ヒト認知特性の研究に基づく映像認知モデルの構築と提案

～映像解析理論モデルとの比較において～

Psychological Study of Cognitive Model of Movie Recognition and Comparison with Theoretical Movie-analysis Models

正会員 大山潤爾[†], 渡邊克巳^{††}
Junji Ohyama[†] and Katsumi Watanabe^{††}

Abstract We proposed a new cognitive visual model of movie recognition based on our previous findings of psychophysical phenomena. Our cognitive model suggested two important functions of movie recognition process. First, a continuous movie sequence was divided and perceived as serial event segments of short scenes. The movie would be coded for each of the segments and structuralized as a contextual association of each segments. Second, the knowledge structure of the context of previously viewed movies was used to predict the ongoing movie context and the online segmentation. We compared our cognitive model with a previously proposed theoretical model of movie processing. The results of our experiments supported our hypothesis: an adaptive learning mechanism of online movie segmentation would be effective for an intelligent knowledge-based structure of a future movie analysis system.

キーワード：映像認知, 視覚的文脈, 時系列情報処理, 分節化, イベント知覚

1. ま え が き

高速なネットワークや大容量データベース, 安価で高性能な映像機器の発展と普及により, 防犯カメラ映像記録システムや, YouTube¹⁾などの動画サイトのように, 膨大な「映像」コンテンツが利用可能になりつつある. このため, 膨大な映像コンテンツの記録や情報の検索に特化した特別な映像解析技術が求められている. 映像情報は, 空間的な「画像」情報であるとともに, その2次元情報が時間的な順序を持って連続的に変化する「時変」情報でもある. つまり, 映像の内容を解析するシステムの構築には, 映像中の空間的な視覚画像情報(すなわち, "何が映っているか")に加えて, 時系列文脈情報(すなわち, "どのような変化(イベント)がみられるか")の解析方法が重要となる. 膨大な映像を利用可能とするためには, 映像の時空間情報から, 利用者に有用な情報(場面)を効果的に抽出できる新しい映像解析システムが必要である. 近年, 空間情報処理である画像

解析技術や, 時系列情報である音声解析の技術の研究は実用レベルまで来ているが, そうした研究と比較して, 映像の解析システムに関する研究は少ない. 映像解析システム開発の妨げとなる大きな課題の一つとして, ヒトにとって, 映像がどのように認識されているか, どのような場面情報が映像理解に重要か, といったヒトの映像認知特性がはっきりしていないことが挙げられる. 映像から, 文脈的な内容を理解する認知過程を実証的に調べた研究報告は, 音声言語や画像の研究と比較してほとんど存在せず, 音声や画像に代わる, 映像を対象とした新しい実験手法の確立が課題であった. 筆者らは, この問題を解決するための独自の実験手法を開発し, 人の映像の認知特性の解明を目的として研究を行ってきた. 本論文では, 筆者らの認知科学における最新の研究成果を報告するとともに, それに基づいた映像の学習的な認知モデルを提案した. また, ヒトの認知モデルと同様の情報処理構造を映像解析システムに導入した学習型の映像情報処理アルゴリズムを提案する.

2. 映像の認知特性

映像がイベントごとに分節化されて認識されていることは, 日常の経験とよく一致する. 筆者らは, この仮説を元, どのような映像がどのように分節化されて知覚されるかを実証的に調べることのできる新しい実験パラダイムを開発し, 映像の認知特性を検討した.

2.1 方法

音声言語の研究において, Fodorら(1965)は, スピーチ

2011年1月26日受付, 2011年7月13日再受付, 2011年10月5日採録

[†] 東京大学/産業技術総合研究所/日本学術振興会特別研究員

(〒153-8904 目黒区駒場4-6-1東京大学先端科学技術研究センター13号館318号室, TEL 03-5452-5248)

^{††} 東京大学/科学技術振興機構

(〒153-8904 目黒区駒場4-6-1東京大学先端科学技術研究センター13号館318号室, TEL 03-5452-5248)

[†] The University of Tokyo/National Institute of Advanced Industrial Science and Technology/JSPS Research Fellow

(4-6-1, Komaba, Meguro-ku, Tokyo 153-8904, Japan)

^{††} The University of Tokyo/Japan Science and Technology Agency

(4-6-1, Komaba, Meguro-ku, Tokyo 153-8904, Japan)

図1 知覚における時系列情報の分節化
Segmentation of time series information on perception.

