video summarization by learning submodular mixtures of objectives,” in IEEE Conference on Computer Vision and Pattern Recognition, pp. 30903098, 2015.

[20] Y.-L. Lin, V. Morariu, and W. Hsu, “Summarizing while recording: Context-based highlight detection for egocentric videos,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 5159, 2015.

[21] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh,“Gaze-enabled egocentric video summarization via constrained submodular maximization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 22352244, 2015.

[22] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In CVPR, 2011

[23] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In CVPR, 2010. 1

[24] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In CVPRW, 2009.

[25] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. InICCV, 2011

[26] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In ECCV, 2012.

[27] K. Matsuo, K. Yamada, S. Ueno, and S. Naito. An attention-based activity recognition for egocentric video. In CVPRW, 2014.

[28] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato. Coupling eye-motion and ego-motion features for first-person activity recognition. In CVPRW, 2012.

[29] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In CVPR, 2012.

[30] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In CVPR, 2013.

[31] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In CVPRW, 2009

[32] S. Sundaram and W. W. M. Cuevas. High level activity recognition using low resolution wearable vision. In CVPRW, 2009

[33] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In CVPR, 2011.

[34] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In CVPR, 2013.

[35] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In CVPR, 2012.

[36] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In CVPR, 2011.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

[38] Y. Poleg, T. Halperin, C. Arora, and S. Peleg. Egosampling: Fast-forward and stereo for egocentric videos. in IEEE Conference on Computer Vision and Pattern Recognition, pp. 47684776, 2015.

[39] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In ECCV, 2014.

[40] J. Kopf, M. Cohen, and R. Szeliski. First-person hyper-lapse videos. ACM Transactions on Graphics, 2014.

[41] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In ICCV, 2013

[42] Y. Hoshen and S. Peleg. Egocentric video biometrics. CoRR, abs/1411.7591, 2014.

[43] Y. Poleg, C. Arora, and S. Peleg. Head motion signatures from egocentric videos. In ACCV, 2014.

[44] A. Garcia, C. Tan, J. -H. Lim, and A.-H. Tan. Summarization of Egocentrtic Videos: A Comprehensive Survey. In Proc. IEEE Transactions on human-machine systems, Vol. 47., No. 1, Februray 2017, pp. 65-76.

[45] A. G. del Molino, B. Mandal, L. Li, and J. H. Lim, “Organizing and retrieving episodic memories from first person view,” in International Conference on Multimedia and Expo Workshops, pp. 16, IEEE, 2015.

[46] B. Xiong and K. Grauman, “Detecting snap points in egocentric video with a web photo prior,” Computer VisionECCV, pp. 282298, 2014.

[47] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, “Creating summaries

from user videos," in Computer VisionECCV, pp. 505520, Springer, 2014.

[48] B. Zhao and E. Xing, "Quasi real-time summarization for consumer videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 25132520, 2014.

[49] B. Xiong, G. Kim, and L. Sigal, "Storyline representation of egocentric videos with an applications to story-based search," in IEEE International Conference on Computer Vision, pp. 45254533, 2015.

[50] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization,"in Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[51] K. Aizawa, K. Ishijima, and M. Shiina,"Summarizing wearable video,"in International Conference on Image Processing, vol. 3, pp. 398401, IEEE, 2001.

[52] H. W. Ng, Y. Sawahata, and K. Aizawa, "Summarization of wearable videos using support vector machine," in International Conference on Multimedia and Expo, vol. 1, pp. 325328, IEEE, 2002.

[53] V. Bettadapura, D. Castro, and I. Essa, "Discovering picturesque high-lights from egocentric vacation videos," arXiv preprint arXiv:1601.04406, 2016.

[54] E. Talavera, M. Dimiccoli, M. Bolaos, M. Aghaei, and P. Radeva, "R-clustering for egocentric video segmentation," Pattern Recognition and Image Analysis, pp. 327336, 2015.

[55] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," arXiv preprint arXiv:1605.08110, 2016.

[56] M. Okamoto and K. Yanai,"Summarization of egocentric moving videos for generating walking route guidance," Image and Video Technology, pp. 431442, 2014.

