I-045

映像インデキシングのための注目領域検出と言語との対応付けの検討

Detecting Focus of Attention from Images and Transcript and Its Utilization for Video Indexing

山本 治由† 中村 裕一‡ 大田 友一§
Haruyoshi Yamamoto Yuichi Nakamura Yuichi Ohta

1. はじめに

コンピュータやインターネットの進歩により，大量の映像を記録し，扱うことが容易にできるようになった．それにともない，映像の利便性を向上させるために，映像を自動的に要約する方法 [1] や，映像の一覧性を高める方法 [2] が盛んに研究されている．そのためには，映像の全体や部分に対してインデックスを付与することが有効であるが，このような情報を映像中の全フレームにつける作業は非常に手間がかかる．

本研究の目的は，このような問題を解決するために，映像中の重要な部分，つまり人が注目しそうな部分（注目領域）を自動検出し，それによって，インデキシングを自動化することである．その基本的な考え方を図１に示す．画像から注目領域を検出し，クローズドキャプションとして映像に付加された発話情報と関連づけることにより，映像のインデックスとする．以下本稿では，その考え方と実験結果について述べる．

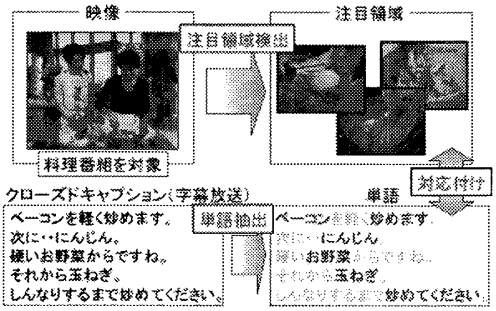

図 1: インデキシング自動化の概要図

2. 注目領域と映像インデキシング

映像には，その制作者が伝えようとした以外の情報も必然的に含まれてしまう．例えば，手に持った物体を説明する映像を撮影する時にも，それを持った人やあまり重要でない背景が写る．そのため，映像制作者が主に伝えたい部分を強調するような映像構成が必要となる．逆に言うと，強調されている部分，目立つ部分が，映像制作者の伝えたい，注目して欲しい部分であると考えられる．

例えば，動き（人が歩いていたり何か作業している姿）は注目を引くため，他に注目すべき対象がある場合に，不用意にこれらの対象を画像中に入れることは少ない．また，画面中央に写っている対象は，その部分がより注目されやすいため，説明の対象となっていることが多い．対象が説明されているトピックに対して重要な特徴 (例えば，色や形など) を持っている場合にも同様のことが言える．

†筑波大学 大学院 システム情報工学研究科
‡京都大学 学術情報メディアセンター
§筑波大学 システム情報工学研究科

本研究では，このように，画像中で視聴者が注目するであろう部分を注目領域と呼び，これを抽出して発話などと関係付けることによりインデックス情報として用いる．

3. 注目領域の検出

本研究では料理や科学実験などの机上作業映像を対象にする．これらは映像マニュアルとして利用価値が高いからである．対象とする料理番組には２種類のショットが良く用いられる．

全体ショット: 料理人とアシスタントが立っている場面全体を入れるようにロングショットで撮影している．

手元ショット: 料理をする人の手元や物体を撮影したクロースアップショット．

全体ショットは漠然とした話をしている時や場面転換で挿入されることが多く，画像自体には作業や材料に関する情報があまり含まれていない．それに対し，手元ショットには作業や材料に関する詳細な情報が含まれるため，本研究ではこのショットのみを抜き出してその後の処理を行う．これらのショットの判別をするためには，フレーム間差分を利用してショットの切り替えを検出する．ショット切り替えを検出した後，その色ヒストグラムによって，どちらのショットに属するかを判定する．

3.1 静止物体の検出

静止物体の検出は，物体がしっかり撮影されている，つまり物体の隠れがなく，動きが少ないフレームに対して処理するのが望ましい．よって，画面の切り替わり直後や動領域が一定時間以上検出されない時に処理を行う．

料理番組に登場する物体には，材料や調理器具がある．その中で，色が特徴的である，野菜・手・皿・鍋を検出する．また，手元ショットには手が含まれていることが多く，手として検出された領域は動作に関係があることが多い．そのために，手も検出する．以下の方法で検出し，注目領域とする．

色分布による検出（図 2）

手（肌色）と野菜（緑色）を検出する．まず領域分割を行い，分割後の各領域の RGB 値の平均を求める．そして，あらかじめ統計的に求めた肌色と緑色の色分布に近い平均値を持つ領域を注目領域とする．

彩度による検出（図 3）

皿や鍋などは白色や灰色であることが多いので，彩度が低い箇所として検出できる．輝度が低いと彩度が大きくなりやすいので，輝度値が一定値以下なら除去する．彩度を二値化し，面積が小さい領域はノイズとして除去する．

