(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7378172号
(P7378172)

(45)発行日　令和5年11月13日(2023.11.13)

(24)登録日　令和5年11月2日(2023.11.2)

(51)国際特許分類		FI		
H04N	*21/8549(2011.01)*	H04N	21/8549	
G06V	*10/82 (2022.01)*	G06V	10/82	
G06N	*3/08 (2023.01)*	G06N	3/08	
G06T	*7/00 (2017.01)*	G06T	7/00	350C
G06N	*3/045(2023.01)*	G06N	3/045	

請求項の数　8　（全24頁）

(21)出願番号　特願2022-72173(P2022-72173)
(22)出願日　令和4年4月26日(2022.4.26)
(65)公開番号　特開2023-129179(P2023-129179A)
(43)公開日　令和5年9月14日(2023.9.14)
審査請求日　令和4年4月26日(2022.4.26)
(31)優先権主張番号　10-2022-0026671
(32)優先日　令和4年3月2日(2022.3.2)
(33)優先権主張国・地域又は機関
韓国(KR)

(73)特許権者　505224569
インハ インダストリー パートナーシップ インスティテュート
Ｉｎｈａ－Ｉｎｄｕｓｔｒｙ Ｐａｒｔｎｅｒｓｈｉｐ Ｉｎｓｔｉｔｕｔｅ
大韓民国 ２２２１２ インチョン ミチョルーグ インハーロ １００
(74)代理人　110001519
弁理士法人太陽国際特許事務所
(72)発明者　チョ・グンシク
大韓民国 ２２２１２ インチョン ミチョルーグ インハーロ １００
(72)発明者　ユン・ウィニョン
大韓民国 ２２２１２ インチョン ミチョルーグ インハーロ １００

最終頁に続く

(54)【発明の名称】　効率的なキーフレーム選択報酬関数を備えた教師なし映像要約方法および装置

(57)【特許請求の範囲】

【請求項１】

フレームレベル映像特徴抽出モジュールによって入力映像からフレームレベル視覚的特徴を抽出する段階、

アテンション基盤の映像要約ネットワークモジュールによってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す段階、

評価モジュールによって前記キーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する段階、

前記予測された重要度点数によって該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、方策勾配アルゴリズム基盤の学習モジュールによって前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ Ｇｒａｄｉｅｎｔ）学習を実行する段階、

前記選択されたキーフレームの重要度点数を用いて前記映像要約ネットワークモジュールによってキーフレームを選択する確率を制御するための正規化および再構成損失を計算する段階、および

前記計算された正規化および再構成損失に基づいて映像要約生成モジュールによって映像要約を生成する段階

を含む、アテンション基盤の映像要約方法。
【請求項２】
前記アテンション基盤の映像要約ネットワークモジュールによってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す段階は、
エンコーダネットワーク、デコーダネットワーク、および前記エンコーダネットワークとデコーダネットワークの間のアテンション層において、拡張ＲＮＮによってパラメータと計算を減少させて時間依存性を抽出し、
前記エンコーダネットワークは、キーフレーム同士のローカルおよびグローバルコンテキストと視覚的類似性をキャプチャし、
前記アテンション層では、エンコーダネットワークの出力と以前のデコーダネットワークの隠れ状態の両方を用いてアテンション加重値を計算し、
前記アテンション加重値は、ソフトマックス関数に基づいて各キーフレームの確率点数に正規化し、
前記アテンション加重値を用いてエンコーダネットワークの出力に掛けてコンテキストベクトルを求め、
前記デコーダネットワークの入力のためにコンテキストベクトルと初期化されたデコーダネットワークの以前の出力を連結して前記デコーダネットワークを学習し、前記デコーダネットワークおよび前記エンコーダネットワークの学習結果を利用して重要度点数を求める
請求項１に記載のアテンション基盤の映像要約方法。
【請求項３】
前記評価モジュールによってキーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する段階は、
前記代表性報酬関数を利用して抽出された特徴を用いて選択したキーフレームと映像のすべてのキーフレームの類似性を計算し、前記代表性報酬関数を利用して映像要約のキーフレームを選択するための重要度点数を予測するように学習し、
前記時間的一貫性報酬関数を利用して代表的なショットレベルキーフレームを選択するために、すべてのキーフレームに対して選択されたキーフレームのうちから最も近い隣りを見つけ出す過程を繰り返し学習する
請求項１に記載のアテンション基盤の映像要約方法。
【請求項４】
前記予測された重要度点数によって該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、方策勾配アルゴリズム基盤の学習モジュールによって前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ　Ｇｒａｄｉｅｎｔ）学習を実行する段階は、
過小評価報酬（Ｕｎｄｅｒ－ａｐｐｒｅｃｉａｔｅｄ　Ｒｅｗａｒｄ：ＵＲＥＸ）方法を探索する探索戦略の目的関数を使用して、目標関数の近似値のためにソフトマックス関数を用いて各エピソードに対する報酬の正規化された重要度加重値集合を計算することによってパラメータ化された方策勾配学習を実行する
請求項１に記載のアテンション基盤の映像要約方法。
【請求項５】
入力映像からフレームレベル視覚的特徴を抽出するフレームレベル映像特徴抽出モジュール、
アテンション基盤の映像要約ネットワークによってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す映像要約ネットワークモジュール、
前記キーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離

による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する評価モジュール、
前記予測された重要度点数によって該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ　Ｇｒａｄｉｅｎｔ）学習を実行する方策勾配アルゴリズム基盤の学習モジュール（前記選択されたキーフレームの重要度点数を用いて前記映像要約ネットワークモジュールによってキーフレームを選択する確率を制御するための正規化および再構成損失を計算する）、および
前記計算された正規化および再構成損失に基づいて映像要約を生成する映像要約生成モジュール
を含む、アテンション基盤の映像要約装置。
【請求項６】
前記映像要約ネットワークモジュールは、
エンコーダネットワーク、デコーダネットワーク、および前記エンコーダネットワークとデコーダネットワークの間のアテンション層において、拡張ＲＮＮによってパラメータと計算を減少させて時間依存性を抽出し、
前記エンコーダネットワークは、キーフレーム同士のローカルおよびグローバルコンテキストと視覚的類似性をキャプチャし、
前記アテンション層では、エンコーダネットワークの出力と以前のデコーダネットワークの隠れ状態の両方を用いてアテンション加重値を計算し、
前記アテンション加重値は、ソフトマックス関数に基づいて各キーフレームの確率点数に正規化し、
前記アテンション加重値を用いてエンコーダネットワークの出力に掛けてコンテキストベクトルを求め、
前記デコーダネットワークの入力のためにコンテキストベクトルと初期化されたデコーダネットワークの以前の出力を連結して前記デコーダネットワークを学習し、前記デコーダネットワークおよび前記エンコーダネットワークの学習結果を利用して重要度点数を求める
請求項５に記載のアテンション基盤の映像要約装置。
【請求項７】
前記評価モジュールは、
前記代表性報酬関数を利用して抽出された特徴を用いて選択したキーフレームと映像のすべてのキーフレームの類似性を計算し、前記代表性報酬関数を利用して映像要約のキーフレームを選択するための重要度点数を予測するように学習し、
前記時間的一貫性報酬関数を利用して代表的なショットレベルキーフレームを選択するために、すべてのキーフレームに対して選択されたキーフレームのうちから最も近い隣りを見つけ出す過程を繰り返し学習する
請求項５に記載のアテンション基盤の映像要約装置。
【請求項８】
前記方策勾配アルゴリズム基盤の学習モジュールは、
過小評価報酬（Ｕｎｄｅｒ－ａｐｐｒｅｃｉａｔｅｄ　Ｒｅｗａｒｄ：ＵＲＥＸ）方法を探索する探索戦略の目的関数を使用し、目標関数の近似値のためにソフトマックス関数を用いて各エピソードに対する報酬の正規化された重要度加重値集合を計算することによってパラメータ化された方策勾配学習を実行する
請求項５に記載のアテンション基盤の映像要約装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、効率的なキーフレーム選択報酬関数を備えた教師なし映像要約方法および装

