様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月２０日現在

機関番号：１４４０１
研究種目：若手研究（B）
研究期間：2008～2010
課題番号：20700087
研究課題名（和文）　ラッシュ映像の編集による映像コンテンツ制作支援システムの構築

研究課題名（英文）　Development of Support System for Video Content Creation by Rush Video Editing

研究代表者
新田　直子（NITTA　NAOKO）
大阪大学・工学研究科・講師
研究者番号：00379132

研究成果の概要（和文）：
デジタルビデオカメラなどで撮影したラッシュ映像を素材とする映像コンテンツ制作に対し，専門家が制作した事例映像からの学習に基づき，ユーザに対する支援を行うプロトタイプシステムを構築した．８名の被験者による 10 点満点の主観評価を行った結果，ラッシュ映像の知覚品質を考慮した従来の支援により制作された映像コンテンツの 3.9 点に対し，事例映像を用いた支援により制作された映像コンテンツは 6.2 点に向上し，提案システムの有効性が示された．

研究成果の概要（英文）：
We developed a prototype system for supporting average users in video content creation from rush videos captured by digital video camcorders based on learning from professionally created video content examples. As a result of subjective evaluations in the scale of 0-10 by 8 subjects, while the video content created with the conventional support considering the perceived quality of rush videos received 3.9, the video content created with the example-based support received 6.2, indicating the effectiveness of the proposed system.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,400,000	420,000	1,820,000
2009 年度	900,000	270,000	1,170,000
2010 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：情報学／メディア情報学・データベース
キーワード：ラッシュ映像，映像編集，コンテンツ制作，パターン学習，事例

１．研究開始当初の背景

インターネットにおいて大量のデータ配信が可能となるのに伴い，配信されるコンテンツの多様化が見られるようになってきた．特に大きな動向として，これまで情報受信側であった一般ユーザが情報配信者として参加するようになり，一般ユーザが制作したコンテンツである消費者発信型メディア（CGM：Consumer Generated Media）が数多く見られるようになった．ウェブサイトなどの主にテキストで構成されるコンテンツの場合，一般ユーザにより高品質のコンテンツ制作が比較的容易であるが，映像コンテンツの場合，一般ユーザが制作したものは専門

家が制作したものに比べて質の差が大きいのが現状である．そこで本研究は，一般ユーザが，効率的に，より質の高い映像コンテンツ制作ができるよう支援するシステムの構築を目的とする．

従来の映像制作に関する研究としては主に自動映像要約が目的とされてきたが，専門家の編集により制作された放送型映像を素材映像としたものが多く，一般ユーザの映像コンテンツ制作において素材映像となることが多いと考えられる，一般ユーザがデジタルビデオカメラなどで撮影したラッシュ映像はほとんど対象とされていなかった．放送型映像はそのジャンル（スポーツ，ニュース，料理など）により，内容がある程度限定され，ショット（一台のカメラにより連続して撮影された部分映像）構成にも一定のパターンが存在するため，自動解析が比較的容易であるが，ラッシュ映像は内容もショット構成も不定であるため，従来手法をそのまま適用することはできない．本研究は，映像コンテンツ制作における各ステップにおいて，ラッシュ映像の特性を考慮しながら，信号処理技術により実現可能な部分の支援を目指す．

２．研究の目的

映像コンテンツ制作は一般に，(1)何を見せるか(What)と同時に，(2)どのように見せるか(How)を考慮しながら，既存の素材映像や音・トランジション効果（カット・フェードイン／アウトなど）などの要素メディアを選択し，組み合わせることにより実現される．本研究では素材映像となるラッシュ映像の特性を考慮し，上記(1)(2)の解決策としてそれぞれ以下の要素技術を持った映像コンテンツ制作支援システムの構築を目指す．

(1) 必要な部分映像の抽出

ラッシュ映像は一般に冗長な部分や不要な部分を多く含むため，映像の内容を理解するために必要な部分映像のみを自動的に抽出する．

(2)-① 要素メディアの適切な構成配置の提示

映像コンテンツにおける部分映像の時系列上の組み合わせや，各部分映像に対して挿入する音やトランジション効果など，異なるメディア間の組み合わせは無限に存在するため，制作される映像コンテンツの質は制作者の持つ技術に大きく依存する．そこで，実際に専門家が制作した事例映像において，どのような性質を持った要素メディアが，どのようなタイミング，組み合わせで用いられているかを統計的に学習し，学習したパターンに基づき，映像の構成方法を提示する．

