特別研究報告書

物体の把持・解放を手がかりとした調理映像からの動作区間検出

指導教員　　美濃 導彦 教授

京都大学工学部情報学科

松村 優樹

平成27年1月30日

物体の把持・解放を手がかりとした
調理映像からの動作区間検出

松村 優樹

内容梗概

近年，インターネットの普及に伴い，レシピ投稿サイトの利用が広まっており，多種多様なレシピがWebを通して閲覧可能である．レシピ中の文章で説明される作業には,「牛蒡をささがきする」,「胡瓜を塩ずりする」等，どのように食材を加工するのか実際に見てみないと理解が難しい作業が存在する．そのような作業の様子を映像クリップで確認できれば，調理者はそれを模倣するだけでよく，直感的でわかりやすい．対応する文章に映像クリップが付与されたレシピは有用である．

このようなレシピの作成において，映像クリップを用意することが必要となる．料理を作る際，食材をどのように加工するのかを理解することが重要である．調理中の「切る」,「炒める」等の加工動作に注目して，映像クリップを用意する．加工動作毎に個別に撮影することが考えられるが，これでは調理の中断が頻発し煩わしい．そこで，固定カメラで調理の開始から終了までを撮影した映像を用いる．以降これを調理映像と呼ぶ．映像クリップとして調理映像から加工動作が行われる映像区間を検出する．これを本研究の目的とする．

従来，このような調理映像からの動作区間検出は，加工動作毎に学習した識別器を用いた動作認識によって行われてきた．しかし，このような方法では調理映像で観測される加工動作を事前に調べる必要がある．これでは適用できる場面が限られてしまう．

本研究では，調理映像を複数の映像区間に分割し，得られた映像区間をクラスタリングすることを提案する．加工動作が行われる時，その加工に応じた特有な動きが観測されると考えられる．各映像区間から抽出される動作特徴の類似性に基づいたクラスタリングによって，そのような動作特有の動きが観測された映像区間を一つのクラスタにまとめあげる．レシピ作成者が各クラスタに加工動作のラベルを割り当てることで，調理映像からの動作区間検出が実現される．この方法は事前学習を必要とせず，どのような調理映像にも適用できる．

調理映像を映像区間に分割することが最初の処理となる．得られる映像区間によって抽出できる動作特徴も異なるので，映像分割の方法が重要となる．単純

な方法としては区間長を固定することが考えられる．固定幅が長すぎると，ある 1 種類の加工動作を含む映像区間にそれ以外の加工動作や加工動作以外が含まれてしまう．逆に，固定幅が短すぎると，ある 1 種類の加工動作が多数の映像区間に分割され，区間内で加工動作特有の動きが観測できなくなってしまう．加工動作はその種類によって作業時間が様々である．調理中の加工動作に対して常に最適であるような固定幅を設定することは難しい．

そこで，区間長を可変とするような映像分割を考える．このような映像分割の方法として物体の把持・解放を利用することを提案する．ここで，物体の把持とは調理者が食材や調理器具を取ることを指し，物体の解放とは逆に把持しているものを置くことを指す．加工動作の前後では物体の把持，あるいは，解放が起こる．逆に，加工動作中には物体の把持，あるいは，解放が起きにくい．そのため，物体の把持・解放を区切り目とした場合，得られる映像区間のうち加工動作が含まれるものについては，その加工動作しか含まれず，かつ，加工動作特有の動きが観測できる区間長を保つことが期待できる．

提案手法の有効性を示す為，KUSK DATASET で公開されているデータのうち，調理者もレシピも異なる 3 つの調理映像を対象として，固定幅を利用した手法と物体の把持・解放を利用した提案手法を比較した．

3 つの調理映像からそれぞれの手法で取得されたクラスタの集合に対して，同種の加工動作を 1 つのクラスタでまとめあげたか，異なる加工動作を区別して別々のクラスタにまとめあげたか，によって検証した．人手で与えたフレーム毎の加工動作の正解ラベルから各クラスタに含まれる加工動作毎の再現率を求め，クラスタ間での再現率を比較した．その結果，提案手法では多くの加工動作で，固定幅を利用した手法よりも少ないクラスタでまとめあげた．また，提案手法は固定幅を利用した手法では区別できなかった加工動作を区別して別々のクラスタにまとめあげた．

今後の課題としては，提案手法のクラスタリング結果に基づいて，レシピ中の文章への映像クリップの付与を自動で行うことが挙げられる．

Cutting Out Clips from a Cooking Video referring to Object Access

Yuki MATSUMURA

Abstract

In recent years, user-generated recipe sites have been widespread with the growth of the Internet, and a wide variety of recipes are available on those sites. Some instructions on recipe texts, such as "puree tomato" or "beat egg white to soft-peak," are easier to understand by watching how to process foodstuffs in video than by reading texts. If a chef watches video clips corresponding to such instructions, he/she can understand the way of the processing intuitively and imitate it. Recipes with the movie clips are useful to many chefs.

To make such recipes, it needs to prepare the clips. When a chef makes a dish, it is important to understand how to process foodstuffs. To prepare the clips, we focus on food processing, such as cutting and stir-frying, because it is interruptive and bothersome for many users to record the clips one-by-one for each type of food processing. Thus, we propose recording a video throughout a chef's cooking activity with mounted cameras. We detect scenes of food processing from the video as the movie clips. Our goal is to obtain the clips corresponding to each type of food processing.

To detect such scenes, a generally accepted practice is to use motion recognition. However, the motion recognition preliminarily needs food processing that is observable. Therefore, videos to which such an approach can be applied are limited.

In this paper, we propose a method that divides the video into segments and clusters the obtained segments. From a preliminary observation, we found that each type of food processing involves own peculiar motion. Hence, extracting motion feature from the video and clustering the extracted feature agglomerate a type of food processing with the peculiar motion into one cluster. When a user labels the cluster as the type of food processing, scenes of food processing are detected from the video. This method doesn't need to restrict types of food processing and is available for any video.

The key point is how to divide the video into segments at the first step

of the method because the motion feature depends on the obtained segment. A simple approach is to divide the video into segments with the same length. When the length is too long, the segment including a type of food processing includes other motions. Conversely, when the length is too short, a type of food processing tends to be separated into a lot of segments.

Because the time for food processing is different by type, the length should be various by type. We focused on the fact that a chef picks up foodstuffs or cooking tools at the start of food processing, and puts them down at the end. Also, the chef hardly picks them up or puts them down during food processing. We divide the video by the moments when a chef picks up or puts down objects. Then, it is expected that each segment includes either a type of food processing or non-processing motion.

In order to evaluate the performance of the proposed method, we compared it to the methods using the same lengths. From KUSK Dataset, which is a public dataset of cooking activity, we used three cooking videos in which the chefs and the recipes are different.

We examined clusters obtained from the videos by each method in terms of whether or not a type of food processing was packed into a cluster, and whether or not a cluster includes a type of food processing. We calculated recall rates of all types of food processing in each cluster based on correct labels of food processing and compared the recall rates among the obtained clusters. As a result, a type of food processing was packed into fewer clusters, and clusters including plural types of food processing was fewer in the proposed method than in the methods with the same lengths.

In future work, we plan to automate the matching between the obtained clips and the instructional texts based on the result of clustering.