置に関する。
【背景技術】
【０００２】
ＹｏｕＴｕｂｅ（登録商標）のようなオンライン動画共有プラットフォームを利用して興味のある映像を検索するために時間を費やす人が多くなった。この時間を短縮するためにプレビューまたは要約映像を使用することで、全体映像コンテンツを効率的かつ迅速に把握することができる（非特許文献１）。ここ数年間にわたって映像要約が重要視されるようになっており、映像コンテンツを検索したり、長い映像から短い形式の要約映像を製作したりするための研究が積極的に行われている。しかし、映像要約では、映像のフレームレベルまたはショットレベルの重要度点数を予測するための困難な作業が問題となっている（非特許文献２）。映像要約は、明らかな視聴覚パターンや意味の規則がない、抽象的かつ主観的なマルチモード作業であるためである。映像のフレームが興味深いものであるか有益なものであれば、フレームの重要度点数が高くなければならない。また、このような高い点数のフレームが、映像要約を製作するために選択される。
【０００３】
最近は多様な方法が提案されているが、このような方法はディープラーニング（非特許文献３、４、５）を利用することで高い性能を発揮する。ディープラーニング基盤の映像要約方法は、教師あり学習基盤の方法と教師なし学習基盤の方法とに分けられる。教師あり学習基盤の方法の場合は、ラベルが指定されたデータセットを生成するのが極めて難しい。さらに、多様なドメインや場面を包括した大規模なデータセットを生成することも極めて難しい。このような理由により、教師なし映像要約方法の開発に力が注がれてきた。
【０００４】
強化学習（Ｒｅｉｎｆｏｒｃｅｍｅｎｔ　Ｌｅａｒｎｉｎｇ：ＲＬ）基盤の映像要約方法が従来技術（非特許文献６）で提案され、優れた成果を示した。特に、ＲＬを用いて深層ニューラルネットワークを学習させるために、報酬関数であるキーフレームを選択するための効率的かつ明示上な評価方法がある。さらに、深層ニューラルネットワークは、評価方法を利用して、代表性、多様性、統一性などの映像の多様な特徴を効率的に学習する。ＲＬを使用しながら、部分線形補間法を使用する従来技術（非特許文献３）でＩｎｔｅｒ－ＳＵＭを提案した。補間法を用いることとでネットワークの出力を減らし、高い分散問題を緩和して性能を改善した。しかし、多くの映像において、キーフレームが特定の場面からしか選択されなかったり、興味深いキーフレームが適切に選択されなかったりした。
【０００５】
さらに、以前のＲＬ基盤の映像要約方法にはいくつかの短所がある。一つ目に、深層ニューラルネットワークによっては視覚的および時間的コンテキストの捕捉が難しい。二つ目に、多くの方法は、キーフレームの時間的分布を考慮せず、キーフレーム同士の視覚的な差を計算してネットワークを学習させるために報酬関数または損失関数を使用する。したがって、一律的に映像を容易に理解できるようにキーフレームを選択して演出者のストーリーラインを生かすことのできる映像要約方法が求められている。
【先行技術文献】
【非特許文献】
【０００６】
【文献】Ｅｊａｚ，Ｎ．；Ｍｅｈｍｏｏｄ，Ｉ．；Ｂａｉｋ，Ｓ．Ｗ．Ｅｆｆｉｃｉｅｎｔ ｖｉｓｕａｌ ａｔｔｅｎｔｉｏｎ ｂａｓｅｄ ｆｒａｍｅｗｏｒｋ ｆｏｒ ｅｘｔｒａｃｔｉｎｇ ｋｅｙ ｆｒａｍｅｓ ｆｒｏｍ ｖｉｄｅｏｓ．Ｊ．Ｉｍａｇｅ Ｃｏｍｍｕｎ．２０１３，２８，３４－４４．
【文献】Ｇｙｇｌｉ，Ｍ．；Ｇｒａｂｎｅｒ，Ｈ．；Ｒｉｅｍｅｎｓｃｈｎｅｉｄｅｒ，Ｈ．；Ｇｏｏｌ，Ｌ．Ｖ．Ｃｒｅａｔｉｎｇ ｓｕｍｍａｒｉｅｓ ｆｒｏｍ ｕｓｅｒ ｖｉｄｅｏｓ．Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ Ｅｕｒｏｐｅａｎ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｃｏｍｐｕｔｅｒ Ｖｉｓｉｏｎ（ＥＣＣＶ））；Ｓｐｒｉｎｇｅｒ，２０１５，ｐｐ．５０５－５２０．

【文献】Yoon, U, N.；Hong, M. D.；Jo, G. S. Interp-SUM：Unsupervised Video Summarization with Piecewise Linear Interpolation. Sensors 2021. vol. 21, no. 13, 4562.
【文献】Apostolidis, E.；Adamantidou, E.；Metsai, A.；Mezaris, V.；Patras, I. Unsupervised Video Summarization via Attention-Driven Adversarial Learning. In International Conference on Multimedia Modeling (MMM)；Springer：Daejeon, Korea, 5-8 January 2020, pp. 492-504.
【文献】Jung, Y. J.；Cho, D. Y.；Kim, D. H.；Woo, S. H.；Kweon, I. S. Discriminative feature learning for unsupervised video summarization. AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, 27 January-1 February 2019, pp. 8537-8544。
【文献】Zhou, K.；Qiao, Y.；Xiang, T. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. AAAI Conf. Artif. Intell. 2018, 32, 7582-7589.
【文献】Song, Y.；Vallmitjana, J.；Stent, A.；Jaimes, A. Tvsum：Summarizing web videos using titles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7-12 June 2015, pp. 5179-5187.
【文献】Feng, L.；Li, Z.；Kuang, Z.；Zhang, W. Extractive Video Summarizer with Memory Augmented Neural Networks. MM, Seoul, Republic of Korea, 22-26 October 2018, pp. 976-983.
【文献】Zhang, K.；Chao, W. L.；Sha, F.；Grauman, K. Video Summarization with Long Short-term Memory. In Proceedings of the European Conference on Computer Vision (ECCV)；Springer：Amsterdam, Netherlands, 2016；pp. 766-782.
【文献】Zhang, Y.；Kampffmeyer, M.；Zhao, X.；Tan, M. DTR-GAN：Dilated Temporal Relational Adversarial Network for Video Summarization. In Proceedings of the ACM Turing Celebration Conference (ACM TURC), Shanghai, China, 18 May 2018.
【文献】Ji. Z.；Xiong. K.；Pang. Y.；Li. X. Video Summarization with Attention-Based Encoder-Decoder Networks. IEEE Transactions on circuits and systems for video technology, June 2020, vol. 30, no. 6. pp. 1709-1717.
【文献】Mahasseni, B.；Lam, M.；Todorovic, S. Unsupervised Video Summarization with Adversa

rial LSTM Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu、Hawaii, USA, 22−25 July 2017, pp. 202−211.
【文献】Yuan, L.；Tay, F. E.；Li, P.；Zhou, L.；Feng, F. Cycle−SUM：Cycle−consistent Adversarial LSTM Networks for Unsupervised Video Summarization. In Proceedings of the Thirty−Third AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, 27 Jan−1 Feb 2019, Volume 33, pp. 9143−9150.
【文献】Kaufman, D.；Levi, G.；Hassner, T.；Wolf, L. Temporal Tessellation：A Unified Approach for Video Analysis. In Proceedings of the IEEE International Conference on Computer Vision（ICCV）, 2017, pp. 94−104.
【文献】Rochan, M.；Ye, L.；Wang, Y. Video Summarization Using Fully Convolutional Sequence Networks. In Proceedings of the Eu−ropean Conference on Computer Vision（ECCV）；Springer：Munich, Germany, 2018, pp. 347−363.
【文献】Silver, D.；Lever, G.；Heess, N.；Degris, T.；Wierstra, D.；Riedmiller, M. Deterministic Policy Gradient Algorithms. In Pro−ceedings of the 31st International Conference on International Conference on Machine Learning（ICML）, Beijing, China, 21−26 June 2014, pp. 387−395.
【文献】Yu, Y. Towards Sample Efficient Reinforcement Learning. In Proceedings of the Twenty−Seventh International Joint Con−ference on Artificial Intelligence（IJCAI）, Stockholm, Sweden, 13−19 July2018, pp. 5739−5743.
【文献】Lehnert, L.；Laroche, R.；Seijen, H. V. On Value Function Representation of Long Horizon Problems. In Proceedings of the Thirty−Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, 2−7 Feb 2018, pp. 3457−3465.
【文献】Szegedy, C.；Liu, W.；Jia, Y.；Sermanet, P.；Reed, S.；Anguelov, D.；Erhan, D.；Vanhoucke, B.；Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition（CVPR）, Boston, MA, USA, 7−12 June 2015, pp. 1−9.
【文献】Luong. T.；Pham. H.；Manning C. D.；Effective Approaches to Attention−based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natu

ｒａｌ Ｌａｎｇｕａｇｅ Ｐｒｏｃｅｓｓｉｎｇ（ＥＭＮＬＰ），Ｌｉｓｂｏｎ，Ｐｏｒｔｕｇａｌ，１７－２１ Ｓｅｐｔｅｍｂｅｒ ２０１５，ｐｐ．１４１２－１４２１．
【文献】Ｎａｃｈｕｍ，Ｏ．；Ｎｏｒｏｕｚｉ，Ｍ．；Ｓｃｈｕｕｒｍａｎｓ，Ｄ．Ｉｍｐｒｏｖｉｎｇ Ｐｏｌｉｃｙ Ｇｒａｄｉｅｎｔ ｂｙ Ｅｘｐｌｏｒｉｎｇ Ｕｎｄｅｒ－Ａｐｐｒｅｃｉａｔｅｄ Ｒｅｗａｒｄｓ．ａｒＸｉｖ ２０１６，ａｒＸｉｖ：１６１１．０９３２１．
【文献】Ｐｏｔａｐｏｖ，Ｄ．；Ｄｏｕｚｅ，Ｍ．；Ｈａｒｃｈａｏｕｉ，Ｚ．；Ｓｃｈｍｉｄ，Ｃ．；Ｃａｔｅｇｏｒｙ－ｓｐｅｃｉｆｃ ｖｉｄｅｏ ｓｕｍｍａｒｉｚａｔｉｏｎ．Ｅｕｒｏｐｅａｎ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｃｏｍｐｕｔｅｒ Ｖｉｓｉｏｎ（ＥＣＣＶ），Ｚｕｒｉｃｈ，Ｓｗｉｔｚｅｒｌａｎｄ，Ｓｅｐ ２０１４，ｐｐ．５４０－５５５．
【文献】Ｒｏｃｈａｎ，Ｍ．；Ｗａｎｇ，Ｙ．Ｖｉｄｅｏ Ｓｕｍｍａｒｉｚａｔｉｏｎ ｂｙ Ｌｅａｒｎｉｎｇ ｆｒｏｍ Ｕｎｐａｉｒｅｄ Ｄａｔａ．Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ＩＥＥＥ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｃｏｍｐｕｔｅｒ Ｖｉｓｉｏｎ ａｎｄ Ｐａｔｔｅｒｎ Ｒｅｃｏｇｎｉｔｉｏｎ（ＣＶＰＲ）２０１９，ｐｐ．７９０２－７９１１．

