キー画像によるニュース映像アーカイブの意味的構造に基づく可視化

孟　洋† 山岸　史典†† 井手　一郎††† 片山　紀生† 佐藤　真一†
坂内　正夫†

† 情報・システム研究機構国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
†† 東京大学大学院情報理工学系研究科　〒 101–8430 東京都千代田区一ツ橋 2–1–2
††† 名古屋大学大学院情報科学研究科　〒 464–8601 名古屋市千種区不老町 1
E-mail: †{mo,katayama,satoh,sakauchi}@nii.ac.jp, ††fuminori@ nii.ac.jp, †††ide@is.nagoya-u.ac.jp

あらまし　近年の蓄積装置の大容量化、低価格化は、ホームビデオサーバを現実のものとし、オンデマンドな映像視聴をはじめ、映像情報の幅広い活用を可能にしつつある。蓄積された映像情報への高度なアクセスを実現するためには、意味内容に基づき映像情報を構造化することが重要である。映像情報は、画像、音声、文字からなるマルチメディア情報であるため、各情報を用いることで、異なる視点、あるいは複合的な視点から映像情報の構造化が可能である。本研究では、ニュース映像アーカイブの可視化を目指し、画像と文字の両視点から意味構造の解析を行うことで、映像内容を説明するキー画像の取得を試みたので報告する。

キーワード　映像アーカイブ，映像解析，意味構造，ニューストピック，キー画像

Key Image Extraction Based on the Semantic Structure of a News Video Archive

Hiroshi MO†, Fuminori YAMAGISHI††, Ichiro IDE†††, Norio KATAYAMA†, Shin'ichi SATOH†, and Masao SAKAUCHI†

† National Institute of Informatics, Research Organization of Information and Systems
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
†† Graduate School of Information Science and Technology, The University of Tokyo
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
††† Graduate School of Information Science, Nagoya University
1 Furo-cho, Chikusa-ku, Nagoya, 464–8601, Japan
E-mail: †{mo,katayama,satoh,sakauchi}@nii.ac.jp, ††fuminori@ nii.ac.jp, †††ide@is.nagoya-u.ac.jp

Abstract　Recently, it has become possible to handle a large amount of video data with video archive system. It is very important that a video data is structured based on semantics for useful access to large video archive. The video data is consisted of images, sounds and texts. Therefore video data can be structured by using its information in multi-modality. In this paper, we introduce a method for key image extraction based on the semantic structure of news video archive.

Key words　Video Archive, Video Analysis, Semantic Struture, News Topic, Key Image

1. はじめに

近年の情報通信技術や情報蓄積技術の急速な発展は、放送された様々な映像情報を長期間にわたって蓄積することを現実のものとし、蓄積された大量の映像情報をオンデマンドで利用することを可能にしつつある。蓄積された大量の映像情報を有効に活用するためには、映像情報の内容解析によるインデキシングの実現はもちろんのこと、意味的な分割や関連付けを行う知的構造化、明示されていない情報を抽出するデータマイニング技術などに基づき、高度な映像アクセス機能を実現することが強く求められる。

現在までに、映像情報の意味構造の抽出に対しては様々な手

法が検討されてきているが、その多くはショットやシーンの分割、同定を中心とした単一映像内の構造化にとどまるものであった [1], [2]。しかし、大量に蓄積された映像情報の構造化という視点で考えてみると、高次な意味情報の抽出において、複数映像間の関連性の抽出が重要な意味を持つようになってくる。特に、ニュース映像のように、同一あるいは関連する話題が継続的に取り上げられるものについては、話題の分岐や連続性の判別など、複数映像間の構造解析が必須となる。

また、従来、映像情報の構造化は、画像情報あるいは文字情報を利用するなど、多くの場合、単一情報の側面から検討されてきた [3]。しかし、映像情報は、画像、音声、文字からなるマルチメディア情報であるため、各情報を用いることで、相補的に、異なる視点、あるいは複合的な視点からの映像情報の構造化が可能である。