物体の把持・解放を手がかりとした
調理映像からの動作区間検出

目次

第1章　はじめに

調理とは，料理を作る為に行われる活動である．一般的に，調理の手順は作りたい料理によって異なる．調理者によってはその調理手順がわからない料理が存在することは十分考えられる．例えば，テレビ等のメディアで知った料理を家庭で再現しようとしたとき，食べたことはあるが調理したことはない料理に挑戦するとき等が挙げられる．このような状況において，レシピが利用できる．

近年，インターネットの普及に伴い，個人ユーザが作成したレシピが公開される，レシピ投稿サイト [1][2] の利用が広まっている．レシピ投稿サイトで公開されているレシピは，一般的に調理途中，あるいは完成した料理の静止画と文章で記述されている．このようなレシピを参照することで，調理者は作りたい料理に必要な調理手順を知ることができる．

静止画や文章で記述されたレシピでは，説明されている調理手順から具体的にどのような作業をすればよいのかを把握する必要がある．これに対して，そのような作業の様子を記録した映像クリップがあれば，調理者はそれを模倣するだけでよく，直感的で分かりやすい．

AmKitchen[3] では，このような映像クリップが付与されたレシピを参照することができる．しかし，AmKitchen で公開されているレシピではプロの映像編集者によって映像クリップが付与されている．これに対して，レシピ投稿サイトには様々な個人ユーザによって多種多様なレシピが投稿され，実際，レシピ投稿サイトである COOKPAD には 196 万ものレシピが存在する．これら大量のレシピに映像クリップを付与する場合，AmKitchen のようにプロの映像編集者が行うことは現実的ではない．そこで，個人ユーザがレシピへの映像クリップの付与を行うことを考える．

このようなレシピの作成において，映像クリップを用意することが必要となる．料理を作る際，食材をどのように加工するのかを理解することが重要である．調理中の「切る」,「炒める」等の加工動作に注目して，映像クリップを用意する．加工動作毎に個別に撮影することが考えられるが，これでは調理の中断が頻発し煩わしい．そこで，固定カメラで調理の開始から終了までを撮影した映像を用いる．以降これを調理映像と呼ぶ．映像クリップとして調理映像からの加工動作が行われる映像区間を検出する．これを本研究の目的とする．

従来，このような調理映像からの動作区間検出は，加工動作毎に学習した識

別器を用いた動作認識 [4][5] によって行われてきた．しかし，このような方法では調理映像で観測される加工動作を事前に調べる必要がある．これでは適用できる場面が限られてしまう．

本研究では，調理映像を複数の映像区間に分割し，得られた映像区間をクラスタリングすることを提案する．加工動作が行われる時，その加工に応じた特有な動きが観測されると考えられる．各映像区間から抽出される動作特徴の類似性に基づいたクラスタリングによって，加工動作特有の動きが観測された映像区間を 1 つのクラスタにまとめあげる．レシピ作成者が各クラスタに加工動作のラベルを割り当てることで，調理映像からの動作区間検出が実現される．この方法は事前学習を必要とせず，どのような調理映像にも適用できる．

調理映像をどのような映像区間に分割するかが最初の処理となる．得られる映像区間によって抽出できる動作特徴も異なるので, 映像分割の方法が重要となる．単純な映像分割の方法として区間長を固定することが考えられる．固定幅が長すぎると，ある 1 種類の加工動作を含む映像区間にそれ以外の加工動作や加工動作以外が含まれてしまう．逆に，固定幅が短すぎると，1 種類の加工動作が多数の映像区間に分割され，区間内で加工動作特有の動きが観測できなくなってしまう．加工動作はその種類によって作業時間が様々である．調理中の加工動作に対して常に最適であるような固定幅を設定することは難しい.

そこで，区間長を可変とするような映像分割を考える．このような映像分割の方法として物体の把持・解放を利用した手法を提案する．ここで，物体の把持とは調理者が食材や調理器具を取ることを指し，物体の解放とは逆に把持しているものを置くことを指す．加工動作の前後では物体の把持，あるいは，解放が起こる．逆に，加工動作中には物体の把持，あるいは，解放が起きにくい．そのため，物体の把持・解放を区切り目とした場合，得られる映像区間のうち加工動作が含まれるものについては，その加工動作しか含まれず，かつ，加工動作特有の動きが観測できる区間長を保つことが期待できる．

本稿の構成は以下の通りである．まず，2 章で関連研究に対する本研究の位置づけを説明する．次に，3 章で調理映像からの動作区間検出の方法を説明し，その技術的問題点と解決策を述べる．さらに，4 章で提案手法の実装方法を説明する．5 章で実験の結果と考察を述べ，6 章でまとめと今後の課題について述べる．

第2章　関連研究

加工動作ごとに学習した識別器を用いた動作認識によって，調理映像からの動作区間検出を行う方法が従来考えられてきた [4][5]．このような動作認識を用いて，未知の調理映像から動作区間検出を行う場合，その調理映像で観測される加工動作を事前に調べる必要がある．これでは適用できる場面が限られてしまう．

これに対して，久原ら [6] は加工動作の多くが「切る」,「混ぜる」などの繰り返し動作であると考え，調理映像からの動作区間検出として繰り返し動作区間を検出している．繰り返し動作に注目した動作区間検出を行う研究は他にも存在し [7][8]，調理映像の要約や料理レシピと調理映像との対応付けで成果をあげている．しかし，これらの研究では繰り返しの周期がパラメータで制限されており，その周期に合った加工動作だけが検出対象となってしまう．

一方，山肩ら [9] はレシピに登場する加工動作を統計的に調べ，5つのカテゴリ（「加える」,「焼く・炒める」,「煮る・茹でる」,「切る・剥く」,「揚げる」）のいずれかに分類できるものだけを検出対象としている．これにより，調理者の位置や赤外線カメラ用いることで，様々な調理映像を対象とした動作区間検出が可能となる．このような動作区間検出の結果と，レシピと調理映像を木構造で表現することによってその対応付けを実現している．しかし，5つのカテゴリに含まれない加工動作は考えていない．

本研究では，個人ユーザが作成するレシピに映像を付与させることを想定するので，個人ユーザがそのレシピに沿って行う調理の様子を記録した映像を対象として動作区間検出を行う．このとき，個人ユーザによって調理の仕方は様々であり，調理映像で観測される加工動作は不確定なため，検出対象となる加工動作をあらかじめ定めてしまうのは好ましくない．そこで，本研究では，調理映像を複数の映像区間に分割し，レシピ作成者が各クラスタに加工動作のラベルを割り当てることを想定して得られた映像区間をクラスタリングする．この方法は事前に加工動作を限定せず，どのような調理映像にも適用できる．

第3章　調理映像からの動作区間検出

本章では，調理映像からの動作区間検出において，問題となる映像の分割法として，物体の把持・解放を利用することの有用性を示す．まず．3.1 節で問題

を定式化する．次に，3.2 節で映像分割における問題点を述べる．最後に，3.3 節で本研究で提案する物体の把持・解放による映像分割を説明する．

3.1 動作区間検出の定式化

本研究では，調理映像を複数の映像区間に分割し，得られた映像区間から抽出される動作特徴の類似性に基づいたクラスタリングによって，加工動作特有の動きが観測された映像区間を一つのクラスタにまとめあげることを考えた．各クラスタへの加工動作のラベル付けを行うことで，調理映像からの動作区間検出は実現される．本節ではこれを定式化する．

調理映像をフレームの列として，k 番目のフレームを f_k とする．また，総フレーム数 K の調理映像を $F = \{f_1, ..., f_K\}$ と表す．次に，調理映像 F に対する区切り目の集合を $\{s_0, ..., s_M\}$ とする．ただし，$s_0 = 0,\ s_M = K$ である．このとき，調理映像は M 個の映像区間に分割される．m 番目の映像区間をその区間の開始となる区切り目と終了となる区切り目によって $S_m = \{s_{m-1}, s_m\}$ と表す．ここで，ある映像分割法 g によって得られる調理映像 F に対する映像区間の集合を $g(F) = \{S_1, ..., S_M\}$ とする．

映像区間 S_m から特徴量 x_m を抽出する関数を $h : S_m \rightarrow x_m$ とする．映像区間ごとに抽出された特徴量の集合 $\{x_1, ..., x_M\}$ をその類似性に基づいてクラスタリングする．このとき，各クラスタ内の特徴量 x_m と映像区間 S_m は一対一で対応する．従って，得られたクラスタ数を N 個とし，各クラスタに属する x_m に対応した S_m の集合で区間クラスタ $C_n(0 < n \leq N) \subset g(F)$ を定める．任意の n で区間クラスタ C_n に属する映像区間がある 1 種類の加工動作か加工動作以外のいずれか一方に対応すれば，本研究の目的は達成される．

3.2 調理映像の分割における問題点

前節で定式化した手法では，調理映像を映像区間に分割することが最初の処理となる．得られる映像区間によって抽出できる動作特徴は異なるので，映像分割法 g が重要となる．加工動作が行われる時，その加工に応じた特有の動きが観測されると考えられる．映像区間で加工動作特有の動きが観測されたかどうかを判断するためには，ある程度の区間長が必要であると考えられる．また，映像区間から動作特徴を抽出するので，映像区間はある 1 種類の加工動作，あるいは，加工動作以外で構成されるべきである．