[57] S. Yeung, A. Fathi, and L. Fei-Fei,"Videoset: Video summary evaluation through text," arXiv preprint arXiv:1406.5824, 2014.

[58] Y. Le Cun, L. Bottou, and Y. Bengio. Reading checks with multilayer graph transformer networks. In Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3431-3440.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing ststems, pp. 675-678. ACM.

[60] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutinal Networks for Semantic Segmentation," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440, 2015.

[61] G. Farneback. Two-frame Motion Estimation based on Polynomial Expansion. 13th Scandinavian Conference, SCIA 363-370, 2003 Halmsted, Sweden, June 2, 2003.

[62] Pivothead http://pivothead.com/.

[63] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems, pages 568576, 2014.

食材の位置と撮影者の動きに着目したCNNの融合による一人称調理映像の自動要約

三重大学大学院 工学研究科 情報工学専攻
ヒューマンインターフェース研究室
島田 尚宜

研究背景

一人称　　映像

教育[1]

レシピ動画[2]

一人称カメラを用いた調理の撮影

[1] How to Tie Basic Tie, First person view, fpv How To Tie Four In Hand Knot, hacer corbata https://www.youtube.com/watch?v=cwqzy2tiiOE
[2] ～レンジで簡単～もやしと豆苗の豚巻きレンジ蒸し https://www.youtube.com/watch?v=GJB7CSS9F7E

研究目的と一人称映像の課題

目的：一人称調理映像から調理内容とその方法理解に必要な映像の抽出 ➡ **要約**

- 入力: 一人称調理映像
- 出力: 調理動作フレームの集合

課題：一人称調理映像は多くの無駄が存在

関連研究：　Optical-flowを用いた手法[3][4]

- Optical-flow：連続するフレーム間での動きをベクトル表現したもの

➡ **美術館における一人称映像の要約を実現**

[3] Buschek, D., Spitzer, M. and Alt, F. Video-Recording Your Life: User Perception and Experiences. CHI EA '15, pp.2223-2228 (2015).
[4]長徳 将希，小泉 直也，苗村 健. ミュージアムにおける一人称動画短縮のための場面抽出－自身での振返りと他者との共有. 情報処理学会論文誌, Vol.57, No.12, pp.2516-2525 (2016).

研究目的と提案する要約システム１

目的：一人称調理映像から調理内容とその方法理解に必要な映像の抽出 ➡ **要約**

- 入力: 一人称調理映像
- 出力: 調理動作フレームの集合

課題：一人称調理映像は多くの無駄が存在

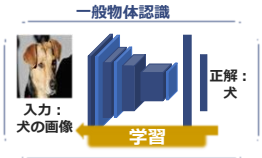
提案手法１：Optical-flowとCNN

Convolutional Neural Network (CNN)

CNNとは

概説：一般物体認識と呼ばれる画像認識のタスクで優れた性能を持つアルゴリズム

- 認識に有効な特徴
  - 学習に基づく特徴抽出
- 大量のデータを学習



提案手法１：Optical-flowとCNN

Convolutional Neural Network (CNN)

実験概要

データセット：複数メニューを同時に調理した映像をウェアラブルカメラで撮影

学習データ

テストデータ

データ	全体のフレーム数	要約映像に含むフレーム数
学習用	10,300	5,150
テスト用	14,890	8,100

実験概要

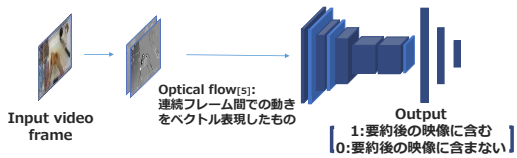
評価方法：フレーム単位での*Precision*，*Recall*を求め，*F*値を算出

- ***Precision***：要約映像から得られる情報の質
- ***Recall***：要約映像の見やすさ

$$Precision = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN},\ F = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$$

- ✓TP：要約映像に含むフレームを提案手法で正しく分類したフレーム数
- ✓FP：要約映像に含まないフレームを提案手法で誤分類したフレーム数
- ✓FN：要約映像に含むフレームを提案手法で正しく分類できなかったフレーム数