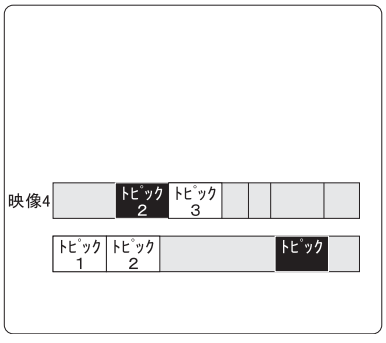
そこで、本研究では、ニュース映像アーカイブの可視化を目指し、画像と文字の両視点から、単一映像内、複数映像間の意味構造の解析を行うことで、ニュース映像を代表するキー画像とともにトピック構造の抽出を試み、トピック内容を説明するキー画像の取得手法について検討を行ったので報告する。具体的には、ニュース映像アーカイブにおいて繰り返し利用されるショットに基づきキー画像の取得を、文字放送字幕テキストの文間の関連性に基づきトピック構造の抽出を行った。

2. ニュース映像アーカイブにおける意味構造

2.1 文字情報に基づき抽出可能な構造

映像情報の構造としては、映像内と映像間の大きく二つの関係構造を考えることができる。ニュース映像アーカイブの場合、トピック（話題）を単位として意味的に構造化される。図 1 にトピックに基づく映像内・映像間の構造の例を示す。

日々放送されるニュース映像は、複数の話題から構成され、同一あるいは関連する話題を継続的に取り上げているという特徴を持っている。このため、複数の映像情報の間には、話題の分岐や連続性など、単一の映像情報からは得られない話題のつながりの構造を考えることができる。



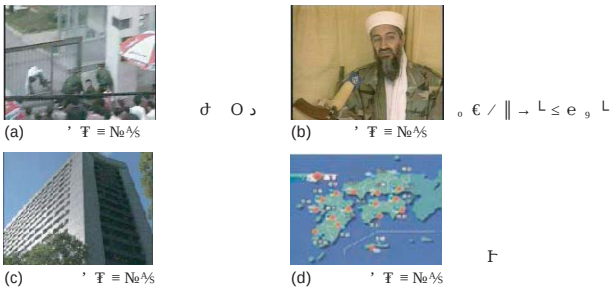


図 3　繰り返し利用される映像ショットの例

なお、番組内容を区切る定型的なショットを利用すれば、例えば、オープニング、ニューストピック、スポーツ、天気予報、エンディングなどといった番組構造を画像的に抜き出すことも可能と考えられる。

3.　映像アーカイブからのキー画像の取得

3.1　意味構造に基づくキー画像の取得の枠組

2. で述べたように、ニュース映像アーカイブの話題としての意味構造は、話題の関連性に基づき構成されるスレッドによりあらわすことができる。一方、重要な映像、貴重な映像、あるいは資料的な映像は、繰り返し利用される傾向にあるため、ニュース映像アーカイブ内で複数回出現するショットの画像、つまり画像的な意味構造として求めることができる同一映像ショット群を構成するショットの画像は、トピック内容を代表するキー画像となり得る。

本研究では、文字情報を解析することでスレッドを、画像情報を解析することで同一映像ショット群を抽出し、あるトピックから派生するスレッドを構成するトピック群（トピッククラス）に含まれる同一映像ショット群を求めることで、トピック内容を説明するキー画像の取得を試みる。スレッドと同一映像ショットに基づくキー画像の抽出の枠組を図 4 に示す。