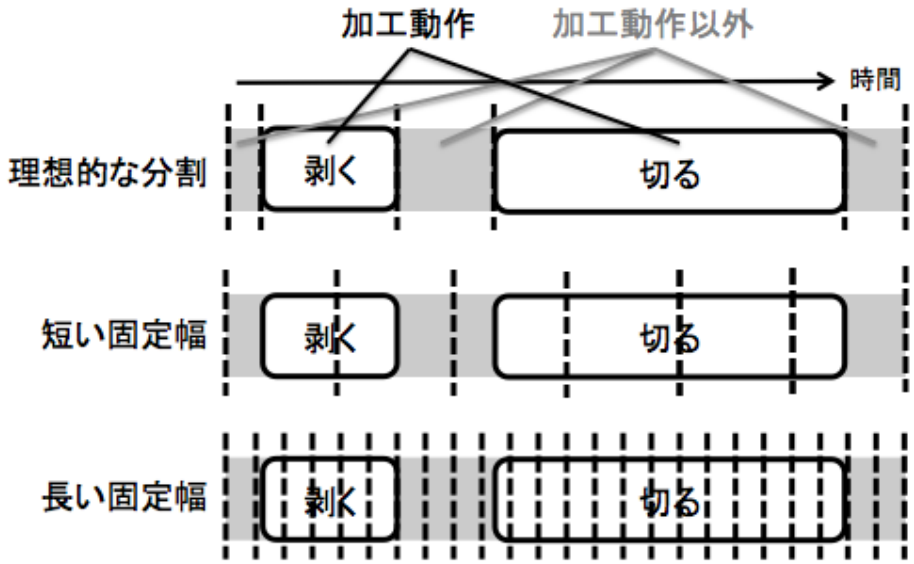

図 1: 調理映像の固定幅による分割

このとき，映像分割法 g としては，映像区間の長さを固定する方法が考えられる．ここで，映像区間 S_m の長さを $l_m = s_m - s_{m-1}$ で表す．映像区間の長さが固定幅のとき，その長さを l とする．このとき，m によらず $l_m = l$ となる．ここで，l が短すぎると，1 種類の加工動作が多数の映像区間に分割され，区間内で加工動作特有の動きが観測できなくなってしまう．逆に，l が長すぎると，ある 1 種類の加工動作を含む映像区間にそれ以外の加工動作や加工動作以外が含まれてしまう．

図 1 では，加工動作として「剥く」,「切る」が行われた調理映像の固定幅による分割の様子を示している．映像区間の区切れ目は点線で表現されている．固定幅が短いと,「切る」のように作業時間の長い加工動作は多数の映像区間に分割され,「切る」特有の動きが観測できない可能性が生じる．逆に，固定幅が長いと,「剥く」を含む映像区間に加工動作以外が含まれてしまう．

このように調理で観測される加工動作はその種類によって作業時間が様々である．そのため，固定幅による映像分割では最適な幅を定めることが困難である．本稿では物体の把持・解放を映像の区切り目として提案し，調理映像の可変幅な分割を行う．

3.3　映像分割のための物体の把持・解放

調理において加工動作の前後では，加工の対象となる食材や加工のための調理器具が持ち替えられる．この持ち替えのタイミングを利用することで加工動作の長さに応じて可変幅な映像区間を得ることが期待される．そこで，本研究で提案する動作区間検出のための映像分割法 g では，持ち替えのタイミングとして物体の把持と解放を利用する．

物体の把持とは調理者が食材や調理器具を取ることを指し，物体の解放とは逆に把持しているものを置くことを指す．ここで，物体を把持しているとはその物体を調理者が直接手に持っている状態だけを示すものではない．例えば，調理者が菜箸で食材を持ち上げたり，食材の入ったボウルを手で持った場合，調理器具としての菜箸やボウルを把持しているだけでなく，同時に食材も把持している．よって，物体を把持しているとは直接・間接を問わず，調理者がその物体に関与していることを表し，解放はその関与がなくなることを表す．このような物体の把持と解放によって持ち替えのタイミングを表現し，調理映像の区切り目とする．

加工動作の前後では物体の把持，あるいは，解放が起こると考えられる．そのため，加工動作とそれ以外の加工動作や加工動作以外の境界付近では，物体の把持，あるいは，解放による区切り目が存在すると想定される．よって，それぞれ映像区間はある 1 種類の加工動作，あるいは，加工動作以外で構成されると考えられる．

また，加工動作を行っている最中，物体の把持，あるいは，解放は起こりにくいと考えられる．そこで，物体の把持，あるいは，解放を区切り目とすれば，作業時間の長い加工動作が多数の単位区間には分割されないと想定される．これにより，加工動作特有の動きが観測できなくなってしまうことを防ぐことが期待できる．

提案手法では，先行研究である橋本らの手法 [10] を利用して，区切り目となる物体の把持と解放を調理映像から検出する．図 2 は，調理映像で観測される物体の把持と解放，さらにその検出結果の例を示す．このとき，図 2 の検出結果のように，検出逃しや検出誤り，さらに把持と解放を間違えて検出してしまう検出結果誤りが考えられる．

これに対して，本稿では，物体の把持と解放を区別せずに映像の区切り目と

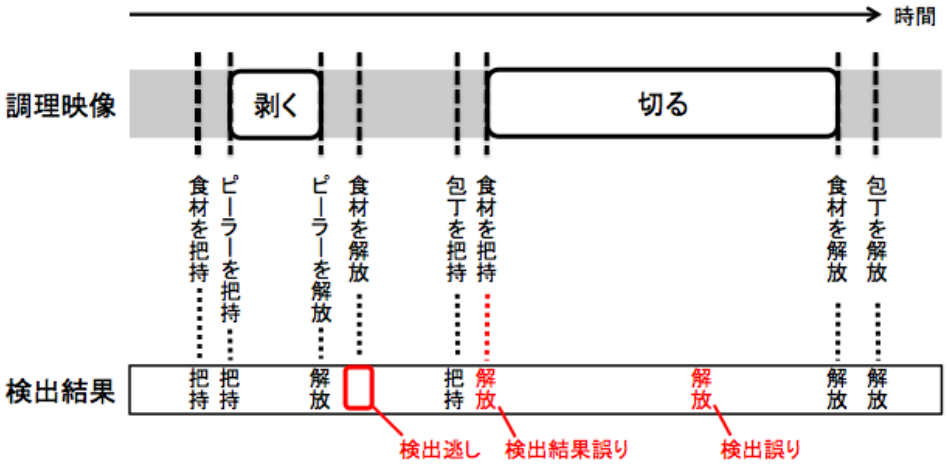

図 2: 検出される物体の把持・解放

することで対処するので，検出結果誤りは考えなくてよい．図 2 のように，加工動作以外の映像区間では，物体の把持，あるいは，解放が加工動作の映像区間よりも頻繁に観測されるため，加工動作の前後での区切り目の候補となる物体の把持，あるいは，解放は複数存在する．これにより，検出逃しの影響が抑えられる．また，検出誤りが集中して発生しない限り，加工動作が多数の映像区間に分割されてしまうことはない．

以上から，調理映像の可変幅な分割において，物体の把持，あるいは，解放を区切り目とすることは有用であると考えられる．調理映像で観測される物体の把持と解放の総数を M とし，物体の把持，あるいは，解放が観測されたタイミングを区切り目の集合 $\{s_1, ..., s_M\}$ として得る．このように物体の把持と解放を区別しないことを以降，物体の把持・解放で表す．本稿では，物体の把持・解放を区切り目とした映像分割を利用して動作区間検出を行う．

第 4 章　物体の把持・解放による動作区間検出

本章では，調理映像からの動作区間検出を行うための具体的な処理について説明する．まず 4.1 節では，本研究で用いる調理映像の撮影環境を説明する．次に 4.2 節で，橋本ら [10] の手法を用いた調理映像からの物体の把持・解放の検出手法を述べる．さらに 4.3 節では，映像区間からの動作特徴の抽出手法を述べる．最後に 4.4 節で，動作特徴の類似性に基づいたクラスタリングにより区間クラスタを取得する手法を述べる．