【発明の概要】

【発明が解決しようとする課題】

【０００７】

本発明が達成しようとする技術的課題は、新たな時間的一貫性報酬関数および代表性報酬関数（ＴＲ－ＳＵＭ）を備えた強化学習基盤の映像要約フレームワーク、および重要度点数を正確に予測するためのアテンション基盤の映像要約方法および装置を提案することにある。より詳細には、長い映像のコンテキストを効率的にキャプチャするために、アテンション基盤のエンコーダ－デコーダアーキテクチャをもつ映像要約ネットワークおよび関心のあるキーフレームを効率的かつ均一に選択するための時間的一貫性報酬関数および代表性報酬関数である新たな報酬関数を提案する。

【課題を解決するための手段】

【０００８】

一側面において、本発明で提案する効率的なキーフレーム選択報酬関数を備えた教師なし映像要約方法は、フレームレベル映像特徴抽出モジュールによって入力映像からフレームレベル視覚的特徴を抽出する段階、アテンション基盤の映像要約ネットワークモジュールによってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す段階、評価モジュールによって前記キーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する段階、前記予測された重要度点数によって該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、方策勾配アルゴリズム基盤の学習モジュールによって前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ Ｇｒａｄｉｅｎｔ）学習を実行する段階、前記選択されたキーフレームの重要度点数を利用して前記映像要約ネットワークモジュールによってキーフレームを選択する確率を制御するための正規化および再構成損失を計算する段階、および前記計算された正規化および再構成損失に基づいて映像要約生成モジュールによって映像要約を生成する段階を含む。

【０００９】

前記アテンション基盤の映像要約ネットワークモジュールによってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す段階は、エンコーダネットワーク、デコーダネットワーク、および前記エンコーダネットワークとデコーダネットワークの間のアテンション層において、拡張ＲＮＮによってパラメータと計算を減少させて時間依存性を抽出し、前記エン

コーダネットワークは、キーフレーム同士のローカルおよびグローバルコンテキストと視覚的類似性をキャプチャし、前記アテンション層ではエンコーダネットワークの出力と以前のデコーダネットワークの隠れ状態の両方を用いてアテンション加重値を計算し、前記アテンション加重値は、ソフトマックス関数に基づいて各キーフレームの確率点数に正規化し、前記アテンション加重値を用いてエンコーダネットワークの出力に掛けてコンテキストベクトルを求め、前記デコーダネットワークの入力のためにコンテキストベクトルと初期化されたデコーダネットワークの以前の出力を連結して前記デコーダネットワークを学習し、前記デコーダネットワークおよび前記前記エンコーダネットワークの学習結果を利用して重要度点数を求める。

【００１０】

前記評価モジュールによってキーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する段階は、前記代表性報酬関数を利用して抽出された特徴を用いて選択したキーフレームと映像のすべてのキーフレーム同士の類似性を計算し、前記代表性報酬関数に基づいて映像要約のキーフレームを選択するための重要度点数を予測するように学習し、前記時間的一貫性報酬関数を利用して代表的なショットレベルキーフレームを選択するために、すべてのキーフレームに対して選択されたキーフレームのうちから最も近い隣りを見つけ出す過程を繰り返し学習する。

【００１１】

前記予測された重要度点数によって該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、方策勾配アルゴリズム基盤の学習モジュールによって前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ　Ｇｒａｄｉｅｎｔ）学習を実行する段階は、過小評価報酬（Ｕｎｄｅｒ－ａｐｐｒｅｃｉａｔｅｄ　Ｒｅｗａｒｄ：ＵＲＥＸ）方法を探索する探索戦略の目的関数を使用し、目標関数の近似値のためにソフトマックス関数を用いて各エピソードに対する報酬の正規化された重要度加重値集合を計算することによってパラメータ化された方策勾配学習を実行する。

【００１２】

また他の側面において、本発明で提案する効率的なキーフレーム選択報酬関数を備えた教師なし映像要約装置は、入力映像からフレームレベル視覚的特徴を抽出するフレームレベル映像特徴抽出モジュール、アテンション基盤の映像要約ネットワークによってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す映像要約ネットワークモジュール、前記キーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する評価モジュール、前記予測された重要度点数によって該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ　Ｇｒａｄｉｅｎｔ）学習を実行する方策勾配アルゴリズム基盤の学習モジュール（前記選択されたキーフレームの重要度点数を使用して、前記映像要約ネットワークモジュールによってキーフレームを選択する確率を制御するための正規化および再構成損失を計算する）、前記計算された正規化および再構成損失に基づいて映像要約を生成する映像要約生成モジュール映像要約生成モジュールを含む。

【発明の効果】

【００１３】

本発明の実施形態によると、新たな時間的一貫性報酬関数および代表性報酬関数（ＴＲ－ＳＵＭ）を備えた強化学習基盤の映像要約フレームワークおよび重要度点数を正確に予測するためのアテンション基盤の映像要約方法および装置を提案する。提案する教師なし

映像要約方法は、アテンション基盤のエンコーダ－デコーダアーキテクチャをもつ映像要約ネットワークによって長い映像のコンテキストを効率的にキャプチャすることができ、時間的一貫性報酬関数および代表性報酬関数である新たな報酬関数を利用することで関心のあるキーフレームを効率的かつ均一に選択することができる。

【図面の簡単な説明】

【００１４】

【図１】本発明の一実施形態における、アテンション基盤の映像要約ネットワークの概念を説明するための図である。

【図２】本発明の一実施形態における、時間一貫性報酬および代表性報酬関数を利用した強化学習基盤の映像要約フレームワークを示した図である。

【図３】本発明の一実施形態における、効率的なキーフレーム選択報酬関数を備えた教師なし映像要約装置の構成を示した図である。

【図４】本発明の一実施形態における、効率的なキーフレーム選択報酬関数を備えた教師なし映像要約方法を説明するためのフローチャートである。

【図５】本発明の一実施形態における、アテンション加重値を計算するためのアテンション基盤の映像要約ネットワークの構成を示した図である。

【図６】本発明の一実施形態における、代表性報酬関数を説明するための図である。

【図７】本発明の一実施形態における、時間的一貫性報酬関数を説明するための図である。

【発明を実施するための形態】

【００１５】

本発明は、効率的なキーフレーム選択報酬関数を備えた教師なし映像要約方法を提案する。効率的なキーフレームの選択のために、新たな時間的一貫性報酬および代表性報酬関数（ＴＲ－ＳＵＭ）を備えた強化学習基盤の映像要約フレームワークを設計する。映像のキーフレームレベル重要度点数を予測するために、アテンション基盤のエンコーダ－デコーダアーキテクチャとして映像要約方法および装置を提案する。本発明の実施形態によると、重要度点数を使用して、関心のあるキーフレームを効率的かつ均一に選択するために役立つ報酬関数に基づいて報酬を計算する。以下、本発明の実施例について、添付の図面を参照しながら詳しく説明する。

【００１６】

図１は、本発明の一実施形態における、アテンション基盤の映像要約ネットワークの概念を説明するための図である。

【００１７】

本発明では、新たな時間的一貫性報酬および代表性報酬関数（ＴＲ－ＳＵＭ（ｖｉｄｅｏ ｓｕｍｍａｒｉｚａｔｉｏｎ ｆｒａｍｅｗｏｒｋ ｗｉｔｈ ｔｈｅ ｎｅｗ ｔｅｍｐｏｒａｌ ｃｏｎｓｉｓｔｅｎｃｙ ｒｅｗａｒｄ ａｎｄ ｒｅｐｒｅｓｅｎｔａｔｉｖｅｎｅｓｓ ｒｅｗａｒｄ ｆｕｎｃｔｉｏｎｓ））を備えた強化学習基盤の映像要約フレームワークを提案し、図１に示すように、重要度点数を正確に予測するためのアテンション基盤の映像要約ネットワークを提案する。

【００１８】

提案する教師なし映像要約方法は、アテンション基盤のエンコーダ－デコーダアーキテクチャをもつ映像要約ネットワークによって長い映像のコンテキストを効率的にキャプチャすることができ、時間的一貫性報酬関数および代表性報酬関数である新たな報酬関数を利用することで関心のあるキーフレームを効率的かつ均一に選択することができる。

【００１９】

映像要約方法は、教師あり学習基盤の方法と教師なし学習基盤の方法とに分けられる。２つの方法すべて、多数の使用者が注釈をつけた映像のフレームレベルまたはショットレベルの重要度点数を含む映像要約データセットを使用する（非特許文献２）。教師あり学習基盤の方法は、重要度点数を予測するための入力として映像のフレームレベルまたはショットレベル特徴を用いてモデルを学習する。この方法は、データセットを使用して、予測された重要度点数と注釈がついた重要度点数の差によって費用を計算する。また、この