なお、このように取得されたキー画像は、トピック内容を画像で提示するなど、映像アーカイブの可視化に利用することができる。

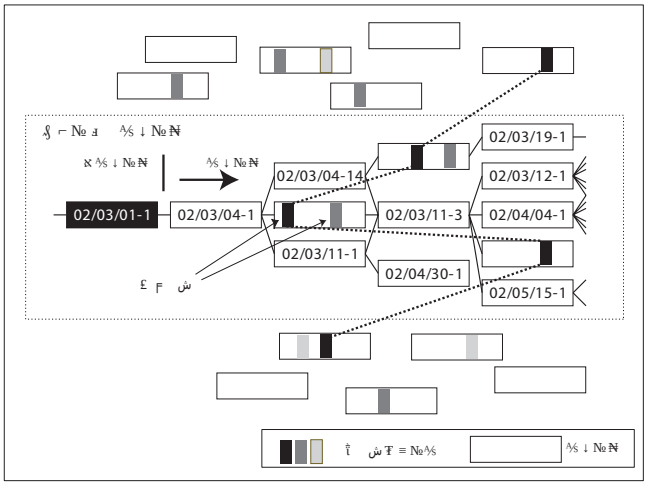


図 4　スレッドと同一映像ショットに基づくキー画像の抽出の枠組

3.2　文字情報に基づくスレッドの抽出

ニュース映像アーカイブにおけるスレッド（トピック構造）の抽出は、(1) トピック分割、(2) トピック追跡・スレッド構築という手順 [5] で行う。

トピック分割は、ニュース映像とともに放送されている文字放送字幕テキストにおける文間のキーワードベクトルの類似性に基づいて行う。また、トピック追跡とスレッド構築は、分割されたトピック間のキーワードベクトルの類似性をもとに、意味的関連性、時間的関連性を考慮して行う。

3.2.1　トピック分割

トピック分割は、文字放送字幕テキストに対し、下記の処理を適用して行う。

（1）　文字放送字幕テキストの形態素解析 [6] を行い、各文から名詞列を抽出する。

（2）　名詞列の語義属性（一般，人物，場所・組織，時相）の解析 [7] を行い、文単位で属性別の出現頻度付きキーワードベクトル $(\vec{k_g}, \vec{k_p}, \vec{k_l}, \vec{k_t})$ を作成する。

（3）　各文の境界で、前後 w 文を結合したキーワードベクトル間の類似度を評価する。文 i と $i+1$ の間のにおける類似度は、次式のように定義する。

$$R_{S,w}(i) = \frac{\sum_{m=i-w+1}^{i} \vec{k_S}(m) \cdot \sum_{n=i+1}^{i+w} \vec{k_S}(n)}{\left|\sum_{m=i-w+1}^{i} \vec{k_S}(m)\right| \left|\sum_{n=i+1}^{i+w} \vec{k_S}(n)\right|} \quad (1)$$
$$(i = w, w+1, ..., i_{max} - w)$$

ここで、$S = \{g, p, l, t\}$ であり、i_{max} は、1 番組中の文字放送字幕テキストの文数である。実験では、$w = 1, 2, ..., 10$ とした。

（4）　各語義属性別の類似度を重み付き和の形で統合し、閾値 θ_{seg} を下回る場合に、文 i と $i+1$ の間をトピックの境界として検出する。

$$R(i) = \sum_{S=\{g,p,l,t\}} a_S \max_w R_{S,w}(i) \quad (2)$$

実験では、各重みを、$(a_g, a_p, a_l, a_t) = (0.23, 0.21, 0.48, 0.08)$ とした。これは訓練データを用いて重回帰分析により決定した。

（5）　トピック毎にキーワードベクトル $\vec{K_S}$ を作成し、次式で定義される隣接するトピック間の類似度が、閾値 θ_{cat} を上回る場合に、トピック i と j を結合する。

$$R(i,j) = \sum_{S=\{g,p,l,t\}} a_S \frac{\vec{K_S}(i) \cdot \vec{K_S}(j)}{\left|\vec{K_S}(i)\right| \left|\vec{K_S}(j)\right|} \quad (3)$$

予備実験の結果から、各閾値は、$\theta_{seg} = 0.28, \theta_{cat} = 0.08$ とした。

3.2.2　トピック追跡・スレッド構築

トピックの追跡は、式 (3) を用いて、分割された全てのトピック相互間の類似度を評価して行う。トピック i と j の類似度 $R(i,j)$ がある閾値 θ_{trk} を超えた場合、これらのトピックは関連するとみなすことにし、以下の手順を再帰的に実行することで、関連するトピックを階層的に構造化し、スレッドを構築する。

（1） あるトピック（親）に関連する全てのトピック（子）を関連付ける。

（2） ある子トピックに対して、兄弟や子孫で関連するトピックがあれば、時系列的に最も近いものの子として関連付けをし直す。

スレッド構築の目的は、関連したトピックを時系列に連鎖させることで、話題の分岐や連続性を構造化し、特定のニューストピックに関する情報を抽出しやすくするのと同時に、利用者が容易に理解できるように話題の流れを示すことにある。特に、大量のニュース映像が蓄積された大規模なニュース映像アーカイブを用いる場合、興味のあるニューストピックや、その話題の流れを追うことは大きな負担となるため、スレッドの構築は重要な問題となる。