図 3: 調理映像のシーン（左から「シンク」,「調理台」,「コンロ」）

4.1 調理映像の撮影環境

本研究では．レシピを作成する個人ユーザが行った調理の様子を記録した映像を対象とする．このような調理は主に「シンク」,「調理台」,「コンロ」の 3 箇所で行われると考えられる．そこで，これら 3 箇所の設備で行われる調理を撮影する．このとき，調理状況に応じて適切な角度で撮影することが考えられる．このようなカメラワークを個人ユーザが行う場合，調理者は作業に集中するだけでなく，カメラの角度についても常に意識しなくてはならない．これに対して，固定カメラで撮影すれば個人ユーザは作業だけに集中することができる．そこで，本研究では，固定カメラによって撮影された調理映像を用いる．

個人ユーザによってカメラの設置場所は異なると考えられる．しかし，調理状況を把握する為には，調理全体を通して，人の行動，及び，物体の把持・解放が正しく観測できるような位置・角度に設置されたカメラによる調理映像が有効である．調理の様子を横から撮影すると，食材や調理器具等が置かれている位置や調理者の立ち位置によって，観測したい物体や動きが遮蔽物の影響を受けることが考えられる．また，カメラとの距離によって動きや物体の大きさが変化してしまい，類似性判定が困難になる．そこで，本研究では調理の様子を真上から撮影した調理映像を用いる．調理映像のシーンを図 3 に示す．なお，これら 3 つとは異なる角度からの調理映像が必要であれば別に用意し，時間の同期を行えば問題ない．

4.2 物体の把持・解放の検出

提案手法の実現のためには，まず調理映像を映像区間に分割する必要がある．このとき，調理映像 F の区切り目の集合 $\{s_1, ..., s_M\}$ を物体の把持・解放が行われたフレームの検出結果によって得る．物体の把持・解放の検出は，調理映像中の食材や調理器具が占める領域を把握し，そのような領域が把持されてい

るかどうかを判定することで実現できる．

食材や調理器具の占める領域の把握は，映像中の物体を検出することで実現される．ここで，物体認識に基づく手法が考えられるが，物体認識によって物体の占める領域を見つけるためには，その物体の形状や色に関して事前に用意して学習しておく必要がある．これでは，事前知識なしに調理映像から動作区間検出を行う提案手法に合わない．

事前知識を必要としない物体検出として，類似した特徴を持つ画素の集まりであるスーパーピクセルを利用した手法が考えられる．Li ら [11] は，映像に対して，スーパーピクセルを用いて検出された物体領域と肌色抽出によって検出された作業者の領域との位置関係から作業者が手に持っている物体を検出する手法を提案している．このとき，物体の把持は手で直接持っていることを前提としている．これに対して，本研究で対象となる調理映像中の作業では，菜箸等によって間接的に物体を持つ場合も物体の把持に含まれる．従って，この手法は本研究での物体の把持を検出する手法としては適さない．

間接的な物体の把持も含めて物体の把持・解放を検出する手法を提案した研究として橋本ら [10] の研究がある．この手法での物体検出は，背景との非類似性を評価する背景差分によって行われている．このとき，事前知識として背景画像が必要となる．しかし，本研究での映像は固定カメラによる撮影を前提とするので，背景画像の取得は容易である．よって，提案手法の実装において問題とならない．ここで，背景画像と一致しない画素の集合からなる領域を前景領域と呼ぶ．

この手法では，作業台の外から連結した前景領域を人物領域として定義している．この定義では，例えば，作業者が食材の入ったボウルを手に持った場合，ボウルだけでなく食材も人物領域に含まれる．物体の把持は，作業台に置かれていた物体領域が人物領域と重なり，かつ，移動することで検出される．人物領域は把持している物体領域も含むため，間接的な物体の把持も検出される．従って，この手法によって物体の把持が検出されたフレーム番号を本研究での物体の把持の検出結果とする．

物体の解放は，人物領域に含まれていた物体領域が人物領域と非連結な前景領域となることで検出される．このとき，作業者が本当に解放したかどうかを判定する為に，このような前景領域に対して一定時間の静止判定が行われる．そのため，この手法で検出される物体の解放のタイミングは実際に解放が行われ

図 4: 物体の把持・解放が検出されたフレーム

たタイミングよりも遅くなる．そこで，物体の解放が検出されたフレーム番号から静止判定で使用されるフレーム数を差し引いたものを本研究での物体の解放の検出結果とする．

図 4 は，この手法で検出された調理台での物体の把持・解放のフレームの様子を示す．物体が把持された瞬間，解放された瞬間ではないが，その付近で物体の把持・解放が検出されていることがわかる．このようにして得られた物体の把持・解放に対するフレーム番号の集合を，その番号が小さいものから順に並べることで，区切り目の集合 $\{s_1, ..., s_M\}$ を得る．この区切り目によって調理映像を映像区間 S_m に分割する．

4.3 映像区間からの動作特徴抽出

映像区間 S_m から抽出される動作特徴 x_m を得る．様々な動作特徴が提案されているが，本研究では任意の長さの映像区間から抽出可能な特徴を用いる必要がある．そのため，各フレーム f_k から抽出される特徴量 v_k を映像区間内全体で統合する手法を用いる．本稿では，フレーム単位の特徴量 v_k として，立体高次局所自己相関 (CHLAC;Cubic Higher-order Local Correlation) 特徴を利用する．

CHLAC 特徴は，注目画素の近傍領域における画素値の変化の有無のパターンを画像中のすべての画素から数え上げることで得られる．このとき，数え上げは注目画素の位置の区別なく行われる．そのため，数え上げの範囲を指定することで，画面上で観測される動きの位置依存性を自由に扱うことができる．

4.1 節で述べたように，調理は主に「シンク」,「調理台」,「コンロ」の 3 箇所の設備で行われる．これら 3 箇所で行われる加工動作はその設備によって，そ

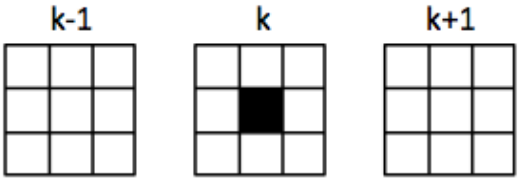

図 5: 0 次のパターン

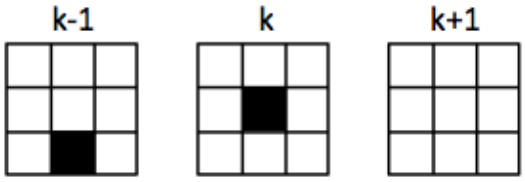

図 6: 1 次のパターンの例

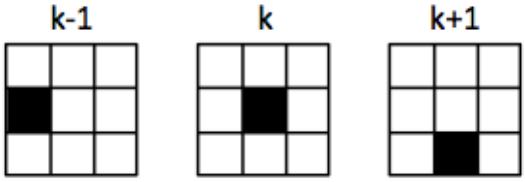

図 7: 2 次のパターンの例

れぞれ目的が異なる．そのため，加工動作特有の動きをまとめあげるためには同じ設備で観測される動きだけを比較すればよい．CHLAC 特徴の数え上げの範囲を設備毎にすれば，その範囲内での動きの大きさや方向に対する特徴を抽出することができる．

CHLAC 特徴を抽出するためには，まず前処理として，対象となる映像のフレーム間差分画像を二値化する必要がある．この前処理によって，動きに関する成分以外の画像情報が除去される．

注目画素の近傍の領域は，画素値の位置を 2 次元，時間方向を 1 次元とする 3 次元の時空間中の $3 \times 3 \times 3$ の局所領域で表される．このような局所領域に関するパターンの例を図 5，6，7 に示す．パターンの次数は注目画素以外に参照する画素の数に対応している．それぞれのパターンに対して，そのパターンが参照する画素すべてで動きが観測される局所領域の数を数え上げることで特徴量を抽出する．このとき，変位パターンの組み合わせは，0 次が 1 通り，1 次が 13 通り，2 次が 237 通りである．よって，CHLAC 特徴は 251 次元のベクトルで構成される．

3 箇所の設備ごとに抽出される CHLAC 特徴はそれぞれ 251 次元のベクトルとなる．調理中には設備の境界付近で行われる作業が存在する．このような作業に対応する為，各設備で抽出された CHLAC 特徴をつなげ，753 次元の結合された CHLAC 特徴を得る．

ここで，CHLAC 特徴では，参照する画素が重複しているパターンも多数存在する．例えば，0 次のパターンが参照する注目画素は，すべてのパターンで参照される．そのため，CHLAC 特徴のパターンの共起には強い相関がある．このとき，主成分分析 (PCA;Principal Component Analysis) による白色化が有効である．PCA で白色化された CHLAC 特徴は，無相関な主成分のベクトルで表される．結合された CHLAC 特徴を PCA によって変換することで，各フレーム f_k に対する特徴量 v_k として参照する．