(2)-② 特定フレーミング，カメラワークで撮影した映像の擬似生成

一般のユーザが撮影したラッシュ映像はフレーミング（クローズアップ，ロングショット等）やカメラワーク（ズーム，パン等）などの点で専門家が撮影した映像に比べ品質が劣る場合が多く，コンテンツ制作に有用な映像が存在しない可能性がある．そこで，ユーザが指定した特定のフレーミング，カメラワークで撮影した映像を擬似的に生成する．

３．研究の方法

２章に述べた各要素技術を以下の方法で実現する．ただし，(1)，(2)-①については，専門家が制作した映像コンテンツとその素材映像を事例とし，事例からの学習に基づく手法を提案する．

(1) 必要な部分映像の選択

放送型映像は一般に，一台のカメラが連続して撮影した部分映像であるショットの列として扱われ，放送型映像の自動編集においては，編集映像に使用するショットを選択した上で，さらにショット中から重要な部分映像を抽出することが多い．

一方ラッシュ映像は，一般ユーザが，カメラの電源を切ることなく，撮影対象を変えながら連続的に長時間撮影を続けることがあるため，放送型映像に比べて明確な構造を持たない．そこで，撮影対象が変わる際，パン・チルト・ズームなどのカメラ操作が見られることが多いことに着目し，まず映像の動き特徴から推定したカメラワークの種類により，同じ対象を撮影していると思われる部分映像（以下映像クリップ）に分割する．

一方，専門家が制作した編集映像中のショット（素材ショットから選択された部分映像）と素材ショットのセットから隠れマルコフモデル(HMM: Hidden Markov Model)を用いて学習した，選択される部分映像及び選択されない部分映像の動画・音特徴の変化パターンに基づき，編集映像に使用する各映像クリップから重要な部分映像を選択する．

(2)-① 要素メディアの適切な構成配置の提示

映像・音・トランジション効果などの複数要素メディアを組み合わせ，視聴者の興味を引くことを目的に専門家が制作した映像コンテンツを事例とする．映像コンテンツを放送型映像と同様にショットの列とみなし，まず以下の(ア)(イ)のように，ショットの時系列上の並びである時間構造，ショットに対して挿入する音やトランジション効果など異なるメディア間の組み合わせである空間構造に着目し，要素メディアの適切な構成配置を学習する．ただし，以下の処理の前処理として，事例映像を構成するフレーム，ショット，シーンなど各構造に応じたセグメントから，輝度，動き，カメラワークや人物の有無などに関連した動画特徴や，音量，周波数，テンポなどの音特徴を抽出する．

(ア) 各メディアの時間構造

まず，類似した特徴を持つショットを同等に扱うため，抽出した特徴ベクトル列を，ベクトル量子化などにより類似度に基づいたシンボル列に変換する．ここで一般に，複数の連続したショットはシーンを構成し，複数のシーン列により映像コンテンツは構成される．映像コンテンツ及び各シーンを構成するショットの並びには，映像文法に代表されるように，あるパターンが存在すると考えられるため，事例となる映像コンテンツにおけるシンボル列の時間軸上の並び方を，映像のシーン構成，シーンのショット構成として，HMM により学習する．

(イ) メディア間の空間構造

音はある映像セグメントに，トランジション効果は連続した 2 つの映像セグメント間に付与される．そこで，以下の 2 種類の空間構造について検討する．

(I)セグメント間の関係：ここでは付与する音を音楽に限定する．専門家の制作した映像コンテンツにおいて，一つの音楽クリップは映像中のシーンに挿入されることが多い．また，各シーンには印象が適合した音楽クリップが選択されると考えられる．そこで，事例となる映像コンテンツ中のシーンと付与された音楽クリップの組み合わせから，シーンと音楽クリップの相関関係を学習する．具体的には，動画特徴空間と相関をもつよう音楽特徴空間を変換する非線形写像をニューラルネットワークに基づいた学習モデルにより推定する．この結果，動画特徴空間において距離の近いシーンに印象が適合する音楽クリップは，変換後の音楽特徴空間において距離が近くなると考えられる．

(II) セグメント境界間の関係：トランジション効果は，連続したショット間に挿入され，ショット境界付近における音・動画特徴の変化に応じたトランジション効果が選択されると考えられる．そこで，各トランジション効果（カット，フェードなど）が付与されたショット列に対し，ショット境界付近のフレーム列における画像・音特徴変化の変化パターンを，マルコフ連鎖モデル(MCM)により空間構造として学習する．

