2F5-1in1

プロット中の重要文を情報源とする映画の要約支援方式

A Method of Assisting Movie Summarization based on Key Sentences of the Plot

李 雪山[*1] 宇津呂 武仁[*2]
Xueshan Li Takehito Utsuro

[*1]筑波大学大学院システム情報工学研究科
Grad. Sch. Sys. & Inf. Eng, Univ. of Tsukuba

[*3]筑波大学システム情報系
Fclty. Eng, Inf. & Sys, Univ. of Tsukuba

This paper proposes a method of assisting movie summarization using plot information. A plot of a movie available at Wikipedia contains a major story of the movie. From such a plot of a movie, we extract several important sentences as the content of summary. For summarizing movie, the key work is finding the best alignment between sentences of plot and shots which are segmented from a movie. There are two cues used to measure the similarity between a sentence and a shot. One is based on character appearing in both a sentence and a shot, another is based on words matching. Then, an alignment method based on dynamic programming is applied to optimize the alignment. Finally, an experimental evaluation on movie *Roman Holiday* shows the effectiveness of this method.

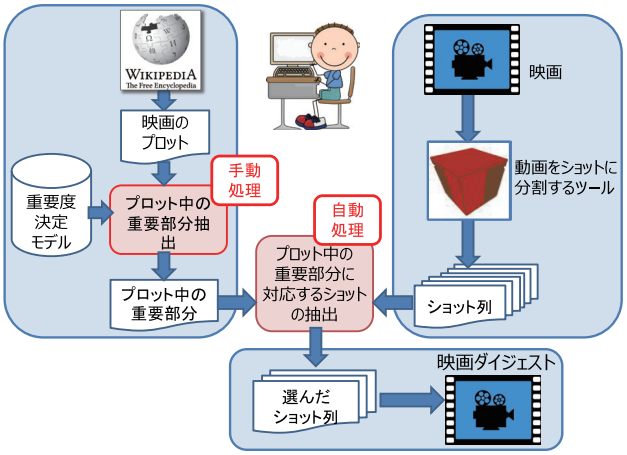


図 1: 映画要約の支援方式の流れ

1. はじめに

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．

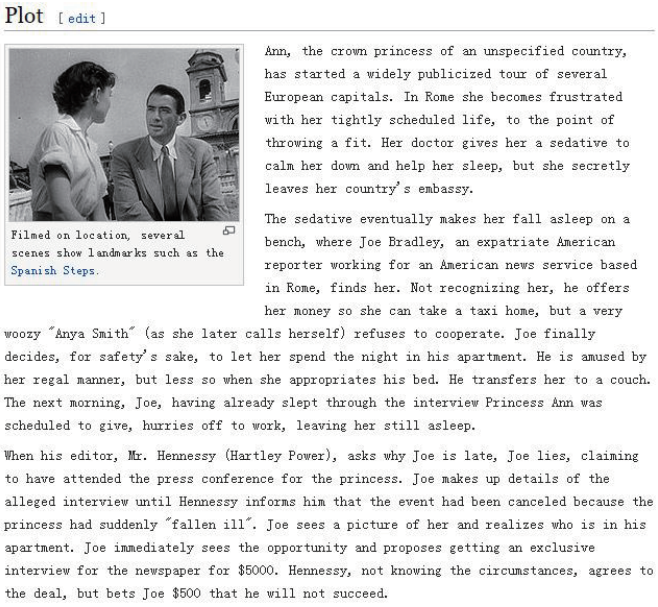
1. すでに一度視聴済の映画に対して，鑑賞レポートを書く必要がある場合．
2. 自分が一度視聴した映画を他人に推薦する場合や映画の宣伝をする場合．
3. これから自分が新たに視聴する映画を選ぶ場合．

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．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．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．

2. 映画要約の支援方式

本論文で提案する映画要約の支援方式の流れを図 1 に示す．本方式においては，まず，Wikipedia 中における当該映画作品のエントリの記事本文から，映画の物語のあらすじ (プロット (plot)) を収集する．次に，あらかじめ定義された文の重要度の基準に基づいて，プロット中の重要文を人手で選択する．一方，映画の動画に対しては，映像途中のカメラ切り替わり個所において動画をショット列に分割するツール [Apostolidis 14] を適用することによって，通常の 2 時間程度の時間長の映画を数百個のショットへと分割する．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，選択されたショット列に対応する断片的動画を結合することにより，映画の要約映像を生成する．

連絡先: 李 雪山，筑波大学大学院システム情報工学研究科，
〒 305-8573 茨城県つくば市天王台 1-1-1, 029-853-5427



Plot [edit]

Filmed on location, several scenes show landmarks such as the Spanish Steps.