次に，特徴量 v_k を統合して映像区間に対する動作特徴 x_m を得る方法を考える．このとき，統合を単純に特徴量の和をとることで行い，$x_m = \sum_{k=s_{m-1}}^{s_m} v_k$ とする方法が挙げられる．フレーム単位の特徴量 v_k は，各次元の組み合わせで表現される．よって，動作特徴 x_m の各次元の値は映像区間内の各フレームから抽出される v_k の対応する次元の和で決定する．この場合，フレーム単位の特徴量 v_k が異なっていても，映像区間内での和が同じになってしまう場合が考えられる．これでは v_k の情報が欠損してしまう．

ベクトル量子化によってこの問題に対処する．特徴量 $\{v_{s_1}, ..., v_{s_K}\}$ をクラスタリングし，それぞれのクラスタを最小単位の動きと考え，動作プリミティブと呼ぶ．各フレームの特徴量が全く異なる場合，この出現回数の和を取ることで，前述の情報欠損をさけて統合することができる．クラスタの総数を I とし，動作プリミティブを $\{p^1, ..., p^I\}$ で表す．

ここで，フレーム単位の特徴量 v_k に対する動作プリミティブを次式で表す．

$$p_k = \underset{0<i\leq I}{\operatorname{argmax}} P(p^i|v_k) \tag{1}$$

このとき，映像区間 S_m 中に含まれる動作プリミティブ $\{p_{s_{m-1}}, ..., p_{s_m}\}$ を数え上げることで得られるヒストグラムを用いて，映像区間に対する動作特徴 x_m を定める．

物体の把持・解放で区切られた映像区間の長さ l_m は加工動作毎の作業時間の長さに対応している．そこで，映像区間の長さ l_m を考慮して v_k を統合する．このとき，動作プリミティブを数え上げることで得られるヒストグラムの要素数

は映像区間の長さ l_m に比例する．しかし，映像区間の長さが一定以上の時，l_m は動きの類似性を判定する上で重要でない．これは，作業時間が一定以上長くなる要因としては，加工動作の種類よりも，加工の対象となる食材の種類や量が想定されるからである．そこで，ヒストグラムの各成分を映像区間の長さ l_m に応じた正規化項 Z_m で割る．このとき，Z_m を映像区間が長いほど大きな値を持ち，かつ，一定以上の長さ以上ではその値の変位が小さくなるような，l_m を変数とした関数で定義する．本稿では，そのような正規化項を次式で定める．

$$Z_m = \log(l_m) \tag{2}$$

このとき，映像区間に対する動作特徴は次式で表せる．

$$\begin{aligned} x_m &= h(S_m) = [x_1^m, ..., x_I^m] \\ x_i^m &= \frac{\sum_{k=s_{m-1}}^{s_m} \mathbb{1}(p^i = p_k)}{Z_m} \end{aligned} \tag{3}$$

ここで, 動作プリミティブの取得の為のクラスタリング手法として, 本研究では Dirichlet Process を用いた混合ガウス分布 (DPGMM;Dirichlet Process Gaussian Mixture Model) へのモデルフィッティング手法を利用する．クラスタは GMM を構成する各正規分布に対応する．しかし，調理映像に対する最適な動作プリミティブは定義されていない．これに対して，DPGMM は自動でクラスタ数を決定してくれる．このとき，DPGMM が決定するクラスタ数は局所解であり，初期値によって得られるクラスタ数にはばらつきが生じる．DPGMM によるクラスタリングを複数回行い，クラスタの総数 I が最小となる結果を利用して，動作プリミティブ $\{p^1, ..., p^I\}$ を得る．

4.4 映像区間のクラスタリング

調理映像 F に対する映像区間の集合 $\{S_1, ..., S_M\}$ を映像区間から抽出される動作特徴 $\{x_1, ..., x_M\}$ の類似性に基づいてクラスタリングする．このクラスタリングによって，抽出される動作特徴が類似した映像区間をまとめあげた区間クラスタ C_n を得る．

まず，映像区間ごとの動作特徴 x_m の類似性を定める．4.3 節より，x_m はヒストグラムで表すことができて，$[x_1^m, ..., x_I^m]$ である．そこで，次式で示される，

x_m と $x_{m'}$ のヒストグラムの交差を利用する．

$$d_{is}(v_m, v_{m'}) = \frac{2\sum_{i=1}^{I} min(x_i^m, x_i^{m'})}{\sum_{i=1}^{I} x_i^m + \sum_{i=1}^{I} x_i^{m'}} \tag{4}$$

このとき，$0 < d_{is}(v_m, v_{m'}) \leq 1$ である．また，$d_{is}(v_m, v_{m'})$ が 1 に近いほど，x_m と $x_{m'}$ は類似している．そこで，経験的に得られたしきい値 t_h を用いて，抽出される動作特徴が類似していることを次のように判定する．

$$d_{is}(v_m, v_{m'}) > t_h \tag{5}$$

同種の加工動作が行われる映像区間では抽出される動作特徴が類似する．ここで，ある映像区間 S_{m_1} とそれとは別の映像区間 S_{m_2} で同種の加工動作が行われていれば，$d_{is}(v_{m_1}, v_{m_2}) > t_h$ となる．このとき，映像区間 S_{m_3} でもその加工動作が行われていれば, $d_{is}(v_{m_1}, v_{m_3}) > t_h$ と $d_{is}(v_{m_2}, v_{m_3}) > t_h$ がともに成立する．

このような処理を行うことができるクラスタリング手法として，クラスタ間の類似度を最遠隣法によって求める階層的クラスタリングが挙げられる．本研究ではこれを用いる．新たに併合されるクラスタのクラスタ間類似度が t_h より小さい時，クラスタリングを終了する．その時点でのクラスタをそれぞれ区間クラスタとして取得する．

第5章　実験

4 章では，調理映像からの動作区間検出を行う手法の具体的な処理を述べた．本章では，提案手法を用いた実験とその結果・考察について述べる．

5.1　実験目的

提案手法によって，抽出される動作特徴が類似した映像区間は 1 つの区間クラスタにまとめあげられる．このとき，任意の区間クラスタ C_n に属する映像区間がある 1 種類の加工動作か加工動作以外のいずれか一方に対応していれば，本研究の目的は達成される．これに対して，本稿では，動作区間検出における映像分割の方法として物体の把持・解放を提案した．そこで，映像分割法として物体の把持・解放を利用した場合と固定幅を利用した場合とで得られる区間クラスタを比較し，提案手法の有用性を確かめる．

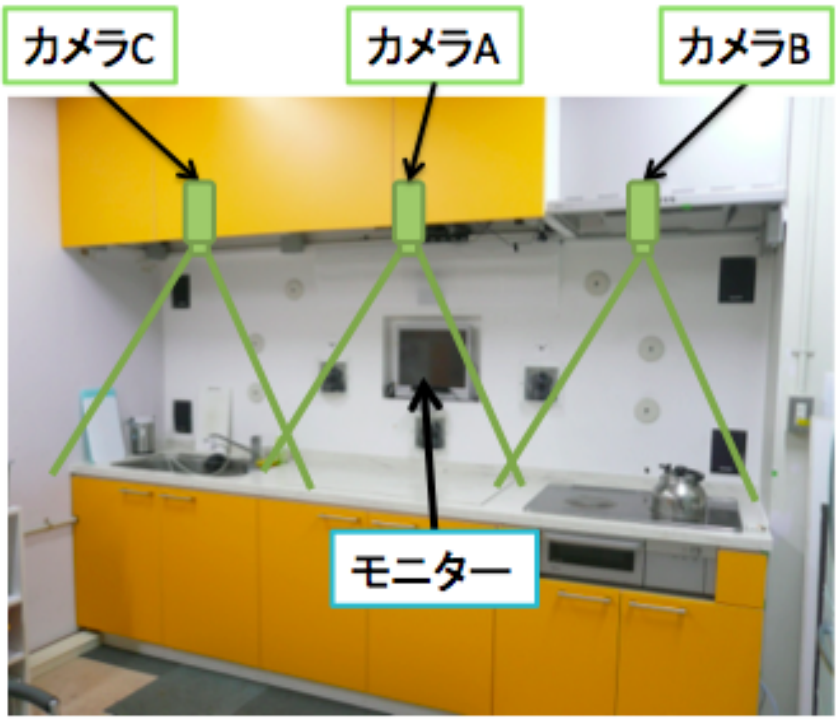

図 8: KUSK DATASET の観測環境

5.2 実験方法

5.2.1 データセット

本実験では，データセットとして KUSK DATASET[12] を利用した．このデータセットの調理映像には，1 人の調理者があるレシピに従って行った調理の様子が調理に必要な食材や調理器具を調理台に置くところから，料理が完成したことを調理者が判断した時点まで記録されている．調理の様子は異なるカメラ A，B，C によって，図 3 の「調理台」,「コンロ」,「シンク」ごとのシーンを撮影している．また，KUSK DATASET では撮影時のストレージの節約のため，映像中で動きのない部分は除去されている．従って，提案手法に適用する際はフレーム間に動きのない画像を補完することで対処した．