ような方法は、最適のモデルを探索するための費用を最小化する。

【００２０】

非特許文献８では、メモリ増強映像要約方法が提案された。メモリネットワークは、全体映像から抽出した支援知識を効率的に提供する。現在は、ショットの重要度点数を予測するために、グローバルアテンションメカニズムを用いて元の映像に対する全体的な理解によって点数を調整する。非特許文献９では、学習の代表性と多様性のために、フレームをサンプリングする確率をエンコードする決定的点過程（Ｄｅｔｅｒｍｉｎａｎｔａｌ Ｐｏｉｎｔ Ｐｒｏｃｅｓｓ：ＤＰＰ）のあるＬＳＴＭ基盤のネットワークを提示した。非特許文献１０では、映像フレーム同士の時間的コンテキストの表現を強化するために、生成器の拡張時間関係（Ｄｉｌａｔｅｄ Ｔｅｍｐｏｒａｌ Ｒｅｌａｔｉｏｎａｌ：ＤＴＲ）装置を提示した。映像の最上の要約を予測するためのネットワークを学習させるために、敵対的学習方法を３－プレイヤ（ｔｈｒｅｅ－ｐｌａｙｅｒ）損失関数とともに使用する。非特許文献１１では、キーショット選択のための重要度点数を予測するために、アテンション基盤のエンコーダ－デコーダネットワークが提案された。このネットワークは、双方向ＬＳＴＭネットワークがあるエンコーダとアテンションメカニズムがあるデコーダを用いて映像表現を学習する。

【００２１】

しかし、教師あり学習基盤の方法は、多様なドメインの映像を含んでいるため、人間がラベリングした映像要約データセットを生成することが極めて難しいという問題を抱えている。この反面、教師なし学習基盤の方法は、人間がラベリングしたデータセットを必要としない。

【００２２】

非特許文献４で、アテンションオートエンコーダ（Ａｔｔｅｎｔｉｏｎ Ａｕｔｏｅｎｃｏｄｅｒ：ＡＡＥ）ネットワークは、ＳＵＭ－ＧＡＮで提案された敵対的オートエンコーダ（Ａｄｖｅｒｓａｒｉａｌ Ａｕｔｏｅｎｃｏｄｅｒ：ＡＡＥ）の学習の効率性と性能を高めるためにＳＵＭ－ＧＡＮでＶＡＥ（Ｖａｒｉａｔｉｏｎａｌ Ａｕｔｏｅｎｃｏｄｅｒ）に替わる。ネットワークを学習させる間、映像を要約するための興味深いフレームに加重値を付与する。非特許文献５では、映像の要約を適切に予測するための映像のローカルおよびグローバルコンテキストを効率的に学習するために、ＶＡＥとＧＡＮ（Ｇｅｎｅｒａｔｉｖｅ Ａｄｖｅｒｓａｒｉａｌ Ｎｅｔｗｏｒｋｓ）アーキテクチャを基盤としたＣＳＮｅｔ（Ｃｈｕｎｋ ａｎｄ Ｓｔｒｉｄｅ Ｎｅｔｗｏｒｋ）を提案した。非特許文献１２では、敵対的オートエンコーダ（ＡＡＥ）基盤の映像要約モデルが提案された。選択器ＬＳＴＭは、映像の入力フレームレベル特徴からフレームを選択する。この後、ＶＡＥは、選択したフレームを用いて再構成された映像を生成する。全体ネットワークを学習させるために、判別器は、再構成された映像と原本入力映像とを区別する。モデル学習には４つの損失関数が使用される。非特許文献１３ではＳＵＭ－ＧＡＮの変形であるＣｙｃｌｅ－ＳＵＭを提案したが、要約映像で原本映像の情報を保存するために２つのＶＡＥ基盤の生成器と２つの判別器がある周期敵対的生成網を採択した。非特許文献１４で提案された、テッセレーション接近方式を使用する映像要約方法は、視覚的に類似するクリップを探索し、グラフ基盤の方法であるＶｉｔｅｒｂｉアルゴリズムを使用して時間的一貫性を維持するクリップを選択する。非特許文献１５では教師なし学習基盤のＳＵＭ－ＦＣＮを提案した。このような方法は、映像シーケンスを処理するために、空間畳み込みで変換された時間畳み込みをもつ新たなＦＣＮアーキテクチャを提示する。この方法は、デコーダの出力点数を使用してフレームを選択し、要約映像でフレームの多様性を適用するためにリペリング（ｒｅｐｅｌｌｉｎｇ）正規化器によって損失関数を計算する。

【００２３】

深層強化学習は、深層ニューラルネットワークを強化学習方法と結合する（非特許文献１６）。方策勾配（Ｐｏｌｉｃｙ Ｇｒａｄｉｅｎｔ）方法は、モデルがない強化学習方法のうちの１つである。方策勾配方法は、方策を深層ニューラルネットワークモデルにパラメータ化し、確率的勾配降下法（Ｓｔｏｃｈａｓｔｉｃ Ｇｒａｄｉｅｎｔ Ｄｅｓｃｅ

ｎｔ：ＳＧＤ）のような勾配降下方法を用いて、方策によって定義された状態分布に対する報酬を極大化してモデルを最適化する。モデルを学習させるために、この方法は、目的関数として費用を計算して最小化する。しかし、方策勾配方法には、低いサンプル効率性問題（非特許文献１７）と高い分散のようないくつかの問題を抱えている。特に、低いサンプル効率性の問題は、エージェントが人間ほど知能的でないため、環境（状態）において学習行動のための人間の経験のようなサンプルを人間よりも遥かに多く必要とするためである。他の問題は、推定された勾配の高分散である。この問題は、長い水平問題と高い次元の作用空間によって発生する（非特許文献１８）。長い水平の問題は、目標を達成するための長い一連の決定に対する極めて遅延した報酬から始まる。提案する方法では、分散を減らすために基準とともに方策勾配を使用し、サンプル効率性の問題を緩和するためにエピソード数を増やす。

【００２４】

図２は、本発明の一実施形態における、時間一貫性報酬および代表性報酬関数を利用した強化学習基盤の映像要約フレームワークを示した図である。

【００２５】

本発明では、映像要約ネットワークが予測した重要度点数（Ｉｍｐｏｒｔａｎｃｅ Ｓｃｏｒｅ）を用いて映像要約の問題をフレーム選択の問題として公式化する。特に、拡張された（ｄｉｌａｔｅｄ）ＧＲＵエンコーダとアテンションメカニズム（ａｔｔｅｎｔｉｏｎ ｍｅｃｈａｎｉｓｍ）があるＧＲＵデコーダネットワークを用いてネットワークを開発する。このネットワークは、映像表現を学習し、重要度点数をフレーム選択確率によって効率的に予測する。重要度点数は、図２に示すようにベルヌーイ分布（Ｂｅｒｎｏｕｉｉ Ｄｉｓｔｒｉｂｕｔｉｏｎ）を用いることで、要約からキーフレームを選択するためのフレーム選択動作に変換される。

【００２６】

図３は、本発明の一実施形態における、効率的なキーフレーム選択報酬関数を備えた教師なし映像要約装置の構成を示した図である。

【００２７】

提案する効率的なキーフレーム選択報酬関数を備えた教師なし映像要約装置は、フレームレベル映像特徴抽出モジュール３１０、映像要約ネットワークモジュール３２０、評価モジュール３３０、方策勾配アルゴリズム基盤の学習モジュール３４０、および映像要約生成モジュール３５０を含む。

【００２８】

本発明の実施形態に係るフレームレベル映像特徴抽出モジュール３１０は、入力映像からフレームレベル視覚的特徴を抽出し、アテンション基盤の映像要約ネットワークモジュール３２０によってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す。

【００２９】

本発明の実施形態に係る映像要約ネットワークモジュール３２０は、エンコーダネットワーク、デコーダネットワーク、および前記エンコーダネットワークとデコーダネットワークの間のアテンション層において、拡張ＲＮＮによってパラメータと計算を減少させて時間依存性を抽出する。

【００３０】

本発明の実施形態に係る映像要約ネットワークモジュール３２０は、前記エンコーダネットワークはキーフレーム同士のローカルおよびグローバルコンテキストと視覚的類似性をキャプチャし、前記アテンション層ではエンコーダネットワークの出力と以前のデコーダネットワークの隠れ状態の両方を用いてアテンション加重値を計算する。

【００３１】

本発明の実施形態に係る映像要約ネットワークモジュール３２０は、前記アテンション加重値はソフトマックス関数によって各キーフレームの確率点数に正規化し、前記アテンション加重値を用いてエンコーダネットワークの出力に掛けてコンテキストベクトルを求