3.3 画像情報に基づく同一映像ショットの抽出

ニュース映像アーカイブ内の異なる映像ソースにおいて、2ヶ所以上存在する同一のショットを同一映像ショットと呼ぶ。同一映像ショットの抽出は、画像的にほとんど同一のフレームを検出 [8] し、輝度情報を用いたショット分割の結果とあわせ、それらをショット単位へ統合、変換することにより実行する。

ショットとしては、画像的にほとんど変化がない区間を対象とし、二つのショット内に1フレームでも同一のものが見つかった場合には、同一映像ショットとする。ここでは、画像的に“ 類似 ”ではなく“ 同一 ”のフレームの検出を目的としている。“ 同一 ”の判定には、全画面での精密な照合を行う必要があるため、フレーム間の正規化相互相関 (Normalized Cross Correlation; NCC) を用いることにした。比較する2枚の画像間の NCC は、次式により定義される。

$$NCC(a,b) = \frac{\frac{1}{n}\sum_{i}^{n}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\frac{1}{n}\sum_{i}^{n}(a_i-\bar{a})^2}\sqrt{\frac{1}{n}\sum_{i}^{n}(b_i-\bar{b})^2}} \quad (4)$$

ここで、a,b は画像、a_i,b_i は画像を構成する n 個の画素の輝度、$\bar{a}$,$\bar{b}$ は各画像の画素の輝度の平均値である。

NCC は、-1 から 1 までの値をとり、全く同一の画像同士では $NCC = 1$ となる。本研究では、相関値がきわめて 1 に近い場合 (0.9 以上) を“ 同一 ”であると判定する。実際の映像で検証した結果、“ 同一 ”であると想定した画像対は、ほとんど過不足なく検出できることを確認している。

3.4 トピックを代表するキー画像の取得

同一の映像ショットは、映像情報上、ある重要な意味を持つために、ニュース映像アーカイブ内で複数回取り上げられている。このため、同一映像ショットは、ニュース映像アーカイブ内のニュース映像を代表する映像ショットであると考えることができる。一方、ニュース映像アーカイブにおいて、各ニューストピックの内容、及びその展開は、その意味的、時間的な関連性から、複数のトピックが分岐、合流するスレッドと呼ばれるトピック構造によりあらわすことができる。このため、各トピックは、そのトピックのみならず、そのトピックから派生するスレッドを構成する複数のトピックの内容をも意味していると考えることができ、各トピックは、そのトピック自身とスレッドにおいて関連付けられた複数のトピックからなるトピッククラスとしてあらわすことが可能である。以上のことを踏まえ、各トピックの内容を代表するキー画像を、そのトピックから派生するトピッククラス内に含まれる同一映像ショットの画像として取得する。

同一映像ショットの種類を $IVS_i(i=1,\ldots,I)$、同一映像ショット IVS_i に属する各ショットを $VS_{ij}(j=1,\ldots,J_i)$ とする。また、トピッククラスの種類を $TPC_p(p=1,\ldots,P)$、トピッククラス TPC_p に属する各トピックを $TP_{pq}(q=1,\ldots,Q_p)$ とする。同一映像ショット IVS_i は2つ以上のショットから、トピッククラス TPC_p は1つ以上のトピックから構成される。このとき、トピックの内容を代表するキー画像は、トピッククラス TPC_p に属する各トピック TP_{pq} と時間的に重なりを持つ、同一映像ショット IVS_i に属する各ショット VS_{ij} の画像として求めることができる。

同一映像ショットには、重要なショット、貴重なショット、資料的なショット、定型的なショットなどがある。重要なショットや貴重なショットはトピックの内容に密接に関係するショットと考えられるが、資料的なショットや定型的なショットは強い関係がないショットである可能性が高い。そこで、得られたキー画像がどの程度トピックの内容を代表するかを示す指標として、トピッククラス TPC_p における同一映像ショット IVS_i の重要度 $S(p,i)$ を次式のように定義する。

$$S(p,i) = M_{ip}\log(N_{IVS}/J_i) \quad (5)$$

ここで、$M_{ip}(M_{ip} \leqq J_i)$ は、同一映像ショット IVS_i に属する各ショット VS_{ij} のうちトピッククラス TPC_p に属するショット数、N_{IVS} は、映像アーカイブ内の同一映像ショットの総数、J_i は、同一映像ショット IVS_i に属するショットの数である。