調理者が参照するレシピには COOKPAD に公開されているレシピが使用されている．一般的に，COOKPAD のレシピには，完成した料理画像，使用する材料の一覧，文章による調理手順が記述されている．調理者は調理台に設置されたモニターでこのようなレシピを確認しながら調理を行う．KUSK DATASET で公開されている調理映像の観測環境を図 8 に示す．

表 1: 調理映像のレシピと観測時間

調理映像	レシピ名	観測時間（分 :秒）
F_1	アツアツとろ〜り白菜と鶏肉のスープ	27:12
F_2	押し麦入り和風とマトスープ。	45:31
F_3	和風のし鶏	43:34

本実験では，KUSK DATASET から調理者，レシピがともに異なる 3 つの調理映像 F_1，F_2，F_3 を対象として提案手法に適用した．それぞれの調理映像について，調理者が参照するレシピと観測時間を表 1 に示す．映像のフレームレートはすべて 30 フレーム/秒である．また，調理映像から CHLAC 特徴を抽出する際の計算を高速化するため，元々は 1040 × 776 であった画面のサイズを 400 × 300 に縮小した．

5.2.2 加工動作のラベル付け

本研究では，加工動作が行われた映像区間をレシピ中のその加工動作に対応する動詞に付与することを想定して，調理映像からの動作区間検出を行った．このとき，実際に付与される映像区間は加工方法や個人ユーザの好みによって異なると考えられる．例えば,「切る」に付与される映像区間として，実際に食材を切っている様子は必要であるが，包丁を食材のほうに持っていく，切った食材をボウルに移す，等は必ずしも必要ではない．

付与される映像区間にはこのような曖昧性が存在し，そのような影響を受けずに区間クラスタを評価するために，本実験では，食材に分断等の非可逆な変化が継続して起きている映像区間だけに注目する．ここで，加工している際に一瞬食材から手を離す等を想定される．そこで，１秒未満の中断は無視して加工の継続を判断した．このような映像区間に加工動作としてのラベルを付ける．

加工動作の種類としてはレシピ中の動詞が考えられる．ここで，4.3 節で述べたように,「調理台」.「コンロ」,「シンク」で行われる加工動作は設備によってそれぞれ目的は異なると考えた．しかし，一部の加工動作では，対応する動詞が同じでも，異なる設備で行われる加工動作が存在した．本研究の動作特徴抽出では，このような加工動作を同じ区間クラスタにまとめあげることを想定していない．そこで，それぞれの加工動作をレシピ中の動詞とその加工を観測したカメラの組み合わせで定める．調理映像ごとのこのような組み合わせの総数

を L とし，それぞれの加工動作を加工動作 $l(1 \leq l \leq L)$ と表す．調理映像中のフレーム $f_k(1 \leq k \leq K)$ へのラベル付けを次式で定める．

$$a_k^l = \begin{cases} 1, & (f_k\text{が加工動作 } l \text{ とラベル付けされた}) \\ 0, & (otherwise) \end{cases} \tag{6}$$

5.2.3 比較対象の映像分割法

比較対象となる映像分割法 g としては，5 秒，つまり，150 フレームを固定幅とした方法 g_{eq150} ，10 秒，つまり，300 フレームを固定幅とした方法 g_{eq300} ，検出された物体の把持・解放による方法 g_{access} の 3 種類を用意した．

5.3 結果・考察

調理映像 F_1，F_2，F_3 から，映像分割法 g_{access}，g_{eq150}，g_{eq300} を用いた場合に取得された区間クラスタを比較した．まず，5.3.1 項で取得された区間クラスタの評価方法を説明する．5.3.2 項で実験結果を示し，その考察を述べる．

5.3.1 評価方法

本研究の目的は，調理映像から取得された任意の区間クラスタに属する映像区間がある 1 種類の加工動作か加工動作以外のいずれか一方に対応することで達成される．そこで，各区間クラスタ $C_n(1 \leq n \leq N)$ に属する映像区間が加工動作 $l(1 \leq l \leq L)$ ，あるいは，加工動作以外に対応しているかどうかを 5.2.2 項の加工動作のラベル付けの結果に基づいて評価し，目的が達成できたかどうかを確かめる．区間クラスタ C_n に属する映像区間中の全フレームにおいて，加工動作 l とラベル付けされたフレームの総数 A_n^l は次式で算出される．

$$A_n^l = \sum_{S_m \in C_n} \sum_{k=s_{m-1}}^{s_m - 1} a_k^l \tag{7}$$

このとき，区間クラスタ C_n に属する映像区間中のフレームに対する A_n^l の割合，すなわち，適合率によって評価することが考えられる．この場合，任意の l で $a_k^l = 0$ となるフレームはその加工を説明する上で不要であると考えている．しかし，そのようなフレームでは食材に非可逆な変化が起きていないだけであり，5.2.2 項でも述べたように，加工方法や個人ユーザによってはその加工を説明する際に必要となる可能性が十分考えられる．従って，このような適合率による評価は本実験には適さない．

ここで，同種の加工動作は 1 つの動詞に対応すると考えられる．そのため，同

種の加工動作はできるだけ 1 つの区間クラスタにまとめあげられることが好ましい．よって，区間クラスタ C_n に属する映像区間が加工動作 l に対応している場合，それ以外の区間クラスタ $C_{n'}(n' \neq n)$ は加工動作 l に対応しないべきである．また，異なる加工動作は別々の動詞に対応すると考えられるので，別々の区間クラスタにまとめあげるべきである．

区間クラスタがこのような性質を持っているかどうかを評価することで，本研究の目的が達成できたかどうかを確かめる．この評価を $1 \leq l \leq L$ での A_n^l の値の比較に基づいて行うことが考えられる．しかし，加工動作が調理映像中で行われる時間の長さはその種類によって異なるので，単純な大小比較ではそのような時間の長さの差異の影響を受けてしまう．そこで，調理映像中の全フレームで加工動作 l とラベル付けされたフレームの総数 A_{total}^l に対する A_n^l の割合，すなわち，再現率 A_n^l/A_{total}^l を比較する．A_{total}^l は次式で算出される．

$$A_{total}^l = \sum_{k=1}^{K} a_k^l \tag{8}$$

5.3.2 実験結果

調理映像 F_1，F_2，F_3 から取得された各区間クラスタの加工動作 l の再現率 A_n^l/A_{total}^l を表 2〜10 に示す．表を見やすくするため，$A_n^l/A_{total}^l \leq 0.1$ となる区間クラスタはまとめあげた．5.2.2 項で述べたように，加工動作はレシピ中の動詞とその加工を観測したカメラの組で表されている．これ以降，加工動作は”切る-カメラ A”のように”動詞-カメラ”で表す．

まず，同種の加工動作ができるだけ 1 つの区間クラスタにまとめあげられるかどうかを評価した．すべての調理映像で行われた”切る-カメラ A”に注目して説明する．固定幅 g_{eq150}，g_{eq300} の場合（表 3，4，6，7，9，10）をみると，”切る-カメラ A”は主に 2 つ，あるいは，3 つの区間クラスタに分かれ，それらの区間クラスタに含まれない”切る-カメラ A”は多数の区間クラスタに散らばってしまった．これに対して，提案手法 g_{access} の場合（表 2，8）をみると，”切る-カメラ A”はほぼ 1 つの区間クラスタにまとめあげられた．

g_{access} でも”切る-カメラ A”が 2 つの区間クラスタ C_1，C_2 に分かれてしまった例として，F_2 に対する結果（表 5）を取り上げる．この場合でも，g_{eq150}，g_{eq300} のように切る-カメラ A”は多数の区間クラスタに散らばることはなかった．F_2 中の”切る-カメラ A”では，切られる食材が F_1，F_3 よりも多く，”切る-カメラ A”が行われた時間が長かった．これは本研究の映像区間のクラスタリングの手