める。
【００３２】
本発明の実施形態に係る映像要約ネットワークモジュール３２０は、前記デコーダネットワークの入力のためにコンテキストベクトルと初期化されたデコーダネットワークの以前の出力を連結して前記デコーダネットワークを学習し、前記デコーダネットワークおよび前記前記エンコーダネットワークの学習結果を利用して重要度点数を求める。
【００３３】
本発明の実施形態に係る評価モジュール３３０は、前記キーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する。
【００３４】
本発明の実施形態に係る評価モジュール３３０は、前記代表性報酬関数を利用して抽出された特徴を用いて選択したキーフレームと映像のすべてのキーフレーム同士の類似性を計算し、前記代表性報酬関数によって映像要約のキーフレームを選択するための重要度点数を予測するように学習する。
【００３５】
本発明の実施形態に係る評価モジュール３３０は、前記時間的一貫性報酬関数を利用して代表的なショットレベルキーフレームを選択するために、すべてのキーフレームに対して選択されたキーフレームのうちから最も近い隣りを見つけ出す過程を繰り返し学習する。
【００３６】
本発明の実施形態に係る方策勾配アルゴリズム基盤の学習モジュール３４０は、前記予測された重要度点数によって該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ　Ｇｒａｄｉｅｎｔ）学習を実行する。
【００３７】
本発明の実施形態に係る方策勾配アルゴリズム基盤の学習モジュール３４０は、過小評価報酬（Ｕｎｄｅｒ－ａｐｐｒｅｃｉａｔｅｄ　Ｒｅｗａｒｄ：ＵＲＥＸ）方法を探索する探索戦略の目的関数を使用し、目標関数の近似値のためにソフトマックス関数を用いて各エピソードに対する報酬の正規化された重要度加重値集合を計算することによってパラメータ化された方策勾配学習を実行する。
【００３８】
本発明の実施形態に係る映像要約ネットワークモジュール３２０は、前記選択されたキーフレームの重要度点数を使用して、キーフレームを選択する確率を制御するための正規化および再構成損失を計算する。本発明の実施形態に係る映像要約生成モジュール３５０は、前記計算された正規化および再構成損失に基づいて映像要約を生成する。
【００３９】
図４は、本発明の一実施形態における、効率的なキーフレーム選択報酬関数を備えた教師なし映像要約方法を説明するためのフローチャートである。
【００４０】
提案する、効率的なキーフレーム選択報酬関数を備えた教師なし映像要約方法は、フレームレベル映像特徴抽出モジュールによって入力映像からフレームレベル視覚的特徴を抽出する段階４１０、アテンション基盤の映像要約ネットワークモジュールによってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す段階４２０、評価モジュールによって前記キーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する段階４３０、前記予測された重要度点数によ

って該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、方策勾配アルゴリズム基盤の学習モジュールによって前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ　Ｇｒａｄｉｅｎｔ）学習を実行する段階４４０、前記選択されたキーフレームの重要度点数を用いて前記映像要約ネットワークモジュールによってキーフレームを選択する確率を制御するための正規化および再構成損失を計算する段階４２０、および前記計算された正規化および再構成損失に基づいて映像要約生成モジュールによって映像要約を生成する段階４５０を含む。

【００４１】

段階４１０で、フレームレベル映像特徴抽出モジュールによって入力映像からフレームレベル視覚的特徴を抽出する。

【００４２】

段階４２０で、アテンション基盤の映像要約ネットワークモジュールによってアテンション加重値を計算し、前記アテンション加重値を利用してキーフレームを選択するためのフレーム追跡確率として重要度点数を示す。

【００４３】

本発明の実施形態によると、エンコーダネットワーク、デコーダネットワーク、および前記エンコーダネットワークとデコーダネットワークの間のアテンション層において、拡張ＲＮＮによってパラメータと計算を減少させて時間依存性を抽出する。

【００４４】

前記エンコーダネットワークは、キーフレーム同士のローカルおよびグローバルコンテキストと視覚的類似性をキャプチャし、前記アテンション層ではエンコーダネットワークの出力と以前のデコーダネットワークの隠れ状態の両方を用いてアテンション加重値を計算する。

【００４５】

前記アテンション加重値は、ソフトマックス関数に基づいて各キーフレームの確率点数に正規化し、前記アテンション加重値を用いてエンコーダネットワークの出力に掛けてコンテキストベクトルを求める。

【００４６】

前記デコーダネットワークの入力のためにコンテキストベクトルと初期化されたデコーダネットワークの以前の出力を連結して前記デコーダネットワークを学習し、前記デコーダネットワークおよび前記前記エンコーダネットワークの学習結果を利用して重要度点数を求める。

【００４７】

段階４３０で、評価モジュールによってキーフレームを選択するためのキーフレーム同士の視覚的類似性距離と時間的距離による時間的一貫性報酬関数と代表性報酬関数を求め、時間的一貫性報酬関数と代表性報酬関数を利用してアテンション基盤の映像要約ネットワークが映像要約のキーフレームを選択するための重要度点数を予測するように学習する。

【００４８】

前記代表性報酬関数を利用して抽出された特徴を用いて選択したキーフレームと映像のすべてのキーフレームの類似性を計算し、前記代表性報酬関数によって映像要約のキーフレームを選択するための重要度点数を予測するように学習する。

【００４９】

前記時間的一貫性報酬関数を利用して代表的なショットレベルキーフレームを選択するために、すべてのキーフレームに対して選択されたキーフレームのうちから最も近い隣りを見つけ出す過程を繰り返し学習する。

【００５０】

段階４４０で、前記予測された重要度点数によって該当のキーフレームを選択して映像要約を生成し、生成された映像要約の品質を評価し、方策勾配アルゴリズム基盤の学習モジュールによって前記アテンション基盤の映像要約ネットワークに対する方策勾配（Ｐｏｌｉｃｙ　Ｇｒａｄｉｅｎｔ）学習を実行する。

【００５１】
本発明の実施形態によると、過小評価報酬（Ｕｎｄｅｒ－ａｐｐｒｅｃｉａｔｅｄ Ｒｅｗａｒｄ：ＵＲＥＸ）方法を探索する探索戦略の目的関数を使用し、目標関数の近似値のためにソフトマックス関数を用いて各エピソードに対する報酬の正規化された重要度加重値集合を計算することによってパラメータ化された方策勾配学習を実行する。
【００５２】
再び段階４２０で、前記選択されたキーフレームの重要度点数を用いて映像要約ネットワークモジュールによってキーフレームを選択する確率を制御するための正規化および再構成損失を計算し、段階４５０で、前記計算された正規化および再構成損失に基づいて映像要約生成モジュールによって映像要約を生成する。
【００５３】
図５～７を参照しながら、効率的なキーフレーム選択報酬関数を備えた教師なし映像要約方法についてさらに詳しく説明する。
【００５４】
図５は、本発明の一実施形態における、アテンション加重値を計算するためのアテンション基盤の映像要約ネットワークの構成を示した図である。
【００５５】
先ず、ＩｍａｇｅＮｅｔデータセットによって学習された強力な深層畳み込みニューラルネットワークであるＧｏｏｇｌｅＮｅｔ（非特許文献１９）を使用して、入力映像からフレームレベル視覚的特徴
$\{x_t\}_{t=1}^{N}$
を抽出する。特徴の抽出は、フレームイメージの視覚的特性を低次元特徴ベクトルで捕捉するのに重要となる。また、抽出された特徴は、映像のフレーム同士の視覚的な差を効率的に計算するのに役立つ。
【００５６】
本発明では、図５に示すように、キーフレームレベルの重要度点数を予測するためのアテンション基盤の映像要約ネットワークを提案する。アテンション基盤要約ネットワークは、エンコーダネットワーク５１０、デコーダネットワーク５２０、および２つのネットワークの間のアテンション層（Ａｔｔｅｎｔｉｏｎ Ｌａｙｅｒ）で構成される。このネットワークは、ＳＵＭ－ＧＡＮＡＡＥ方法に対して提案されたアテンションオートエンコーダ（非特許文献４）でＬＳＴＭネットワークを拡張繰り返しニューラルネットワーク（Ｄｉｌａｔｅｄ ＲＮＮ）に替え、ＧＲＵ（Ｇａｔｅｄ Ｒｅｃｕｒｒｅｎｔ Ｕｎｉｔ）ネットワークをデコーダに替えることでアテンションオートエンコーダを改善する。拡張ＲＮＮは、より少ないパラメータと計算効率性の向上のために拡張スキップ連結とともに実現される。特に、ネットワークは、拡張された繰り返し層を積み、層全体にわたって拡張を幾何級数的に増加させることで複雑な時間依存性を抽出する。また、拡張ＲＮＮの層としてＧＲＵセルを使用する。エンコーダネットワークは、キーフレーム同士のローカルおよびグローバルコンテキストと視覚的類似性をキャプチャする。アテンション層では、エンコーダネットワークの出力と以前のデコーダネットワークの隠れ状態の両方を用いてアテンション加重値を計算する。アテンション加重値を計算するために、非特許文献２０で提案されたアテンションメカニズムを利用する。特に、数式（１）で説明するように、エンコーダネットワーク$Ｅ_{ｏｕｔ}$の出力とデコーダネットワーク$ｈ_{ｔ－１}$の以前の隠れ状態でアテンション加重値を計算するためにコンテンツ基盤の点数関数を使用する。ｔ＝１であるときに、デコーダネットワークｈの出力を０に設定する。$Ｗ_{ａ}$は、学習可能なパラメータであるアテンション加重値マトリックスである。
【００５７】
【数１】
$$Score(E_{out}, h_{t-1}) = E_{out}^{T} W_a h_{t-1}$$