提案手法１



	適合率*P*	再現率*R*	F値
動きの特徴（提案手法１）	54.08%	17.96%	25.85%

考察：調理動作と類似する皿洗いのような動作を含むため，オプティカルフローのみに着目するCNNを用いての要約は困難

[5] G. Farneback, "Two-frame motion estimation based on polynominal expansion", Proc. Of 13th Scandinavian Conf. on Image Analysis, SCIA 2003, 363-370, 2003

研究目的と提案する要約システム２

目的：一人称調理映像から調理内容とその方法理解に必要な映像の抽出 ➡ **要約**

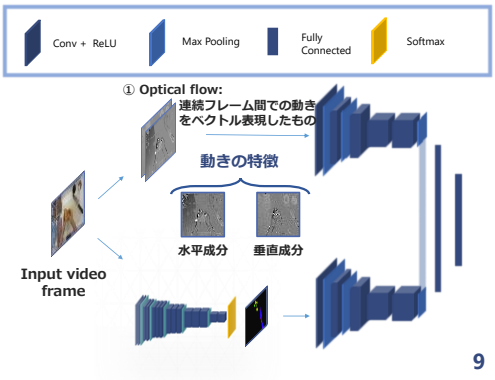
- 入力: 一人称調理映像
- 出力: 調理動作フレームの集合

事前知識：調理動作映像の場合，以下の条件を満たすフレームを要約映像に含むべき

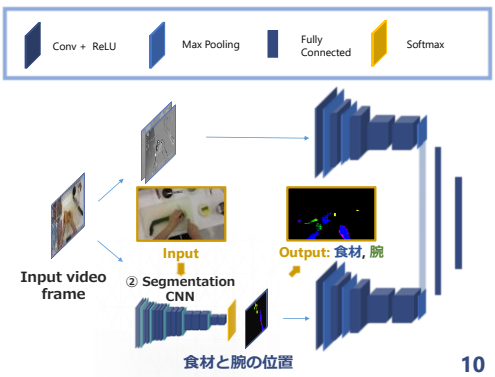
- **食材が写っていること**
- **手に一定の動きがあること**

提案手法２：事前知識を活用した手法を提案

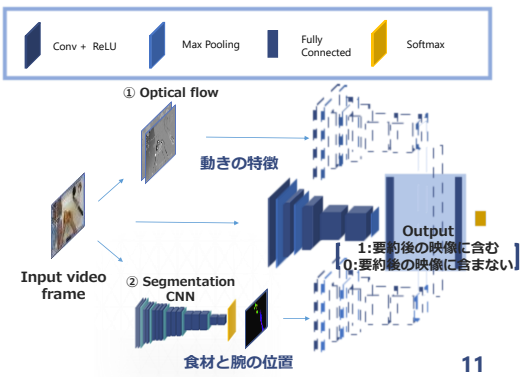
提案手法２



提案手法２



提案手法２



実験結果の比較

評価方法と結果：適合率, 再現率を求め,
*F*値を算出

	適合率*P*	再現率*R*	F値
元画像	34.87%	3.96%	7.11%
動きの特徴 （提案手法１）	54.08%	17.96%	25.85%
食材と腕の位置	54.66%	44.29%	48.93%
提案手法２	**57.45%**	**76.47%**	**65.61%**

12

要約結果の比較

オプティカルフローを用いる手法１

VS.

食材と腕の位置を用いる手法２

考察：一人称映像では撮影者が調理中の場合，
腕は画面下部，そして調理対象食材は画面中央に位置することが多いため，
食材と腕の位置を学習することで，調理動作とそれ以外の動作を区別可能

より自然な要約映像作成を実現

13

まとめ

目的：一人称調理映像から調理の全体的な流れを
視覚的・直感的に理解可能な映像自動要約

提案手法：

Input
一人称調理映像
の1フレーム

CNNs

Output
1:要約後の映像に含む
0:要約後の映像に含まない

評価結果：**オプティカルフローを用いる手法１の
フレーム選択F値25.85%に対し，
食材と腕の位置を用いる手法２では
65.61%に向上した．
また手法２によって，より自然な
一人称調理映像の要約が可能になった．**

14