表 2: F_1 から g_{access} で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	切る カメラ C	まぶす カメラ A	はたく カメラ A	溶かす カメラ A	入れる カメラ B
C_1	**0.92**	0.00	**0.74**	**1.00**	0.00	0.01
C_2	0.00	**0.59**	0.00	0.00	0.00	0.00
C_3	0.00	**0.22**	0.00	0.00	0.00	0.00
C_4	0.00	**0.19**	0.00	0.00	0.00	0.00
C_5	0.05	0.00	**0.26**	0.00	0.00	0.00
C_6	0.00	0.00	0.00	0.00	**1.00**	0.00
C_7	0.00	0.00	0.00	0.00	0.00	**0.34**
C_8	0.00	0.00	0.00	0.00	0.00	**0.24**
C_9,⋯,C_{60}	0.03	0.00	0.00	0.00	0.00	**0.41**

表 3: F_1 から $g_{equal150}$ で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	切る カメラ C	まぶす カメラ A	はたく カメラ A	溶かす カメラ A	入れる カメラ B
C_1	**0.25**	0.00	0.00	0.00	**0.15**	0.02
C_2	**0.20**	0.00	**0.55**	**1.00**	**0.29**	0.05
C_3	**0.11**	0.00	0.09	0.00	**0.26**	0.06
C_4	0.00	**0.36**	0.00	0.00	0.00	0.03
C_5	0.04	**0.33**	0.00	0.00	0.00	0.06
C_6	0.08	**0.31**	0.00	0.00	0.00	0.03
C_7	0.04	0.00	**0.18**	0.00	0.00	0.00
C_8	0.07	0.00	**0.18**	0.00	0.00	0.01
C_9	0.00	0.00	0.00	0.00	**0.29**	**0.17**
C_{10}	0.00	0.00	0.00	0.00	0.00	**0.11**
C_{11}	0.00	0.00	0.00	0.00	0.00	**0.10**
C_{12},⋯,C_{23}	**0.22**	0.00	0.00	0.00	0.00	**0.36**

表 4: F_1 から $g_{equal300}$ で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	切る カメラ C	まぶす カメラ A	はたく カメラ A	溶かす カメラ A	入れる カメラ B
C_1	**0.41**	0.00	**0.36**	0.00	**0.59**	**0.14**
C_2	**0.31**	0.00	0.00	0.00	0.00	0.05
C_3	**0.12**	0.00	**0.19**	**0.51**	0.00	0.03
C_4	0.05	**0.69**	0.00	0.00	0.00	0.00
C_5	0.00	**0.31**	0.00	0.00	0.00	0.06
C_6	0.04	0.00	**0.45**	**0.49**	**0.15**	**0.20**
C_7	0.00	0.00	0.00	0.00	**0.26**	**0.17**
C_8	0.05	0.00	0.00	0.00	0.00	**0.11**
C_9	0.00	0.00	0.00	0.00	0.00	**0.10**
C_{10} ,⋯, C_{13}	0.03	0.00	0.00	0.00	0.00	**0.13**

表 5: F_2 から g_{access} で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	炒める カメラ B	溶かす カメラ B	入れる カメラ B
C_1	**0.59**	0.00	0.00	0.00
C_2	**0.34**	0.00	0.00	0.00
C_3	0.00	**0.63**	0.00	0.00
C_4	0.00	**0.33**	0.00	0.04
C_5	0.00	0.00	**0.96**	**0.29**
C_6	0.00	0.00	0.00	**0.38**
C_7,⋯,C_{53}	0.07	0.04	0.04	**0.29**

表 6: F_2 から $g_{equal150}$ で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	炒める カメラ B	溶かす カメラ B	入れる カメラ B
C_1	**0.21**	0.00	0.00	0.01
C_2	**0.21**	0.00	0.00	0.00
C_3	**0.16**	0.04	0.00	0.01
C_4	0.00	**0.29**	0.06	0.07
C_5	0.00	**0.28**	0.00	0.07
C_6	0.00	**0.19**	0.00	0.09
C_7	0.03	0.09	**0.24**	0.04
C_8	0.07	0.00	**0.18**	0.00
C_9	0.00	0.00	**0.18**	0.00
C_{10}	0.00	0.00	**0.15**	**0.13**
C_{11}	0.06	0.00	**0.12**	0.05
C_{12}	0.01	0.00	0.00	**0.17**
C_{13}	0.00	0.04	0.06	**0.13**
C_{14},⋯,C_{25}	**0.21**	0.02	0.00	**0.13**

表 7: F_2 から $g_{equal300}$ で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	炒める カメラ B	溶かす カメラ B	入れる カメラ B
C_1	**0.31**	0.05	0.00	0.00
C_2	**0.25**	0.04	0.00	**0.27**
C_3	**0.16**	0.00	0.00	0.07
C_4	0.00	**0.44**	**0.27**	0.04
C_5	0.00	**0.36**	0.00	0.08
C_6	0.00	**0.11**	**0.12**	0.04
C_7	0.02	0.00	**0.61**	0.00
C_8	0.06	0.00	0.00	**0.14**
C_9	0.03	0.00	0.00	**0.13**
C_{10}	0.03	0.00	0.00	**0.12**
C_{11},⋯,C_{15}	**0.16**	0.00	0.00	**0.12**

表 8: F_3 から g_{access} で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	ささがきする カメラ A	ささがきする カメラ C	混ぜる カメラ A	ふりかける カメラ A
C_1	**0.97**	0.00	0.00	0.00	0.00
C_2	0.00	**1.00**	0.00	0.00	**0.44**
C_3	0.00	0.00	**1.00**	0.00	0.00
C_4	0.00	0.00	0.00	**1.00**	0.00
C_5	0.00	0.00	0.00	0.00	**0.32**
C_6	0.02	0.00	0.00	0.00	**0.25**
C_7,$\cdots$,C_{36}	0.01	0.00	0.00	0.00	0.00

表 9: F_3 から $g_{equal150}$ で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	ささがきする カメラ A	ささがきする カメラ C	混ぜる カメラ A	ふりかける カメラ A
C_1	**0.43**	0.00	0.00	**0.29**	0.00
C_2	**0.31**	0.00	0.00	0.08	0.00
C_3	0.00	**0.70**	**0.33**	0.00	**0.19**
C_4	0.00	**0.24**	**0.43**	0.00	0.00
C_5	0.05	0.00	**0.12**	**0.32**	0.00
C_6	0.05	0.00	0.00	**0.20**	0.00
C_7	0.00	0.00	0.00	0.00	**0.36**
C_8	0.02	0.00	0.00	0.00	**0.32**
C_9	0.00	0.00	0.00	0.00	**0.14**
C_{10},$\cdots$,C_{20}	0.08	0.00	0.08	0.08	0.00

表 10: F_3 から $g_{equal300}$ で取得された区間クラスタの各加工動作の再現率

区間クラスタ	切る カメラ A	ささがきする カメラ A	ささがきする カメラ C	混ぜる カメラ A	ふりかける カメラ A
C_1	**0.56**	0.00	0.00	**0.40**	0.00
C_2	**0.23**	0.00	0.00	0.00	0.00
C_3	**0.12**	0.00	0.00	**0.17**	0.00
C_4	0.01	**0.70**	**0.23**	0.00	0.00
C_5	0.00	**0.19**	**0.37**	0.00	0.00
C_6	0.00	**0.11**	0.00	0.03	0.00
C_7	0.05	0.00	**0.20**	0.08	**0.64**
C_8	0.01	0.00	**0.17**	**0.32**	0.00
C_9	0.00	0.00	0.00	0.00	**0.36**
C_{10},$\cdots$,C_{13}	0.02	0.00	0.03	0.01	0.00