最後に，以上の学習結果に基づき映像コンテンツ制作の支援を行う．ユーザはインタフェース上で，映像クリップのストーリーボードへの移動や音楽の選択により，映像コンテンツを制作するものとし，インタフェース上で支援を行う．具体的には，(ア)で学習した映像のシーン構成，シーンのショット構成をテンプレートとしてインタフェース上に提示し，ユーザがテンプレート上で選択したショットに応じて適切な映像クリップ候補を提示する．映像クリップの適切さは，学習に用いた動画特徴から得られる各ショットへの適合度，カメラワークや照明条件に関連する知覚品質の 2 つから算出される．また，選択したショットで構成されるシーンに対して，(イ)(I)で学習したシーンと音楽クリップの相関関係に従い，適切な音楽クリップ候補を提示する．トランジション効果については，(イ)(II)の学習結果により，各ショット境界に対して適切なトランジション効果が一意に決まるため，自動的に付与されるものとする．

(2)-② 特定フレーミング，カメラワークで撮影した映像の擬似生成

(2)-①において得られたショットに適合する特徴を持つ映像クリップが存在しない場合，特定フレーミング，カメラワークを持つ映像に変換し，異なる特徴を持つ映像クリップを疑似的に生成する．前提として，ラッシュ映像においてクローズアップで撮影された映像区間などにおいて，撮影されていない背景（欠落領域）の大部分が他の区間において撮影されているものとする．このような映像区間に対し，まず，構成するフレーム列を用いた画像モザイキングによる欠落領域の補完により，撮影環境全体が撮影された画像

を生成する． 生成された画像に各フレームを張り付けることにより，撮影環境全体が撮影されたロングショット映像が生成できる．生成された動画に対し，カメラワークの切り替え点及びその時点のフレーミング（切り出し矩形）を指定することにより，指定したフレーミング，カメラワークを持つ映像が疑似的に生成される．

４．研究成果

３章で述べた各研究課題について，成果を示す．事例映像としては，映像・音楽・トランジション効果を用いて専門家が制作した映像コンテンツである映画予告映像と，その素材映像である映画をジャンルをアクションに限定し，61 本収集した．これを手作業によりショット・シーンに分割した後，パターン学習を行った．ただし，シーン境界は音楽の変化点とした．素材となる映像は一般ユーザがデジタルカメラで撮影した映像，素材となる音楽クリップは，映画のサウンドトラック，クラシック，ポピュラー音楽，洋楽，邦楽などさまざまなジャンルの楽曲を 123 曲収集し，曲調の変化点で人手により分割した合計 180 個の音楽クリップを用いた．

(1)-① 必要な部分映像の選択

まず，12 本の映画予告映像に用いられた素材ショット 69 個に提案手法を適用し，72.5%(50 個）の素材ショットに対し，実際に使用された部分映像との誤差が 5 フレーム以内の部分映像を抽出できることを確認した．また，抽出された部分映像の適切さについて主観評価した結果，80%がランダムに選択した部分映像より評価が高く，53.9%については実際に使用された部分映像と同等またはそれ以上の評価を得た．次に，ラッシュ映像から得た 5 個の映像クリップと，テンポや曲の雰囲気が異なる 6 曲の音楽から選択した 1 曲を入力とし，音楽に合わせた長さで選択された部分映像を繋ぎ合わせ制作したシーンに対し，10 人の被験者に 5 段階評価させた結果を表 1 に示す．

平均で 3.6 点と，先頭から自動的に選択した部分映像（2.8 点）よりも高評価を得た．

(2)-① 要素メディアの適切な構成配置の提示

表 1：選択された部分映像から構成されるシーンの主観

	音楽の特徴			平均得点	
	強いビート	テンポ	雰囲気	生成映像	比較映像
I	少ない	遅い	暗い	3.4	3.2
II	普通	普通	普通	3.6	2.7
III	普通	早い	明るい	3.5	2.6
IV	普通	早い	普通	4.1	2.8
V	普通	遅い	暗い	3.5	1.9
VI	多い	早い	明るい	3.4	3.4
全体				3.6	2.8