図2 認知実験に用いた刺激映像
Sequence of movie stimuli for the cognitive experiment.

のランダムなタイミングに短いクリック音を挿入し、被験者に音刺激が知覚されたタイミングを答えさせる実験を行った。その結果、スピーチの言語的な文節の前後で呈示した音刺激が、文節時点とずれているにもかかわらず、文節時点と同時であったように知覚されることを発見した(図1(左))。この研究は、音声言語が文法的・言語文脈的な分節(Chunk)に区切られて知覚されることを示した著名な研究である²⁾。筆者らは、この検査刺激を利用した音声言語の分節化の実験パラダイムを視覚実験に応用した。すなわち、映像が音声言語と同様に分節に区切られて知覚されているならば、映像の分節時点とずらして呈示した短いフラッシュ刺激が、映像の分節時点と同期して呈示されたように知覚されると思った(図1右)。実験では3種類のヒトの行動映像を用いた(図2)。まず、被験者が一般的に日常環境で目にしてしている行動の例として歩行を選んだ(平地歩行映像条件)。歩行は、実際には周期的な動作が繰り返されていると見ることができ一方で、筆者らは通常一定の歩行を見て、一つの連続的な状況として認識している。筆者らは開発した実験手法を用いて、実際にどのように分節化して認識されているかを調べた。また、歩行中の連続的な動作の中で、地面に不連続な段差があった場合について、映像の分節的な認知に対する影響を検討した(段差上り映像条件)。さらに、段差上り映像の逆再生映像を用意した。逆再生映像は、行動映像としては不自然な状況であるが、物理的な情報量としては、順再生と変わらず、呈示される画像の呈示順序が逆順になっているだけである。順再生と逆再生の映像でフラッシュの分節同期効果を比較することで、分節化が、より基礎的で物理的な情報処理に関係する

のか、より認知的な情報処理に関係するのかを確認した。理論的には、歩行の周期、段差に上る場面、逆再生の段差を降りる場面などで、映像が分節化されて知覚されると予測された。映像は、100HzのCRTモニタ上に40ミリ秒/フレーム(25Hz)で呈示した。一つの試行では、一回映像が再生された。映像再生中の16番から48番フレームの中の偶数フレーム(17カ所)の中のランダムな1カ所で、映像の上下に白いバー刺激を1フレーム(40ミリ秒間)呈示する。実験全体では、17カ所それぞれに6回ずつフラッシュが呈示される。したがって、一人の被験者につき全試行数は306試行である(3映像条件×17カ所×各6回)。実験には7名の被験者が参加した。

2.2 結果

実験の結果を図3に示した。横軸にはフラッシュの知覚されたタイミング、縦軸にはフラッシュが知覚されたと答えた頻度を示した。フラッシュはどの場面でも同じ回数(各6回)呈示されているため、すべて正確なタイミングで知覚されれば、頻度は、1.0の線に沿うはずであった。二元配置分散分析(映像の種類: $F[2, 12] = 0.44, p = 0.63$; タイミング: $F[16, 96] = 5.34, p < 0.01$; 交互作用 $F[32, 192] = 2.43, p < 0.01$) および多重比較検定(Tukey's HSD $p < 0.05$)の結果から、段差上り映像の順再生条件の38番フレームで、フラッシュの知覚頻度が有意に高く、その前後の知覚頻度が低いことがわかった。段差上り映像の38番フレームの画像を図3グラフ内に示した。

本結果は、段差に足を乗せた時点と時間的にずれたタイミングで呈示したフラッシュが、段差に足を乗せた場面と同時にあったように知覚されることを示した。つまり、段

図3 認知実験結果

図中の画像は段差上り映像における38番フレームを示す (Ohyama & Watanabe (2010) 図4より改変³⁾).

(a) The result of the cognitive experiment.

The picture in this figure shows the 38th frame in the gap-walking movie (Modified from Ohyama & Watanabe, 2010, fig.4³⁾).