・・・（１）
【００５８】
次に、加重値は、ソフトマックス関数に基づいて各キーフレームの確率点数に正規化される。アテンション加重値を用いてエンコーダネットワークの出力に掛けてコンテキストベクトルＣを求める。
【００５９】
デコーダネットワークＤ$_{in}$の入力のために、コンテキストベクトルと０に初期化されたデコーダネットワークの以前の出力を連結し、デコーダネットワークを学習させる。
【００６０】
この後、出力Ｄ$_{out}$を取得し、エンコーダネットワークによって学習された豊富な情報を再使用するためにＥ$_{out}$と連結することで、長いシーケンスの映像に対する性能を高める。そして、線形関数とともに、次の段階ｔ＋１でデコーダネットワークに伝達するための特徴の大きさの次元を減らす。最後に、全結合層およびシグモイド関数を用いて出力の次元を減らし、重要度点数

$$S = \{s_t\}_{t=1}^{N}$$

で出力を生成する。すなわち、重要度点数は、映像の要約としてキーフレームを選択するためのフレーム追跡確率（０．ｔｏ　１．）である。
【００６１】
図６は、本発明の一実施形態における、代表性報酬関数を説明するための図である。
【００６２】
本発明の実施形態に係る効率的な映像要約のために非特許文献６の多様性報酬関数を採択しながら、効率的なキーフレーム選択のための２つ新たな報酬を提案する。提案された報酬関数は、キーフレーム同士の視覚的類似性距離と時間的距離を考慮した時間的一貫性報酬関数と代表性報酬関数である。
【００６３】
多様性報酬関数Ｒ$_{div}$数式（２）は、抽出された特徴とともにフレーム選択作業によって選択されたキーフレーム同士の差を計算する。この報酬関数に基づき、ネットワークは、要約のキーフレームから多様なフレームを選択するための重要度点数を予測するように学習される。このようなキーフレームで構成された要約は、映像の内容を容易に把握できるようにする。映像のストーリーラインを維持しながら演算の複雑性を減らすために、選択されたキーフレーム同士の差を計算するための時間距離を２０に制限する。この制限がなければ、フラッシュバック場面や類似する場面が選択したキーフレームと遠く離れていても、多様なフレームを選択するときにこのような場面を無視することができる。
【００６４】
選択されたキーフレームの指数を

$$\mathcal{I} = \{i_k | a_{i_k} = 1,\ k = 1,2,\dots,|\mathcal{I}|\}$$

とする。多様性報酬関数は次のとおりとなる。
【００６５】
【数２】

$$R_{div} = \frac{1}{\mathcal{I}(\mathcal{I}-1)} \sum_{t \in \mathcal{I}} \sum_{\substack{t' \in \mathcal{I} \\ t \neq t'}} \left(1 - \frac{x_t^T x_{t'}}{\|x_t\|_2 \|x_{t'}\|_2}\right)$$

・・・（２）
【００６６】
代表性報酬関数Ｒ$_{rep}$数式（３）は、抽出された特徴を用いて選択したキーフレームと

映像のすべてのキーフレームの類似性を計算する。この報酬関数に基づいて、ネットワークは、映像を示す要約のキーフレームを選択するための重要度点数を予測するように学習される。キーフレームで構成された要約は、映像の主題を容易に把握できるようにする。本発明では、映像の要約を適切に生成して性能を高めるために、数式（３）で示すＤ×ｓｔのように、代表性報酬関数として重要度点数Ｓを適用する新たな技法を提示する。ネットワークを学習させるためには、代表性報酬関数を増やし、報酬関数で距離Ｄ数式（４）を最小化しなければならない。

【００６７】

【数３】

$$R_{rep} = exp\left(-\frac{1}{N}\sum_{t=1}^{N} D_t \times s_t\right)$$

・・・（３）

【００６８】

【数４】

$$D_t = \min_{t' \in J} \|x_t - x_{t'}\|_2$$

・・・（４）

【００６９】

図６を参照すると、Ａは、選択されたキーフレームとすべてのキーフレームとの距離が長い例であり、Ｂは、選択されたキーフレームとすべてのキーフレームとの距離が短い例である。キーフレームの重要度点数Ｓが高ければ、要約としてキーフレームを選択することができる。この反面、キーフレームの重要度点数が低ければ、要約としてキーフレームが選択される確率は低くなる。ＡとＢはすべてＤ×Ｓの平均値であり、図６に示すように

$\overline{A}$

は

$\overline{B}$

よりも高いため、提案された報酬関数を基準にＢよりもＡに対する報酬がより低い。

【００７０】

重要度点数を使用するトリックの効果を説明する３つの事例がある。

【００７１】

１．キーフレームの距離Ｄが短い場合、重要度点数が変更しても報酬関数が高い報酬を返還する。

【００７２】

２．キーフレームの距離Ｄが長くてキーフレームの重要度点数Ｓが高ければ、報酬関数が低い報酬を返還する。また、低い報酬でネットワークを学習させた後、選択を阻むために、ネットワークはキーフレームの重要度が低い点数を予測する。

【００７３】

３．キーフレームの距離Ｄが長くてキーフレームの重要度点数Ｓが低ければ、報酬関数が中間報酬を返還する。しかし、キーフレームのほとんどは重要度点数が低いため、要約に選定されないであろう。

【００７４】

図７は、本発明の一実施形態における、時間的一貫性報酬関数を説明するための図である。

【００７５】

代表的なショットレベルキーフレームを効率的かつ均一に選択するために、時間的一貫性報酬関数Ｒｃｏｎ数式（６）が提案される。

【００７６】
図７に示すように、先ずは
$V_i^{summary}$

の
$V_i^{3'}$

のような選択されたキーフレームと
V_i^{all}

の他のフレームとの類似性を計算した後、自身を除いて最も類似するフレーム
V_i^3

を選択する。
【００７７】
時間的一貫性報酬について説明するために、図７のＢに含まれたキーフレームのような代表的なキーフレームが周辺フレームの周りに類似する場面をもつと定義する。この反面、図７のＡに含まれたキーフレームは、選択したキーフレームと一時的に距離が遠い類似の場面をもっている。演出者が意図したストーリーラインにより、Ａに含まれたキーフレームと類似する場面は少し後に登場することもあるし、一回だけ示されることもある。しかし、Ａに含まれたキーフレームの問題は、これらで生成された要約が映像コンテンツに対する使用者の理解を妨げるという点にある。したがって、要約からこのキーフレームを取り除く。このようなキーフレームを取り除くことによる他の長所は、一側だけで過度に選択される問題を防ぐことができるという点にある。すなわち、キーフレームを均一に選択することができるのである。本発明では、すべてのキーフレームの数
$|\mathcal{J}|$

まで、すべてのキーフレーム
$\{x_t\}_{t=1}^{N}$

から選択したキーフレーム
j_k

のうちから最も近い隣りを見つけ出す過程を繰り返す。本発明では、
i_k

と
j_k

との距離を最小化することによって要約の時間的一貫性を学習することができる。距離を最小化するために、報酬関数を次のように計算する。
【００７８】
【数５】
$$j_k = \left(\underset{k\in|\mathcal{J}|}{argmin}\left\| x_{i_k} - x \right\|\right)$$

・・・（５）
【００７９】
【数６】

$$R_{con} = \frac{1}{\log\left(\sum_{k=1}^{J}(i_k - j_k)^2/|\mathcal{J}|\right)}$$

・・・（６）
【００８０】
このような報酬を正規化するために、距離を
$|\mathcal{J}|$

で割ってログ確率を使用する。
【００８１】
キーフレームを要約として選択するために、重要度点数Ｓをフレーム追跡動作
$A = \{a_t | a_t \in \{0,1\}, t = 1,..,N\}$

に変換するための離散確率分布であるベルヌーイ分布を使用する。フレームのフレーム選択動作が１であれば、このキーフレームを要約として選択する。ベルヌーイ分布は、フレーム選択動作の変形をランダムに生成するため、映像の多様な要約に対する探索を促進する。
【００８２】
【数７】

$$A \sim Bernoulli(a_t; s_t) = \begin{cases} s_t, & for\ a_t = 1 \\ 1 - s_t, & for\ a_t = 0 \end{cases}$$

・・・（７）
【００８３】
また、生成された要約の品質を報酬の合計で評価する。この報酬により、アテンション基盤の映像要約ネットワークを、方策勾配方法を用いてパラメータ化された方策で学習する。方策勾配は、勾配降下アルゴリズムを用いてより効率的な要約を得るために行動戦略を探索するための強化学習方法のうちの１つである。行動戦略の不足を防ぐために、過小評価報酬（Ｕｎｄｅｒ－ａｐｐｒｅｃｉａｔｅｄ　Ｒｅｗａｒｄ：ＵＲＥＸ）方法を探索する探索戦略の目的関数を使用する（非特許文献２１）。方策下で動作
$\pi_\theta(\boldsymbol{a_t} \mid \boldsymbol{h_t})$

のログ確率
$\log \pi_\theta(\boldsymbol{a_t} \mid \boldsymbol{h_t})$

が報酬
$r(\boldsymbol{a_t} \mid \boldsymbol{h_t}) = \frac{R_{rep}+R_{div}}{2} + R_{con}$