表 2：選択された音楽の主観評価

	Video1	Video2	Video3	Video4	Video5
平均	6.1	4.4	7.2	5.0	4.5

表 3：トランジション効果付与結果

		決定結果			
		カット	フェード・ディゾルヴ	その他	合計
正解	カット	84	4	0	88
	フェード・ディゾルヴ	6	6	0	12
	その他	1	0	0	1
	合計	91	10	0	101

まず，**(イ)(I)**の学習方法の評価を行うため，10 個の入力シーンに対し，提案手法により選択した音楽クリップ，提案手法により印象が適合しないと判断された音楽クリップ，実際に付与されていた音楽クリップ，音楽空間を学習した変換行列により変換せずに印象が適合すると判断された音楽クリップ，同様に音楽空間を変換せずに印象が適合しないと判断された音楽クリップをそれぞれ付与したものを Video 1～5 とし，10 名の被験者に 10 段階評価させた平均点を表 2 に示す．Video1 が Video3 に次いで高い評価を得ており，Video 1 が Video 4 より高く，Video 1 と Video 2 の点差が Video 4 と Video 5 の点差より大きいことから，学習された非線形写像，及び特徴空間中の距離に基づく音楽選択手法の有効性が示された．

また，10 本の映画予告映像からランダムに選択した 101 組の連続した 2 ショット間に対し，**(イ)(II)**の学習結果により，トランジション効果を決定した．実際に付与されていたトランジション効果を正解とした結果を表 3 に示す．89.1%（90/101）のテストサンプルに対し，正しくトランジション効果を選択できたが，事例映像自体にカットが極端に多いため，今後，カット以外のトランジション効

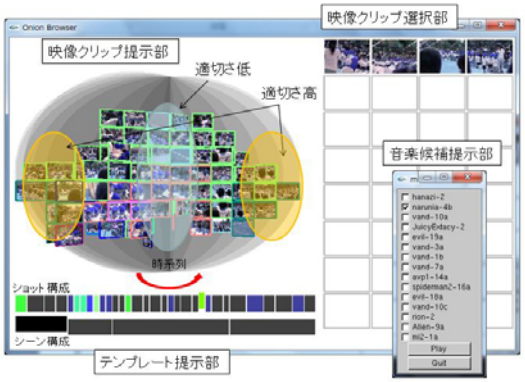

図 1：ユーザインタフェース

果を含む事例を増やして評価する必要がある．

図 1 に構築したユーザインタフェースを示す．まず，テンプレート提示部に，(ア)で学習した映像のシーン構成，シーンのショット構成が提示される．上がショット構成，下がシーン構成で，各矩形が一つのショット，シーンを表す．また，映像の内容把握に重要と考えられる素材映像の時系列情報を考慮した上で，コンパクトな表示を行うため，3 次元フォトブックを作成し，映像クリップ提示部に提示する．ユーザがテンプレート上のショットを選択すると，選択されたショットに対する適合度及びカメラの動き，照明条件などに関連した知覚品質に基づき，映像クリップが各ページに提示される．ここでは特に，半楕円形の平面をフォトブックの 1 ページとし，ユーザから見えにくい縦軸側に適切でない映像クリップ，横軸の最も円周側に適切な映像クリップが配置されるよう，配置場所を決定した．これにより，ユーザは，最も目に付きやすい位置である各ページの最も円周側から映像クリップを選択すればよいことになる．さらに，各シーンを構成するショットに対し，映像クリップをすべて選択した後，テンプレート上で該当シーンを選択すると，(イ)(I)で学習した相関関係に基づき，適切な音楽候補が提示される．最後に，候補からユーザが選択した音楽がシーンに挿入される．

幼稚園の運動会をデジタルビデオカメラで撮影した 45 分のラッシュ映像をカメラワークに基づき分割した 265 個の映像クリップを素材映像とし，構築したプロトタイプシステムを用いて制作した映像コンテンツの品質に対し，8 名の被験者による 10 点満点の主観評価を行った．従来手法で用いられるカメラワークや照明条件に関連する素材映像の知覚品質のみを考慮した支援により制作した映像コンテンツの評価点が 3.9 点であったのに対し，事例映像から学習したテンプレートに対する適合度も考慮することにより 6.2 点に向上し，事例を用いた学習に基づく映像コンテンツ制作支援システムの有効性が示された．