図4 映像のオンライン分節化が知覚に及ぼす影響

The effect of on-line movie segmentation on perception.

差に足を乗せた場面において、フラッシュの分節同期効果が見られた。本研究結果から、連続的な映像情報が音声言語と同様に、分節化して知覚されていることが示唆された³⁾。一方、映像が逆再生の条件では、フラッシュの知覚頻度は、すべての場面で実際の呈示頻度と同程度であり、フラッシュが特定の場面に有意に偏って知覚される分節同期効果は見られなかった。逆再生映像の個々の場面の空間的特徴は、順再生映像と同じであった。したがって、順再生条件と逆再生条件での分節同期効果の違いは、前後の場面を含んだ時間的な文脈の中での場面特性によると考えられる。さらに、二つの条件で時間的な特性だけを比較しても、映像の変化量が同程度であることから、逆再生の映像における段差を降りる場面で、順再生の段差を上る場面と同様に分節化して知覚されると考えられる。しかし、逆再生条件では、どの時点でも分節同期効果が見られなかった。このように、物理的にはほとんど同じ条件である順再生と逆再生の映像の差異を明確に説明する要因を考えると、ヒトの経験や学習による内容の理解度の違いが影響している可能性が高い。つまり、分節化は基礎的な映像の変化量だけ

によるのではなく、経験的・学習的な文脈知識を用いて予測的に行われていると考えられる (図4)。本結果は、これまでの情報理論における分節化から予測できなかったヒトの分節化特性であり、映像の認知的な解析モデルを考える上で重要な性質と考えられる。こうした認知科学の実験結果は、映像の認知特性について重要な二つの性質を科学的に裏付けている。

- (1) 映像は、その文脈内容に応じて分節ごとに区切られて認識されている
- (2) 映像の分節化は、経験や学習によって可能となる

3. 認知科学に基づく映像解析モデル

筆者らが、明らかにした二つの映像の認知的特性に基づいて、知能的な映像解析システムには二つの機能を導入することが有効であると提案できる。

- (1) 映像を、文脈情報に応じて、可変長の短い映像に区切って符号化する。
- (2) 映像の分節は、過去の知識を利用し、学習的に決められる。

興味深いことに、現在、映像などの時系列解析において、優れた性能を示しているいくつかの情報処理モデルは、(1)と同様の機能を有している。例えば、HB-SOINN⁴⁾では、可変長に分節化された連続時系列情報をHMMによって符号化し、自己組織化追加学習によって知識データベース化する手法を提案している。こうした研究結果は、認知科学の実験に基づく筆者らの提案する映像解析システムの構造と整合する。一方、(2)の機能については、効果的なオンライン分節化手法⁵⁾が提案されているものの、未だ決定的なシステムが存在しない。これまでのシステムは、映像の物理的な変化量から分節を抽出することを目指してきたが、ヒトにおける映像の分節化を、ボトムアップの情報変化量だけから抽出することは難しいと考えられる。こうした課題に対し、筆者らの研究は、映像の分節化が、過去の知識を利用したトップダウンの処理に頼る部分が大きいことを示唆している。筆者らの実験結果は、映像の解析において、ある程度映像の文脈構造をデータベースとして記述し、それを利用して分節化を行うことが、効率的なだけでなく、認知科学の観点からも妥当であることを示している。これまで研究が進んできた時系列情報の符号化とデータベース化のモデルにしたがって構造化された学習済みのデータ構造を、新規映像のオンライン分節化に利用することで、より有効な分節化が学習的に獲得され、適応的な映像解析システムとなるだろう。

膨大な「映像」情報を取り扱う上で、映像そのものが持っている物理的な刺激変化を捉えるだけでは、人間が見て有用な情報を取出すのは難しい。筆者らが提案するヒトの認知特性に基づく情報処理モデルは、映像を人間が見た際に、どのような認知的処理を行うか、といった認知心理の結果を教師信号として映像解析処理に加えることで、人間にと

図5 提案する適応的な映像認知モデル
Proposed adaptive model of movie recognition.