を過小評価する場合、該当の動作は探索戦略によってより探求されるであろう。
【００８４】
$\mathrm{UREX}\,\mathcal{O}_{UREX}$

の目的関数を計算するために、エピソード
$\mathcal{J}$

に動作と報酬のログ確率を維持する。
$\mathcal{O}_{UREX}$

は、ＲＡＭＬ（Ｒｅｗａｒｄ Ａｕｇｍｅｎｔｅｄ Ｍａｘｉｍｕｍ Ｌｉｋｅｌｉｈｏｏｄ）目的関数の合計である報酬
$R(\boldsymbol{a}_t \mid \boldsymbol{h}_t)$

の期待値である。本発明では、ＲＡＭＬ目標関数の近似値のために、ソフトマックス関数を用いて各エピソードｊに対する報酬の正規化された重要度加重値集合を計算する。

【００８５】
【数８】

$$\mathcal{O}_{UREX}(\boldsymbol{\theta};\tau) = \mathbb{E}_{h\sim p(h_t)}\left\{\sum_{\boldsymbol{a}\in\mathcal{A}} R(\boldsymbol{a}_t \mid \boldsymbol{h}_t)\right\}$$

・・・（８）

【００８６】
本発明では、勾配推定値の分散を減らして計算効率性を高めるために、方策勾配に重要な技術である基準を用いる。基準は、今まで経験した報酬の移動平均で計算される。多様な映像の移動平均を用いて多様性を改善するために、以下のように各映像b_1に対する基準とすべての映像b_2に対する基準を合算して基準を計算する。最後に、ネットワークを学習させるための費用としてL_{rwd}を極大化する。

【００８７】
【数９】

$$\mathcal{B} = 0.7 \times b_1 + 0.3 \times b_2$$

・・・（９）

【００８８】
【数１０】

$$L_{rwd} = \mathcal{O}_{UREX}(\boldsymbol{\theta};\tau) - \mathcal{B}$$

・・・（１０）

【００８９】
本発明では、重要度点数を用いてキーフレームを選択する確率を制御するために、非特許文献６で提案された正規化タームL_{reg}を使用する。重要度点数のほとんどが１か０に近ければ、誤ったキーフレームを要約して選択する確率が高まる。したがって、学習中に重要度点数を０．５に近くするときにL_{reg}を用いる。重要度点数が０．５に早く収斂されることを防ぐために、これに０．０１を掛ける。

【００９０】
【数１１】

$$L_{reg} = 0.01 \times \left(\frac{1}{N} \times \sum_{1}^{N} s_t - 0.5\right)^2$$

・・・（１１）
【００９１】
すべての損失関数を計算した後、映像要約Ｌｓｕｍｍａｒｙに対する最終損失を計算して逆伝播を実行する。
【００９２】
【数１２】

$$L_{summary} = L_{reg} - L_{rwd}$$

・・・（１２）
【００９３】
アルゴリズム１は、方策勾配方法を用いた教師なし映像要約ネットワークの学習手順に関するものである。
【００９４】

Algorithm 1 Training the Network.

1: Input: x_t frame-level features of the video
2: Output: proposed network's parameters (θ)
3:
4: for the number of iterations do
5: $S \leftarrow$ Network(x_t) % Predict the importance score S
6: $A \leftarrow$ Bernoulli Distribution(S) % Action A from the score S
7: % Calculate the reward functions and the loss using A and S
8: $\{\theta\} \overset{+}{\leftarrow} -\nabla(L_{reg} - L_{rwd})$ % Minimization
9: % Update the network using the policy gradient method:
10: end for

【００９５】
本発明の実施形態によると、アテンション基盤の映像要約ネットワークをテストするために、ショット内のフレームレベルの重要度点数を平均化してショットレベルの重要度点数を計算する。他の方法の性能と比べるためには、映像のショットレベルの重要度点数が必要となる。ショットを感知するために、ショットの境界のような変更地点を感知するＫＴＳ（Ｋｅｒｎｅｌ Ｔｅｍｐｏｒａｌ Ｓｅｇｍｅｎｔａｔｉｏｎ）方法を使用する（非特許文献２２）。要約を生成するために、映像の長さの上位１５％に該当する主要ショットを点数で整列して選択する。この段階は、非特許文献６で説明した要約映像の重要性を最大化するための‘０－１ Ｋｎａｐｓａｃｋ問題と同じ概念である。
【００９６】
上述した装置は、ハードウェア構成要素、ソフトウェア構成要素、および／またはハードウェア構成要素とソフトウェア構成要素との組み合わせによって実現されてよい。例えば、実施形態で説明された装置および構成要素は、例えば、プロセッサ、コントローラ、ＡＬＵ（ａｒｉｔｈｍｅｔｉｃ ｌｏｇｉｃ ｕｎｉｔ）、デジタル信号プロセッサ、マイクロコンピュータ、ＦＰＡ（ｆｉｅｌｄ ｐｒｏｇｒａｍｍａｂｌｅ ａｒｒａｙ）、ＰＬＵ（ｐｒｏｇｒａｍｍａｂｌｅ ｌｏｇｉｃ ｕｎｉｔ）、マイクロプロセッサ、または命令を実行して応答することができる様々な装置のように、１つ以上の汎用コンピュータまたは特殊目的コンピュータを利用して実現されてよい。処理装置は、オペレーティングシステム（ＯＳ）およびＯＳ上で実行される１つ以上のソフトウェアアプリケーションを実行してよい。また、処理装置は、ソフトウェアの実行に応答し、データにアクセスし、データを記録、操作、処理、および生成してもよい。理解の便宜のために、１つの処理装置が使用されるとして説明される場合もあるが、当業者であれば、処理装置が複数個の処理要素および／または複数種類の処理要素を含んでもよいことが理解できるであろう。例え

ば、処理装置は、複数個のプロセッサまたは１つのプロセッサおよび１つのコントローラを含んでよい。また、並列プロセッサのような、他の処理構成も可能である。

【００９７】

ソフトウェアは、コンピュータプログラム、コード、命令、またはこれらのうちの１つ以上の組み合わせを含んでもよく、思うままに動作するように処理装置を構成したり、独立的または集合的に処理装置に命令したりしてよい。ソフトウェアおよび／またはデータは、処理装置に基づいて解釈されたり、処理装置に命令またはデータを提供したりするために、いかなる種類の機械、コンポーネント、物理装置、仮想装置（ｖｉｒｔｕａｌ ｅｑｕｉｐｍｅｔｎ）、コンピュータ記録媒体または装置に具現化されてよい。ソフトウェアは、ネットワークによって接続されたコンピュータシステム上に分散され、分散された状態で記録されても実行されてもよい。ソフトウェアおよびデータは、１つ以上のコンピュータ読み取り可能な記録媒体に記録されてよい。

【００９８】

実施形態に係る方法は、多様なコンピュータ手段によって実行可能なプログラム命令の形態で実現されてコンピュータ読み取り可能な媒体に記録されてよい。前記コンピュータ読み取り可能な媒体は、プログラム命令、データファイル、データ構造などを単独でまたは組み合わせて含んでよい。前記媒体に記録されるプログラム命令は、実施形態のために特別に設計されて構成されたものであってもよいし、コンピュータソフトウェアの当業者に公知な使用可能なものであってもよい。コンピュータ読み取り可能な記録媒体の例としては、ハードディスク、フロッピー（登録商標）ディスク、および磁気テープのような磁気媒体、ＣＤ－ＲＯＭ、ＤＶＤのような光媒体、フロプティカルディスク（ｆｌｏｐｔｉｃａｌ ｄｉｓｋ）のような光磁気媒体、およびＲＯＭ、ＲＡＭ、フラッシュメモリなどのようなプログラム命令を格納して実行するように特別に構成されたハードウェア装置が含まれる。プログラム命令の例は、コンパイラによって生成されるもののような機械語コードだけではなく、インタプリタなどを使用してコンピュータによって実行される高級言語コードを含む。

【００９９】

以上のように、実施形態を、限定された実施形態および図面に基づいて説明したが、当業者であれば、上述した記載から多様な修正および変形が可能であろう。例えば、説明された技術が、説明された方法とは異なる順序で実行されたり、かつ／あるいは、説明されたシステム、構造、装置、回路などの構成要素が、説明された方法とは異なる形態で結合されたりまたは組み合わされたり、他の構成要素または均等物によって対置されたり置換されたとしても、適切な結果を達成することができる。

【０１００】

したがって、異なる実施形態であっても、特許請求の範囲と均等なものであれば、添付される特許請求の範囲に属する。

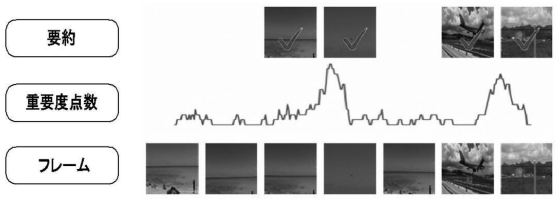
【符号の説明】

【０１０１】

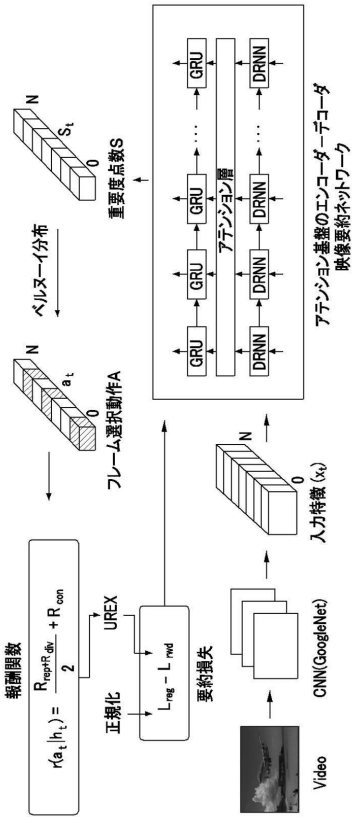
３１０：フレームレベル映像特徴抽出モジュール
３２０：映像要約ネットワークモジュール
３３０：評価モジュール
３４０：方策勾配アルゴリズム基盤の学習モジュール
３５０：映像要約生成モジュール