Ann, the crown princess of an unspecified country, has started a widely publicized tour of several European capitals. In Rome she becomes frustrated with her tightly scheduled life, to the point of throwing a fit. Her doctor gives her a sedative to calm her down and help her sleep, but she secretly leaves her country's embassy.

The sedative eventually makes her fall asleep on a bench, where Joe Bradley, an expatriate American reporter working for an American news service based in Rome, finds her. Not recognizing her, he offers her money so she can take a taxi home, but a very woozy "Anya Smith" (as she later calls herself) refuses to cooperate. Joe finally decides, for safety's sake, to let her spend the night in his apartment. He is amused by her regal manner, but less so when she appropriates his bed. He transfers her to a couch. The next morning, Joe, having already slept through the interview Princess Ann was scheduled to give, hurries off to work, leaving her still asleep.

When his editor, Mr. Hennessy (Hartley Power), asks why Joe is late, Joe lies, claiming to have attended the press conference for the princess. Joe makes up details of the alleged interview until Hennessy informs him that the event had been canceled because the princess had suddenly "fallen ill". Joe sees a picture of her and realizes who is in his apartment. Joe immediately sees the opportunity and proposes getting an exclusive interview for the newspaper for $5000. Hennessy, not knowing the circumstances, agrees to the deal, but bets Joe $500 that he will not succeed.

図 2: Wikipedia に掲載されている「ローマの休日」のプロットの一部

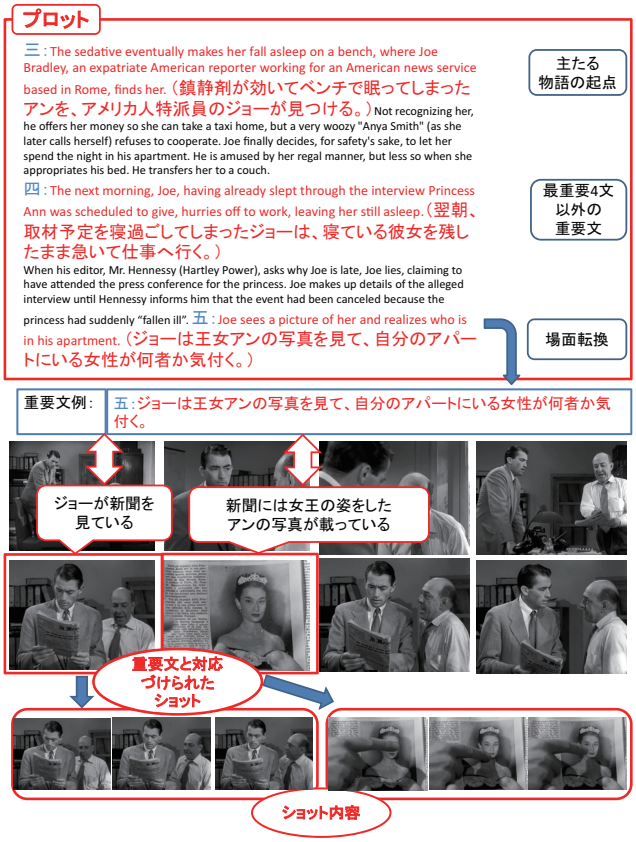


図 3: プロット中の重要文をショットに対応付ける過程 (「ローマの休日」の例)

3. Wikipedia 中の映画のプロット

Wikipedia に掲載されている映画のプロットは，映画のストーリーに沿って重要な内容の概要を記したもので，通常，数百 ～ 千数百語程度の長さで記述される[*1]．一例として,「ローマの休日」のプロットの一部を図 2 に示す.

表 1: 映画例および重要文数

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	41	11	516
ふしぎの国のアリス	43	9	886
白雪姫	15	8	722

4. 映画のプロットからの重要文抽出

具体的な映画の例として,「ふしぎの国のアリス」,「ローマの休日」,「白雪姫」について，プロット中の文の数を表 1 に示す．これから分かるように，Wikipedia 中のプロットは，比較的詳細に書かれている．人手による重要文抽出の際には，映画の物語において相対的に重要な内容を厳選して少数の重要文を抽出する必要がある．そこで，本論文では，重要文選定の際の手順として，以下の二段階の手順を経る.

1. 物語の進行の根幹を形成する最重要文として,
   - オープニング
   - 主たる物語の起点
   - 場面転換
   - 結末

   を抽出する.
2. 物語の進行の細部を把握するために，最重要 4 文以外の重要文を抽出する.