って有用な映像解析処理が行えると考えられる。モデルの有効性は、筆者らの認知科学的実験結果によってサポートされる。

筆者らが提案する知能的な映像解析モデルを図5に示す。 n 番目のフレーム場面（もしくは時刻 n の時点の場面）を X_n とし、すべての場面の集合としての新しい入力映像を V とする。入力映像は、予め学習した映像データベースの構造を教師データとして、オンラインで予測的に分節化して符号化される（図5上）。つまり、過去の映像の場面の時系列部分集合 S を利用して、 V を分節化して符号化することで、映像の文脈内容を解析できる。初期分節構造は予め用意しておくが効果的だが、すべての条件を作り込む必要はない。例えば、防犯カメラ映像のデータベースなどでは、基本的に起こりうるイベントは数通りに決まると考えられる。これらの基本的なイベントの分節化と符号化の構造を、ヒトの認知特性に基づいて、予めデータベース化しておくことは、充分可能である。このようにして分節化された映像情報の構造は、初期データベース構造に反映する（図5下）。図5の場合、入力映像 V は、 $Sa1$, $Sa2$, $Sb3$ という分節の時系列として認識されたことから、新たに $Sa2$ から $Sb3$ が起きるといった文脈構造が学習され、データベース構造に反映される。この後に入力される映像の解析には、新しい情報を反映して変化したデータベース構造が用いられることで、知能的な映像情報処理を自動的に獲得していく適応型映像解析システムを構築できると考えられる。

4. む す び

本論文では、刻々と変化する膨大な視覚情報が、文脈に応じて分節化されて知覚されること、知覚の分節化が経験的・学習的に獲得されることを、認知科学的実験によって明らかにした。また、こうした研究成果を、先行研究報告と合わせて考察し、ヒト認知特性に基づく学習型映像解析モデルを提案した。本研究は、これまでの時系列解析手法の妥当性を実証的にサポートするだけでなく、現行のシステムにヒトの認知特性に基づく機能を追加することで、映

像情報からヒトが内容を理解する上で必要な情報を抽出する知能的な映像解析システムに应用可能であり、膨大な映像コンテンツの活用に貢献すると期待できる。今後は、映像の知覚における分節化の特性だけでなく、データベース構造にあたる映像内容の記憶特性と、それらの学習による変容について調べることで、提案する映像解析モデルをより進め、工学における情報論的解析モデルとの整合性を検討するとともに、効果的な映像解析モデルの提案につなげていきたい。

最後に、本論文の作成にあたり、情報理論的観点からご助言を下った、富山大学戸田英樹氏に感謝致します。

〔文 献〕

- 1) YouTube, <http://www.youtube.com/>
- 2) J.A. Fodor, T.G. Bever: "The psychological reality of linguistic segment", J. of Verbal Learning and Verbal Behavior, 4, pp.414-420 (Oct. 1965)
- 3) J. Ohya, K. Watanabe: "Contextual segmentation in the visual stream produces a temporal synchronization effect on visual perception", Psychologia, 53, 2, pp.125-137 (Mar. 2010)
- 4) S. Furao, O. Hasegawa: "An Incremental Network for On-line Unsupervised Classification and Topology Learning", Neural Networks, 19, 1, pp.90-106 (Jan. 2006)
- 5) J. Kohlmorgen, S. Lemm: "A dynamic hmm for on-line segmentation of sequential data", Advances in Neural Information Processing Systems, 14, pp.793-800 (Dec. 2001)

大山 潤爾 2006年、筑波大学大学院システム情報工学研究科修士取得退学。2008年、筑波大学大学院人間総合科学研究科修士取得退学。2009年～2011年、日本学術振興会特別研究員（東京大学大学院工学系研究科）。2011年、東京大学大学院工学系研究科博士課程修了。2011年より、産業技術総合研究所研究員、修士（工学）。修士（神経科学）。博士（学術）。正会員。

渡邊 克巳 2001年、カリフォルニア工科大学計算科学-神経システム専攻博士課程修了（Ph. D.）。National Eye Institute (NIH) 研究員、産業技術総合研究所研究員、科学技術振興機構ERATO下條潜在脳機能プロジェクト意思決定Gリーダー等を経て、現在、東京大学先端科学技術研究センター認知科学分野准教授。