プロットとショットの対応付けによる映画要約*

李　雪山† 宇津呂 武仁† 上原 宏†‡
筑波大学大学院 システム情報工学研究科† NTT ドコモ 法人営業部‡

1 序論

映画は重要な娯楽文化の一つであり，毎年多数の映画が制作されている．一方，映画を視聴する利用者の側に立つと，膨大な映画作品の中から，自分の興味に合った作品を選択する必要に迫られているのが現状であると言える．そのため，映画の予告編等の要約映像をふまえた上で，視聴する作品を選ぶ作業の必要性は年々高まっているのが現状である．しかし，通常，予告編映像は，激しい場面の組み合わせで構成される場合が多く，映画のストーリーを把握する目的においては，有益とは言い難い．

これらの状況をふまえて，本論文では，図 1 に示す映画要約の支援方式の流れに沿って，Wikipedia 中のプロット中の重要文に対応するショットを抽出し，映画要約結果として出力する方式を提案する．本論文の映画要約方式において想定する利用者像としては，主として以下が挙げられる．

1. すでに一度視聴済の映画に対して，鑑賞レポートを書く必要がある場合．
2. 自分が一度視聴した映画を他人に推薦する場合や映画の宣伝をする場合．
3. これから自分が新たに視聴する映画を選ぶ場合．

本論文の映画要約方式においては，まず，Wikipedia 中のプロットから重要文を抽出する [2]．次に，映画映像をショット列に分割するツールを適用し，数百個のショットに分割する [1]．そして，分割されたショット列において，プロット中の重要文に対応するショットを選定しこれを抽出する．この選定過程においては，時間情報付き字幕 (サブタイトル) およびシーン描写 (スクリプト) を利用してプロットとショットの人物を対応付けること，および，字幕およびプロット中の語の重複を利用してプロットとショットを対応付けることを行う．最後に，抽出されたショットに対応する映画映像を結合することにより，要約映像を作成する．本論文においては，要約映像作成結果を以下の二通りの方式で評価する．

1. プロット文に対応するショットの数が制限しない場合
2. プロット一文に対応するショットの数の上限が 3 の場合

2 プロット中の重要文とショットの自動対応付け

本論文においては，プロット中の重要文とショットを自動的に対応付ける手法として，文献 [5] における方式を用いる．文献 [5] においては，以下に述べる手法によって，Wikipedia 中の映画プロット中の各文に対して，映像中の人物や時間情報付き字幕 (サブタイトル) を介して映像中のショットに対応付ける方式を提案している．

文献 [5] の方式においては，次式によってプロット中の文 s_i とショット t_j の間の類似度 $f_{fus}(s_i, t_j)$ を定義し，この類似度を用いた動的計画法によって，プロット中の文とショットの対応付けを行う．

$$f_{fus}(s_i, t_j) \quad = \quad f_{id}(s_i, t_j) + \alpha \cdot f_{subtt}(s_i, t_j)$$

ここで，$f_{id}(s_i, t_j)$，および，$f_{subtt}(s_i, t_j)$ は，それぞれ，プロット中の文 s_i に含まれる人物名とショット t_j 中に出現する人物の対応付けを利用した類似度 (2.1 節)，および，各ショットを時間情報付き字幕 (サブタイトル) に対応付けた後，字幕およびプロット中の語の重複を集計することによってプロットとショットの対応を測定する類似度 (2.2 節) であり，α は重みパラメータである．

2.1 人物を介した対応付け

プロット中の文 s_i に含まれる人物名，および，ショット t_j 中に出現する人物の対応付けを利用した類似度 $f_{id}(s_i, t_j)$ を算出する際には，ショット中に出現する人

*Movie Summarization by Aligning Plot and Shots
†Xueshan Li, Takehito Utsuro, Hiroshi Uehara, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Hiroshi Uehara, NTT DOCOMO, INC., Corporate Sales and Marketing Division