法が原因であると考えられる．4.4 節で述べたように，本研究では最遠隣法による階層的クラスタリングを用いた．そのため，加工動作の作業時間が長くなるほど，その加工動作は別々のクラスタにまとめあげられてしまいやすくなる性質を持つのである．

”切る-カメラ A”以外の加工動作についても同様の結果が得られた．多くの加工動作について，同種の加工動作をより少ない区間クラスタでまとめあげられることができた g_{access} は g_{eq150}，g_{eq300} よりも有効であると考えられる．例外としては，F_1，F_2 で行われた”入れる-カメラ B”が挙げられる．g_{access} の場合（表 2，5）でも，”入れる-カメラ B”は主に 2 つの区間クラスタに分かれ，それらの区間クラスタに含まれない”入れる-カメラ B”は多数の区間クラスタに散らばってしまった．実際の映像中では，”入れる-カメラ B”は，食材や調味料を鍋に入れる作業に対応した．これは食材や調味料などの物体を別の場所に移すだけで行われる単純な加工であり，映像付与の重要度は低いと考えられる．

次に，異なる加工動作が別々の区間クラスタにまとめあげられたかどうかを評価した．F_2，F_3 について，g_{access} の場合（表 5，8）をみると，ほぼすべての加工動作を種類ごとに別々の区間クラスタにまとめあげることに成功した．例外としては，F_2 で行われた”入れる-カメラ B”と F_3 で行われた”ふりかける-カメラ A”が挙げられる．”入れる-カメラ B”については先ほども述べたようにあまり重要でない．表 8 をみると，区間クラスタ C_2 は”ふりかける-カメラ A”と”ささがきする-カメラ A”を区別できていない．ここで，C_2 では”ささがきする-カメラ A”の再現率が 1.00 であることから，本研究の特徴量では”ふりかける-カメラ A”の一部が”ささがきする-カメラ A”と類似したと考えられる．g_{eq150}，g_{eq300} の場合（表 6，7，9，10）をみると，”入れる-カメラ B”や”ふりかける-カメラ A”以外の加工動作についても区別できていない区間クラスタが存在し，g_{access} の方が有効であることが示された．

F_1 の場合（表 2，3，4）をみると，g_{eq150}，g_{eq300} では区別できていなかった”溶かす-カメラ A”が，g_{access} では区間クラスタ C_6 にまとめあげることに成功していた．この点では，g_{access} の方が有効であると考えられる．しかし，F_1 の加工動作のうち，”切る-カメラ A”，”まぶす-カメラ A”，”はたく-カメラ A”の 3 種類は区別できなかった．g_{access} の場合（表 2）をみると，区間クラスタ C_1 これら 3 つの加工動作がまとめあげられている．g_{eq150}，g_{eq300} の場合（表 3，4）でも，”切る-カメラ A”，”まぶす-カメラ A”，”はたく-カメラ A”を区別できな

い区間クラスタが存在した．

映像中で，これら 3 つの加工動作は調理台に置かれたまな板の上で調理者の位置があまり変わらず連続して行われた．本研究で，調理映像からの特徴量抽出として用いた CHLAC 特徴は動きの大きさが小さいとそのような動きの違いを反映しにくい．そのため，調理者の位置があまり変化しないような加工動作を区別することは難しく，いずれの映像分割法でも区別できない加工動作が存在してしまった．本研究では，単純な RGB カメラを用意するだけで使用可能な CHLAC 特徴を用いた．加工動作の細かい違いを反映できるような特徴量を得る為には，深度カメラを利用したキネクトなどを利用することで対処できると想定される．

第 6 章　おわりに

本研究では，レシピ作成者である個人ユーザが調理中の作業の様子を記録した映像クリップをレシピに付与することを想定した．このとき生じる個人ユーザの負担を軽減する為に，調理映像からの動作区間検出によって映像クリップを得ることを考え，これを本研究の目的とした．

従来の動作認識等を用いた手法では事前に加工動作を限定する必要があった．そこで，本研究では，調理映像を動作特徴を抽出するための映像区間に分割し，レシピ作成者が各クラスタに加工動作のラベルを割り当てることを想定して，抽出された動作特徴の類似性に基づいてクラスタリングした．この方法は事前に加工動作を限定せず，どのような調理映像にも適用できる．

映像分割は最初の処理となり，どのような区切り目の集合 $\{s_0, ..., s_M\}$ で行うかが重要であった．しかし，それぞれの映像区間の長さ $l_m = s_m - s_{m-1}$ を固定幅とした，単純な映像分割では加工動作ごとの長さに対応できない．そこで，本研究では，加工動作の区切れ目となる物体の把持・解放を利用することを提案した．加工動作の前後では，食材や調理器具の把持・解放が起きる．逆に，加工動作中には，そのような物体の把持・解放は起きにくい．そのため，物体の把持・解放を区切り目として調理映像を分割すると，加工動作ごとの長さに対応した映像区間を得ることができる．

実験では，KUSK DATASET の調理映像から区間クラスタを取得する際，提案手法の物体の把持・解放を利用した場合と固定幅を利用した場合を比較する

ことで，提案手法の有用性を確かめた．

今後の課題としては，5.3.2 項で述べた映像区間から抽出される動作特徴の改善が挙げられる．また，提案手法のクラスタリング結果に基づいてレシピ中の動詞と区間クラスタのマッチングを行い，レシピへの映像クリップの付与を自動で行うことが考えられる．

謝辞

本研究を進めるにあたり，多くのご教示を賜りました美濃導彦教授に深く感謝いたします．日頃より研究の方向性について助言を賜りました椋木雅之准教授に深く感謝いたします．さらに，日頃より直接ご指導いただき，多大なるお力添えをいただきました橋本敦史助手に心より感謝致します．最後に，本研究に対して多くの助言を頂きました美濃研究室の皆様に感謝致します．

参考文献

[1] クックパッド株式会社: 毎日の料理を楽しみにCOOKPAD, http://cookpad.com/.

[2] Allrecipes.com: allrecipes.com, http://allrecipes.com/.

[3] INFOCOM CORPOLATION: AmKitchen, http://www.infocom.co.jp/english/.

[4] KSCGR Contest: Kitchen Scene Context Based Gesture Recognition, http://www.murase.m.is.nagoya-u.ac.jp/KSCGR/.

[5] Lei, J., Ren, X. and Fox, D.: Fine-grained kitchen activity recognition using RGB-D, *Proc. of the 2012 ACM Conference on UbiComp2012*, pp. 208–211 (2012).

[6] 久原卓, 出口大輔, 高橋友和, 井手一郎, 村瀬洋: CHLAC 特徴の周期性解析による料理映像中の繰り返し調理動作区間の抽出と識別, 電子情報通信学会技術研究報告. MVE, マルチメディア・仮想環境基礎, Vol. 110, No. 457, pp. 61–66 (2011).

[7] 林泰宏, 道満恵介, 井手一郎, 出口大輔, 村瀬洋: 料理レシピの記述に従った家庭内調理映像の自動要約, 信学技報, Vol. 112, No. 474, pp. 121–126 (2013).

[8] 道満恵介, カイ承穎, 高橋友和, 井手一郎, 村瀬洋: マルチメディア料理レシピ作成のための料理レシピテキストと料理番組映像との対応付け, 電子情

報通信学会論文誌. A, 基礎・境界, Vol. J94-A, No. 7, pp. 540–543 (2011).

[9] 山肩洋子, 角所考, 美濃導彦: 調理コンテンツの自動作成のためのレシピテキストと調理観測映像の対応付け, 電子情報通信学会論文誌. D, 情報・システム, Vol. J90-D, No. 10, pp. 2817–2829 (2007).

[10] 橋本敦史, 舩冨卓哉, 中村和晃, 美濃導彦: 机上物体検出を対象とした接触理由付けによる誤検出棄却, 電子情報通信学会論文誌. D, 情報・システム, Vol. J95-D, No. 12, pp. 2113–2123 (2012).

[11] Li, Y. and Luo, J.: TASK-RELEVANT OBJECT DETECTION AND TRACKING, *ICIP'13*, pp. 3900–3904 (2013).

[12] Hashimoto, A., Sasada, T., Yamakata, Y., Mori, S. and Minoh, M.: KUSK Dataset: Toward a Direct Understanding of Recipe Text and Human Cooking Activity, *Proc. of Workshop on Smart Technologyfor Cooking and Eating Activities in conjunction with UbiComp2014*, pp. 583–588 (2014).