図 2：入力映像

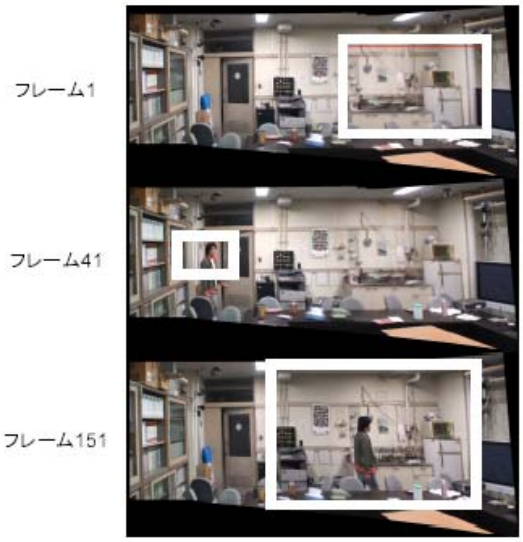

図 3：ロングショット映像と
フレーミングの指定

図 4：生成映像

(2)-② 特定フレーミング，カメラワークで撮影した映像の擬似生成

カメラのパンにより撮影されたラッシュ映像に対し，提案手法により指定したフレーミング，カメラワークで撮影した映像の擬似生成を行った．図 2 に入力映像を示す．画像モザイキングの利用により，図 3 に示すような撮影環境全体が撮影されたロングショット映像が生成され，図に示す 3 フレームにおいてフレーミングを指定した結果，図 4 に示す様に入力映像の冒頭部では撮影範囲外であった部屋の右側から人へとズームインする

映像が生成された．また，被験者 8 人によるカメラワークの自由度に関する主観評価において 5 段階で平均 4.2 の評価が得られた．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）
① N. Nitta and N. Babaguchi, “Example-based Video Remixing,” Multimedia Tools and Applications, vol.51, no.2, pp.649-673, 2011 (査読有).

〔学会発表〕（計 10 件）
① N. Nitta and N. Babaguchi, “Example-based Video Remixing for Home Videos,” International Conference on Multimedia & Expo (ICME2011), Barcelona, Spain, July 12-14, 2011 (accepted).
② 組橋祐亮，新田直子，馬場口登，“スライドショー生成のための事例に基づく画像選択”，電子情報通信学会 2011 年総合大会，D-12-90, pp.193, 東京都市大学，March 17, 2011.
③ 金壮一，新田直子，馬場口登，“事例に基づく映像ショット列に対する音楽ミキシング”，電子情報通信学会技術研究報告，PRMU2010-294，pp.335-340，つくば市，March 11, 2011.
④ 吉田好，新田直子，馬場口登，“事例映像への適合度と知覚品質に基づくホームビデオ編集支援”，電子情報通信学会技術研究報告，PRMU2009-292, pp.347-352, 鹿児島大学，March 16, 2010.
⑤ 金壮一，新田直子，馬場口登，“事例に基づく映像ショット列への音楽付与”，電子情報通信学会技術研究報告, PRMU2009-180, pp.162-172，京都大学，January 21, 2010.
⑥ 金壮一，新田直子，馬場口登，“事例映像に基づくシーンに対する適応的音楽選択”，第 8 回情報科学技術フォーラム(FIT2009), RK-007，pp.93-96，東北工業大学，September 2-3, 2009.
⑦ 栗原陽介，新田直子，馬場口登，“事例映像からの学習に基づくマルチメディア協調型映像編集”，第 5 回デジタルコンテンツシンポジウム，pp.1-5, 幕張メッセ，June 10, 2009.
⑧K. Asano, N. Nitta, and N. Babaguchi, “Virtual Camerawork beyond Original Framing with Longshot Video Generation,” International Workshop on Computer Vision and Its Application to Image Media Processing (WCVIM2009), pp.86-90, Tokyo, January 13, 2009.
⑨K. Kurihara, N. Nitta, and N. Babaguchi, “Automatic Appropriate Segment Extraction from Shots Based on Learning from Example Videos,” Pacific-Rim Symposium on Image and Video Technology (PSIVT2009), pp.1082-1093, Tokyo, January 16, 2009.
⑩ 浅野宏一，新田直子，馬場口登，“ロングショット映像生成による仮想カメラワークの実現”，画像の認識・理解シンポジウム (MIRU2008), pp.1193-1198, 軽井沢, July 31, 2008.

６．研究組織
(1)研究代表者
新田　直子（NITTA　NAOKO）
大阪大学・工学研究科・講師
研究者番号：00379132