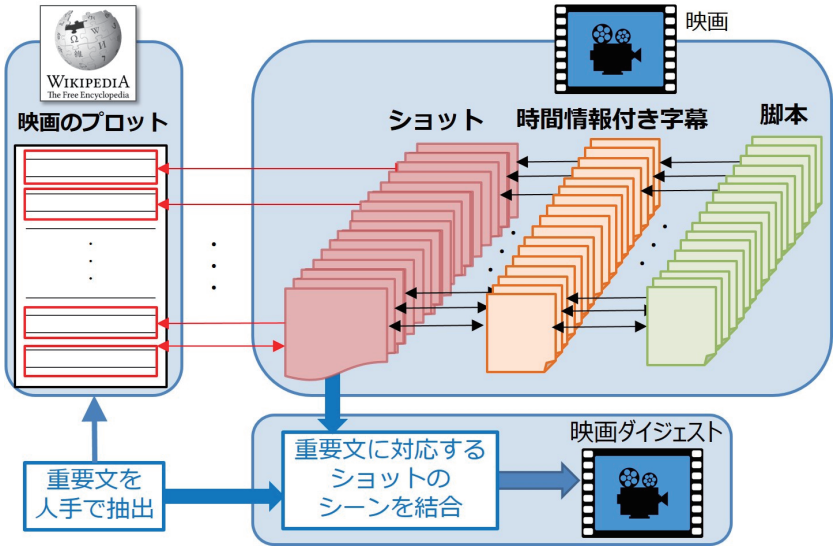

図 1: 映画要約の支援方式の流れ

表 1: 映画要約作業において用いた映画例

映画名	プロット中の文の数	重要文の数	自動分割後のショットの数
ローマの休日	53	11	649
ふしぎの国のアリス	47	10	1,581

物 $c(\in C$，ただし，C は映画中に出現する全人物の集合) とプロット中に含まれる人物名 $d(\in D$，ただし，D はプロット中に含まれる全人物名の集合) の間の対応付けを判定する次式の関数を学習してこれを用いる．

$$\mathrm{align}(c,d) = \begin{cases} 1 & (\text{人物 } c \text{ と人物名 } d \\ & \text{が対応する場合}) \\ 0 & (\text{その他の場合}) \end{cases}$$

この関数の学習は，時間情報付き字幕 (サブタイトル) と人物名を用いたシーン描写 (スクリプト) を対応付けることによって行う．

この関数 $\mathrm{align}(c,d)$ を用いることによって，次式によって，プロット中の文 s_i に含まれる人物名，および，ショット t_j 中に出現する人物の対応付けを利用した類似度 $f_{id}(s_i,t_j)$ を算出する．

$$f_{id}(s_i,t_j) = \sum_{k=j-r}^{j+r} \sum_{c\in C_j} \sum_{d\in D_i} \mathrm{align}(c,d)\cdot I(c)$$

上式においては，ショット t_j に隣接する前後 r ショットずつを含めたショット群中に出現する人物の集合を C_j，プロット中の文 s_i に含まれる全人物名の集合を D_i として，関数 $\mathrm{align}(c,d)$ によって対応する人物 $c(\in C_j)$ と人物名 $d(\in D_i)$ の組数を類似度 $f_{id}(s_i,t_j)$ とする．ただし，各人物 c に対しては，人物 c が出現するショット数を $n_{FT}(c)$ として，逆文書頻度 IDF(inverse document frequency) に相当する次式の重みを付与する．

$$I(c^*) = \frac{\log\left(\max_{c\in C} n_{FT}(c)\right)}{\log\left(n_{FT}(c^*)+1\right)}$$

2.2 字幕およびプロット中の語を介した対応付け

各ショットを時間情報付き字幕 (サブタイトル) に対応付けた後，字幕およびプロット中の語の重複を集計することによってプロットとショットの対応を測定する類似度 $f_{subtt}(s_i,t_j)$ を算出する際には，まず，時間情報付き字幕 (サブタイトル) の時間情報を用いることにより，各ショット t_j に対してサブタイトル $subtt$ の集合を対応付ける．そして，次式によって，プロット中の文 s_i に含まれる語 v とショット t_j に対応付けられたサブタイトル $subtt$ 中の語 w の間の重複を集計し，これを類似度 $f_{subtt}(s_i,t_j)$ として用いる．ただし，次式において，関数 (v,w) は，語 v と語 w が同一の場