【図面】

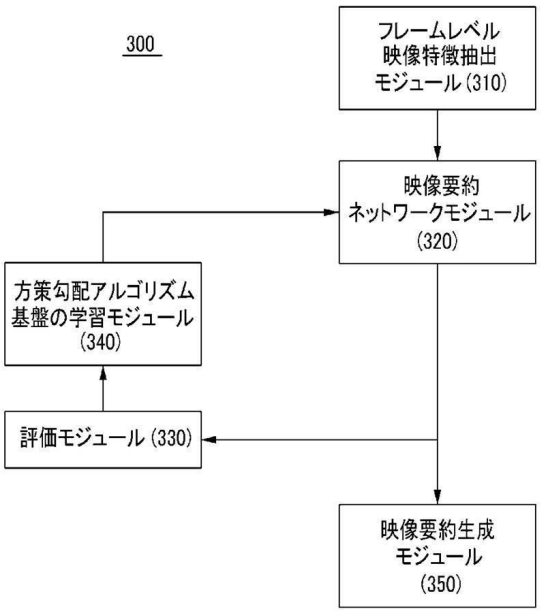
【図１】


要約
重要度点数
フレーム

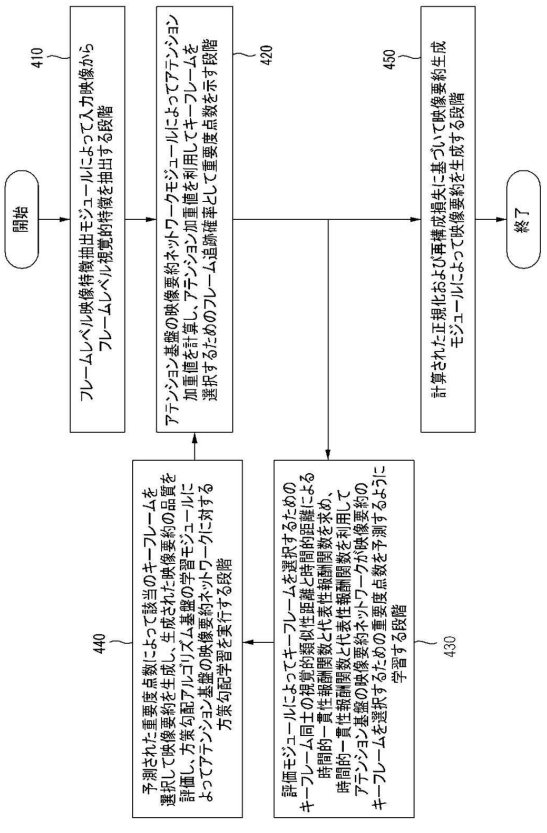
【図２】


報酬関数
$r(a_t|h_t) = \frac{R_{repr}+R_{div}}{2} + R_{con}$
正規化
UREX
$L_{reg} - L_{rwd}$
要約損失
CNN(GoogleNet)
Video
入力特徴 (x_t)
フレーム選択動作A
ベルヌーイ分布
重要度点数S
アテンション基盤のエンコーダーデコーダ
映像要約ネットワーク
アテンション層
GRU
DRNN

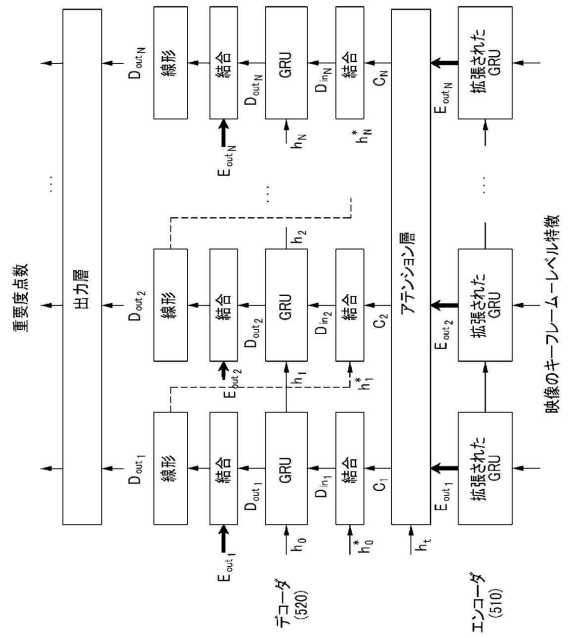
【図３】


300
フレームレベル
映像特徴抽出
モジュール(310)
映像要約
ネットワークモジュール
(320)
方策勾配アルゴリズム
基盤の学習モジュール
(340)
評価モジュール(330)
映像要約生成
モジュール
(350)

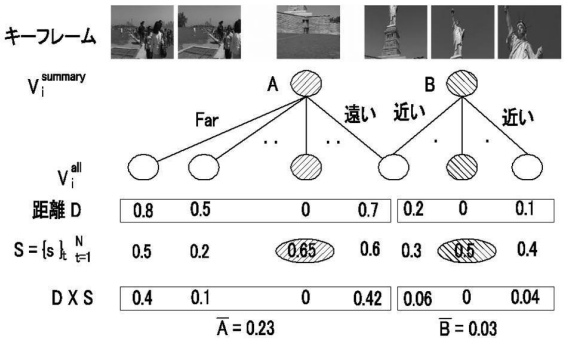
【図４】


開始
410
フレームレベル映像特徴抽出モジュールによって入力映像から
フレームレベル視覚的特徴を抽出する段階
420
アテンション基盤の映像要約ネットワークモジュールによってアテンション
加重値を計算し、アテンション加重値を利用してキーフレームを
選択するためのフレーム追跡確率として重要度点数を示す段階
430
計算された正規化および再構成損失に基づいて映像要約生成
モジュールによって映像要約を生成する段階
終了
440
予測された重要度点数によって該当のキーフレームを
選択して映像要約を生成し、生成された映像要約の品質を
評価し、方策勾配アルゴリズム基盤の学習モジュールに
よってアテンション基盤の映像要約ネットワークに対する
方策勾配学習を実行する段階
430
評価モジュールによってキーフレームを選択するための
キーフレーム同士の視覚的類似性距離と時間的距離による
時間的一貫性報酬関数と代表性報酬関数を求め、
時間的一貫性報酬関数と代表性報酬関数を利用して
アテンション基盤の映像要約ネットワークが映像要約の
キーフレームを選択するための重要度点数を予測するように
学習する段階

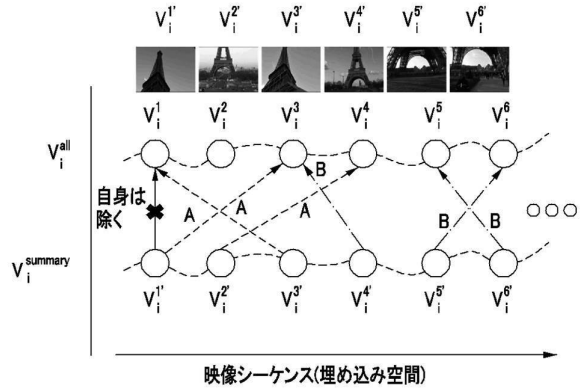
【図５】


重要度点数
出力層
線形
結合
GRU
アテンション層
拡張されたGRU
映像のキーフレームレベル特徴
デコーダ(520)
エンコーダ(510)

【図６】


キーフレーム
Far
遠い
近い
近い
距離 D
0.8 0.5 0 0.7 0.2 0 0.1
0.5 0.2 0.65 0.6 0.3 0.5 0.4
D X S
0.4 0.1 0 0.42 0.06 0 0.04
A = 0.23
B = 0.03

【図７】


自身は除く
映像シーケンス(埋め込み空間)

フロントページの続き

(72)発明者　ホン・ミョンドク
大韓民国 ２２２１２ インチョン ミチョルーグ インハーロ １００

審査官　鈴木 隆夫

(56)参考文献　特開２０２２－１６９００９（ＪＰ，Ａ）
特開２０２１－０６０８７４（ＪＰ，Ａ）
中国特許出願公開第１０９８８５７２８（ＣＮ，Ａ）
Zhong JI et al.，“ Deep Attentive Video Summarization With Distribution Consistency Learning ”，IEEE Transactions on Neural Networks and Learning Systems，Vol. 32，No. 4，2020年05月11日，p.1765-1775，DOI: 10.1109/TNNLS.2020.2991083

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｈ０４Ｎ　２１／００－２１／８５８
Ｇ０６Ｖ　１０／８２
Ｇ０６Ｎ　３／０８
Ｇ０６Ｔ　７／００
Ｇ０６Ｎ　３／０４５