図 3 の上半分には,「ローマの休日」のプロットの一部において，最重要 4 文のうちの「主たる物語の起点」,「場面転換」，および,「最重要 4 文以外の重要文」を選定した様子を示す．また，表 1 においては,「ローマの休日」,「ふしぎの国のアリス」,「白雪姫」について，実際に選定された重要文の数を示す．このうち,「ローマの休日」については，抽出された重要文全 11 文を図 4 に示す.

5. 動画のショット分割

ショットとは，図 3 の最下部において連続するコマの画像を示すように，長時間の映像全体の中で，カメラが切り替わるまでの間の映像の単一断片を指す．通常，一つのショット内部の映像においては，大きな変化が起こらない．例えば，一つのショットの例としては，動かない景色，車の連続の移動，人が水を飲む一連の動作，等が挙げられる．本論文では，一つのショット内部の映像においては大きな変化が起こらないことに着目し，静止画としてのショット画像を閲覧して重要文と対応付けることによって，効率よい映画要約を行う.

本論文では，映画映像のショット分割を行う際には，動画をショット列に分割するツール [Apostolidis 14][*2*3]を適用する.

*1 日本語版 Wikipedia の場合は,「あらすじ」というタイトルの段落において記述される.

*2 http://mklab.iti.gr/project/video-shot-segm

*3 [Apostolidis 14] において報告されているショット分割精度は，88.7%である.

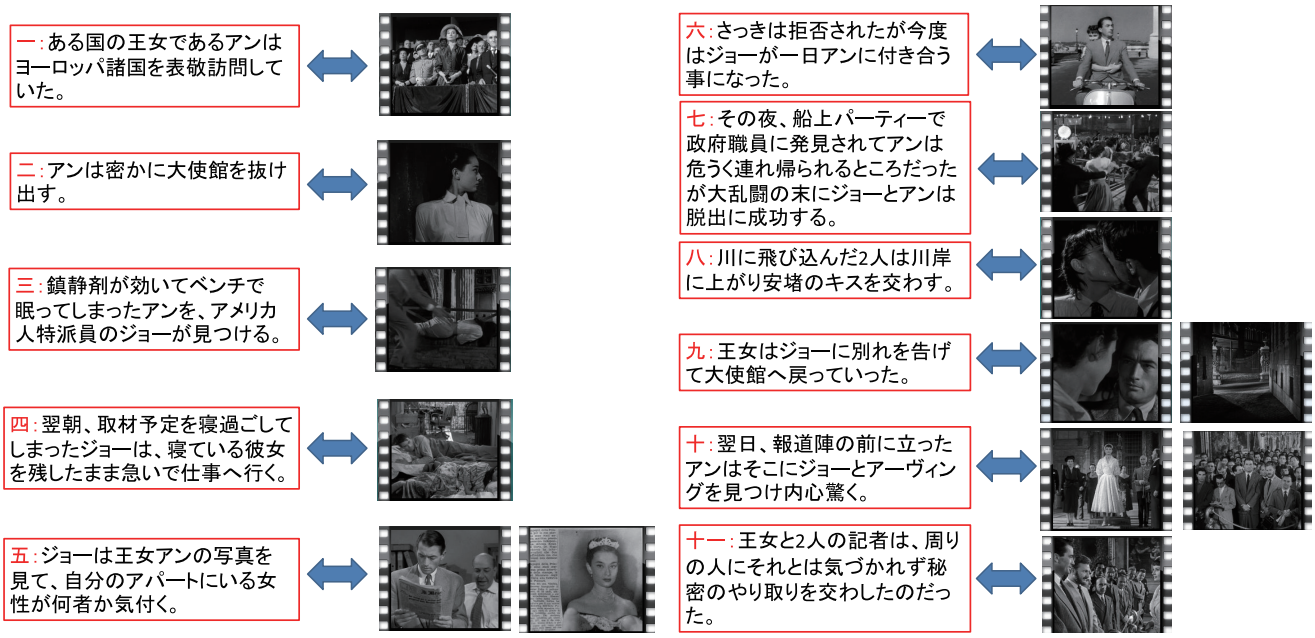


図 4: Wikipedia 中のプロットから選定した重要文とショットの対応付け結果 (「ローマの休日」の例)

ここで，分割結果のショットの時間長が 1 分以上の場合には，ショット分割不足の可能性があるので，同一ツールを用いて再分割を行う．

6. 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，[Tapaswi 15] における方式を用いる．[Tapaswi 15] 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．

[Tapaswi 15] 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 (6.1 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 (6.2 節) であり，α は重みパラメータである．

また，[Tapaswi 15] 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．[Tapaswi 15] 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 ([Tapaswi 15] の方式に従い，本論文では，上限を 83 とする) を設ける．

6.1 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．

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

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．

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

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．

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

6.2 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 word-match(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．