合のみ 1 を返す関数として定義される．

$$f_{subtt}(s_i, t_j) = \sum_{v \in s_i} \sum_{w \in subtt \in t_j} \text{word-match}(v, w)$$

$$\text{word-match}(v, w) = \begin{cases} 1 & (v = w) \\ 0 & (v \neq w) \end{cases}$$

2.3 評価手順

「ローマの休日」と「ふしぎの国のアリス」を対象とする評価を行った．映画例に関するプロットとショットの数は表 1 に示す．ただし，文献 [5] の方式の性能の上限を見積もるために，プロット文中の代名詞について，人手でその照応先の人名に書き換えた後，評価を行なった．

2.3.1 すべてのショットが必ず一つのプロット文に対応

文献 [5] の手法においては，プロットの文数に対して，候補となるショット数が数十倍程度の数であるにも関わらず，全てのショットをプロット中のいずれかの文に対応付けるという制約が課せられている．文献 [5] においては，この制約による弊害を最小限に抑えるために，プロット中の一つの文に対して対応可能なショットの数に上限を設ける方式が提案されている．本論文でも，この方式に従い，プロット中の一つの文に対して対応可能なショットの数に上限を設ける．文献 [5] の方式では，全ショット数を N_T，全プロット文数を N_S として，プロット文一文に対応するショットの平均数 N_T/N_S とパラメータ k の積を求め，これをプロット文一文に対応するショット数の上限 z とする．

$$z = k \cdot N_T / N_S$$

本論文では,「ローマの休日」の場合の設定である $N_T = 649$ および $N_S = 53$ の場合に要約性能が最適となるパラメータ k の値として $k = 5$ を用い，この場合のショット数の上限 $z = 61$ を用いた.「ふしぎの国のアリス」の場合も，パラメータ k の値としては $k = 5$ をそのまま用い，$N_T = 1,581$ および $N_S = 47$ の場合のショット数の上限 $z = 168$ を用いた．

2.3.2 プロット文一文に対して最大 3 ショットのみが対応

全てのショットをプロット中のいずれかの文に対応付けるのではなく，プロットの中の各文に対して，少数のプロットを厳選して対応付ける方式を導入する．

具体的には，システム出力のプロット一文 s_i に対応するショット数が 3 以下の場合にはそのままとし，ショット数が 3 を超える場合には，前節の手順の結果においてプロット文 s_i に対応するショットの数を N_i として，次式によりその中の連続する 3 つのショットの f_{fus} を足し合わせる．そして，この和が最大となるショット列を対応結果 $T_3(s_i)$ とする．

$$T_3(s_i) = \underset{l=0 \sim N_i - 2}{\text{argmax}} \left(\sum_{l=j}^{j+2} f_{fus}(s_i, t_l) \right)$$

2.4 自動評価結果

映画要約結果の自動評価を行うために，まず，重要文に対応するショット最大 3 つを人手で選択し，それを参照用ショット列とする．映画要約結果の自動評価においては，各プロット文に対して，自動要約結果のショット列中のいずれかのショットの位置と，参照用ショット列中のいずれかのショットの位置の間の差が 1 ショット以内の場合に，当該プロット文に対して対応付けられたショット列は正解であると判定する.「ローマの休日」および「ふしぎの国のアリス」を対象とした評価結果を表 2，表 3，および，以下に示す．

1. すべてのショットが必ず一つのプロット文に対応する場合

   「ローマの休日」: $6/11 = 54.5\%$

   「ふしぎの国のアリス」: $8/10 = 80.0\%$

2. プロット一文に対して最大 3 ショットのみが対応する場合

   「ローマの休日」: $2/11 = 18.2\%$

   「ふしぎの国のアリス」: $5/10 = 50.0\%$

2.5 主観評価結果

次に，映画要約結果に対する主観評価として，各プロット文に対して対応付けられたショット列の映像を視聴した場合に，各プロット文の情報が得られるか否かを判定基準として，自動要約結果の主観評価を行なった結果を以下に示す．