表 1: 注目領域と単語の対応付け

	注目領域の全検出数	注目領域からみた単語の対応				
		対応する単語がある				対応する単語がない
		注目領域と同時	注目領域出現前 5 秒以内	注目領域消失後 5 秒以内	その他	
映像 1	74	17	13	6	4	34
映像 2	63	13	10	5	20	25
映像 3	56	11	10	5	9	21

表 2: 単語と注目領域の対応付け

	単語の全検出数	料理に関係するもの	単語からみた注目領域の対応				
			対応する注目領域がある				対応する注目領域がない
			単語と同時	単語検出前 5 秒以内	単語検出後 5 秒以内	その他	
映像 1	137	62	12	3	2	25	20
映像 2	190	63	9	2	2	31	19
映像 3	151	69	10	0	1	31	27

図 2: 色分布で検出（緑色）

図 3: 彩度による検出

3.2 動物体と動作の検出

輝度値のフレーム間差分をしきい値で二値化し，膨張・収縮，ラベリングを行って面積が小さいものをノイズとして除去する．ここで検出された部分を動領域とする．動領域が近接した位置に連続して検出された場合は一連の動作であると考えられる．また，動領域が連続していなくても，消失から発生までの間隔が短い場合は一連の動作であると考えられる．そのため，ショット切り替えと動領域が検出されないフレームで映像を分割する．動領域が検出されるシーケンスを抜き出し，その中で隣接フレームの動領域が重なっているものを一連の動作として検出する．そして一連の動作を示す領域が，短い間隔(本システムでは 0.3 秒以内) で近くで消失・出現している場合，それらをまとめて一連の動作とする．

4. 注目領域の意味付け

本研究ではクローズドキャプションから単語を抜き出し，前節で述べた注目領域と関係付けることによって映像のインデックスとする．実験に使用した料理番組は一人で説明しながら作業を進める場面が多く，その発話がクローズドキャプションとなっている．そこから材料名，動作などを抜き出す．現在は，クローズドキャプションを形態素解析し，人手で修正したものを利用している．

一連の注目領域に対応づける単語は，その時間的に近接しているものとする．一個の領域に対して，複数の単語が対応付けられることや，長時間の動作は対応する単語の候補が増え，対応を取るのが困難になる問題がある．解決策として過去の対応関係の情報を利用して候補から選択する方法もあるが，選択方法については今後の課題である．

5. 実験例

料理番組映像を実際に処理し，注目領域によるインデキシングの可能性を確かめた．使用した映像は，CM を除いた 8 分の料理番組 3 本である．クローズドキャプションは発話のみで，シーンの説明や材料の紹介などの付加情報は含まれていない．クローズドキャプションを形態素解析した結果から物体の名前と動作を手動で抜き出して利用した．

各映像について，注目領域とキャプション中の単語の対応を表 1 に示す．検出した注目領域を，対応する単語があるものとないものに分類した．さらに，対応する単語があるものは，注目領域が検出された時間内に単語があるものと，注目領域の出現前または消失後の 5 秒以内に単語があるものに分類した．逆に，キャプション中の単語に対する注目領域の対応を求めた結果を表 2 に示す．

全単語のうち，料理に関係する単語のみを手動で抽出した．そして各々の単語についても，注目領域が前後 5 秒以内にあるもの，ないものに分類し，その対応を調べた．

注目領域とそれに対応する単語が近接しない原因としては，一連の動作が途切れて複数回として検出されてしまうことがあげられる．例えば，動作の間に別の動作が割り込んだり，カメラが料理をする人の手元を追従する場合である．この問題に対しては，動作の特徴（動領域が円を描く，直線的であるなど）を調べて，連続する動きが同じ動作かどうかを判別するという方法で対応できると考えられる．

表 1 から，注目領域に対応する単語は，注目領域検出と同時かそれよりも前に現れることが多いことがわかる．これは作業行程を確認してから実際の作業をすることが多いからであると考えられる．また，現在は料理と関係の浅い単語も対応付けに利用しているが，これらを辞書やシソーラス等を用いて省くことができれば¶，対応付けの効率が良くなると考えられる．これらは今後の課題となっている．

6. おわりに

視聴者が注目する部分（注目領域）とクローズドキャプションを利用して，インデキシングを自動化する手法を提案した．実際の料理番組に対して，その有用性を示した．今後の課題として，注目領域検出の精度向上が挙げられる．現在は動作の検出に動領域のみを利用しているが，他の情報も取り扱うことで，より精度良く検出できると考えられる．

参考文献

[1] Yu-Fei Ma, Lie Lu, Hong-Jiang Ahang and Mingjing Li: “A User Attention Model for Video Summarization”, Proc. ACM Multimedia 2002,2002.

[2] 村山正司，伊津野克英，中村裕一，大田友一: “ビデオアイコンダイアグラムによる映像内容の構造表現”，信学技報 IE-2001-25,2001.

¶例えば，今回の料理に関係したもののみを選択する処理