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

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

6.3 予備調査

「ローマの休日」を対象とする予備調査を行った．ただし，[Tapaswi 15] の方式の性能の上限を見積もるために，時間情報付き字幕 (サブタイトル) とシーン描写 (スクリプト) の間の対応付けは人手で行った．また，プロット文中の代名詞についても，人手でその照応先の人名に書き換えた後，予備調査を行なった．そして，プロットとショットを自動対応付けした結果のうち，特に，4. 節で選定した重要文 11 文とショットとの対応結果を評価した．この結果においては，11 文中 8 文に対する対応付け結果において，人手で作成した参照用対応ショット (平均約 20 ショット) が含まれていた．これより，約 73%の対応付け精度を達成できた．

ここで，本質的な問題点として，本論文の本来の目的は，映画や映像を要約するための手段としてプロット中の各文とショットを対応付けることにあるのに対して，[Tapaswi 15] の主目的は，プロット中の各文をショットに対応付けた結果を用いることによって，各ショットによって構成される映像を検索することである点が挙げられる．映像検索が主目的の場合には，検索漏れを防ぐために，すべてのショットを検索対象とする必要がある．一方，映像要約が主目的の場合には，むしろその逆に，プロット中の各文に対応する少数のショットを正確に同定する必要がある．そこで，性能を改善するためのもう一つの本質的な方策として，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ことが挙げられる．

7. 関連研究

映画等の映像要約についての関連研究として，[出口 04] においては，映画の文法に基づき，アクション区間，緊迫した区間，落ち着いた区間を抽出し，映画要約の際の特徴量として用いる手法を提案している．また，[吉高 07] においては，映画やドラマなどの撮影・編集上の技法により感性情報が強調される場面に着目し，映像要約の際の手がかりとして検討を行っている．

一方，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，[Liang 09] 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，[Yi 04] 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，[Tsoneva 07] 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，[Tapaswi 15] 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．その他，[中村 97] においては，字幕および映像中の特徴を併用してニュース映像中の重要部分を抽出する手法を提案している．

8. おわりに

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，[Tapaswi 15] において用いられている映画プロットと映像中のショットの対応付け方式を実装するとともに，[Sidiropoulos 11] において提案されている映像中のシーン分割方式との併用を行い，映画プロットと映像中のショットを高精度に対応付けることが挙げられる．また，時間情報付きサブタイトルとスクリプトを対応付けた上で，映像中のショットおよびシーンと対応付けた後，映画プロットと統合して構造化する方式を確立する．

参考文献

[Apostolidis 14] Apostolidis, E. and Mezaris, V.: Fast Shot Segmentation Combining Global and Local Visual Descriptors1, in *Proc. ICASSP*, pp. 6583–6587 (2014)

[出口 04] 出口 嘉紀, 吉高 淳夫：映画の文法に基づく要約映像の生成, 情報処理学会研究報告, Vol. 2004–DBS–132, pp. 33–40 (2004)

[Liang 09] Liang, C., Zhang, Y., Cheng, J., Xu, C., and Lu, H.: A Novel Role-based Movie Scene Segmentation Method, in *Advances in Multimedia Information Processing — PCM2009*, Vol. 5879 of *LNCS*, pp. 917–922, Springer (2009)

[中村 97] 中村 裕一, 金出 武雄：ニュース映像からの重要セグメント抽出 — 画像特徴と言語特徴の相互関係を用いたニュース映像要約, 第 3 回知能情報メディアシンポジウム, pp. 61–68 (1997)

[Sidiropoulos 11] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., and Trancoso, I.: Temporal Video Segmentation to Scenes using High-level Audiovisual Features, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 21, No. 8, pp. 1163–1177 (2011)

[Tapaswi 15] Tapaswi, M., Bäuml, M., and Stiefelhagen, R.: Aligning Plot Synopses to Videos for Story-based Retrieval, *International Journal of Multimedia Information Retrieval*, Vol. 4, No. 1, pp. 3–16 (2015)

[Tsoneva 07] Tsoneva, T., Barbieri, M., and Weda, H.: Automated Summarization of Narrative Video on a Semantic Level, in *Proc. Semantic Computing*, pp. 169–176 (2007)

[Yi 04] Yi, H., Rajan, D., and Chia, L.-T.: Semantic Video Indexing and Summarization using Subtitles, in *Advances in Multimedia Information Processing — PCM2004*, Vol. 3331 of *LNCS*, pp. 634–641, Springer (2004)

[吉高 07] 吉高 淳夫, 田中 壮詩, 平嶋 宗：映画等を対象としたダイジェスト映像生成のための映像特徴に関する検討, 情報処理学会研究報告, Vol. 2007–HCI–124, pp. 79–86 (2007)