1. すべてのショットが必ず一つのプロット文に対応する場合

   「ローマの休日」: $8/11 = 72.7\%$

   「ふしぎの国のアリス」: $8/10 = 80.0\%$

2. プロット一文に対して最大 3 ショットのみが対応する場合

   「ローマの休日」: $6/11 = 54.5\%$

   「ふしぎの国のアリス」: $7/10 = 70.0\%$

表 2: 要約結果の統計情報及び評価結果 (すべてのショットが必ず一つのプロット文に対応)

映画名	要約前の映画の時間長	要約結果の動画の時間長	要約率	精度
ローマの休日	1 時間 58 分 11 秒	18 分 53 秒	15.98%	54.5%
ふしぎの国のアリス	1 時間 15 分 10 秒	14 分 43 秒	19.58%	80.0%

表 3: 要約結果の統計情報及び評価結果 (プロット一文に対して最大 3 ショットのみが対応)

映画名	要約前の映画の時間長	要約結果の動画の時間長	要約率	精度
ローマの休日	1 時間 58 分 11 秒	3 分 3 秒	2.58%	18.2%
ふしぎの国のアリス	1 時間 15 分 10 秒	1 分 23 秒	1.84%	50.0%

この評価結果に示すように，ほとんどの場合において，自動評価結果と比較して改善がみられた．この結果から，提案手法によって，一定の情報を含む映画要約が可能であることが分かった．

3 関連研究

映画の字幕情報と映像情報の対応付けを行い，映画のシーン分割を行う手法の一つとして，文献 [3] では，字幕中の人名と顔画像の対応付けを行う手法を提案している．また，文献 [7] においては，字幕を文書ベクトル化して意味的まとまり区間を抽出することにより映画要約を行う手法を提案している．さらに，文献 [6] では，映画・ドラマにおいて，時間情報付きのサブタイトルとスクリプトを対応付けるとともに，映像のシーン分割を行い，これらの情報を統合した上で重要シーンのランキングを行い，映画・ドラマを要約する手法を提案している．また，文献 [5] においては，Wikipedia 中の映画プロット中の各文に対して，映像中の人物や時間情報付きサブタイトルを介して映像中のショットに対応付けることにより，映像検索を行う手法を提案している．

4 結論

本論文では，Wikipedia 中に掲載されている映画のプロット情報を手がかりとして，プロットを映画の映像から生成したショットに対応付けることにより，映画要約過程を支援する手法を提案した．今後の課題としては，大規模な評価実験を行うことが挙げられる．また，文献 [4] の映像中のシーン分割方式を適用し，プロットとショットを高精度に対応付けることが挙げられる．

参考文献

[1] E. Apostolidis and V. Mezaris. Fast shot segmentation combining global and local visual descriptors1. In *Proc. ICASSP*, pp. 6583–6587, 2014.

[2] 李雪山, 堂元健太郎, 宇津呂武仁. プロット中の重要文とショットの対応付けによる映画要約支援方式. 第 8 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム—論文集, 2016.

[3] C. Liang, Y. Zhang, J. Cheng, C. Xu, and H. Lu. A novel role-based movie scene segmentation method. In *Advances in Multimedia Information Processing — PCM2009*, Vol. 5879 of *LNCS*, pp. 917–922. Springer, 2009.

[4] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal video segmentation to scenes using high-level audio-visual features. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 21, No. 8, pp. 1163–1177, 2011.

[5] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Aligning plot synopses to videos for story-based retrieval. *International Journal of Multimedia Information Retrieval*, Vol. 4, No. 1, pp. 3–16, 2015.

[6] T. Tsoneva, M. Barbieri, and H. Weda. Automated summarization of narrative video on a semantic level. In *Proc. Semantic Computing*, pp. 169–176, 2007.

[7] H. Yi, D. Rajan, and L.-T. Chia. Semantic video indexing and summarization using subtitles. In *Advances in Multimedia Information Processing — PCM2004*, Vol. 3331 of *LNCS*, pp. 634–641. Springer, 2004.