なお、映像アーカイブ内の同一映像ショットの総数 N_{IVS} は、ニュース映像アーカイブ内に存在する同一映像ショット IVS_i の種類の数が I、同一映像ショット IVS_i に属するショットの数が J_i であるので、次式で求めることができる。

$$N_{IVS} = \sum_i J_i \quad (6)$$

これは、あるトピックに対し重要なショットあるいは貴重なショットは、映像アーカイブ全体をとおして出現頻度の割合が低く、かつそのトピックにおいて出現頻度が高くなる、逆に、資料的ショットあるいは定型的なショットは、映像アーカイブ全体をとおして出現頻度の割合が高く、かつそのトピックにおいて出現頻度が低くなる、という傾向があることに基づいたものである。

各トピックに対し、同一映像ショットの画像を $S(p,i)$ の値にしたがって順位付けて表示することで、トピック内容の画像での提示が可能となる。

4. 実　　験

4.1 大規模放送映像アーカイブシステム

我々は、長期間にわたって放送された映像情報の解析を行う

ため、大規模放送映像アーカイブシステムを構築 [9] している。このシステムでは、東京地区の地上波 7 チャンネル分の放送映像を、24 時間、MPEG-1 でキャプチャすることが可能であり、約 10TB（約 1 ヶ月）分のデータを蓄積している。また、ニュース映像に関しては、特に長期的なデータの解析を行うため、毎日 19 時から NHK で放送されている「ニュース 7」という番組のデータを、2001 年より現在まで約 1000 日（約 500 時間）分蓄積している。なお、動画像とともに、主/副音声、文字放送、及び電子番組表（EPG）のデータを同時に記録しており、必要に応じてこれらの情報も利用することが可能になっている。本研究では、このシステムにより取得された映像データを用いて実験を行った。

4.2 ニューストピック閲覧インタフェース

大量に蓄積されたニュース映像を閲覧するため、ニューストピック閲覧インタフェースを構築している。このインタフェースでは、トピックの検索とともに、あるトピックに対する子トピックの一覧表示やスレッドをたどることが可能になっている。インタフェースの画面の例を図 5 に示す。このインタフェースでは、映像とともに、字幕文字放送テキストや各種のトピック情報を表示することができるが、現在、画像に関しては、先頭フレームを表示するにとどまっている。これに対して、トピック内容を代表するキー画像の表示が追加できれば、より直感的に内容を把握できるインタフェースへと拡張することが可能になると考えている。

4.3 キー画像の取得結果

2001 年 9 月 1 日から 2003 年 2 月 5 日までに NHK で放送された「ニュース 7」のニュース映像を用いてトピッククラスに対するキー画像の取得に関する予備的実験を行った。約 500 日分のデータから、オープニングやエンディングのショットを除いて、5623 種類、それぞれ 2〜46ヶ所で出現している同一映像ショットを抽出することができた。

「同時多発テロ」に関するあるニューストピック（2001 年 9 月 28 日#5）（#5 はトピック番号）に対し、派生するトピッククラス（3 階層）を代表すると判定されたキー画像、上位 12 画像を図 6 に示す。この例では、トピッククラスは 16 トピックから構成され、関係する同一映像ショット（キー画像）は 16 種類であった。キー画像としては、オサマ・ビンラディン氏の画像を中心に、タリバン、アフガニスタン、炭素菌などに絡む画像が得らた。比較のため、同じトピッククラスに属する各トピックのスタジオショットを除く最初のショットの画像の例（8 画像）を図 7 に示す。この場合、多くはレポーターなど現場の状況を説明する画像であり、トピックを推定できる画像は少ない。

また、「拉致被害者」に関するあるニューストピック（2002 年 10 月 16 日#1）に対し、派生するトピッククラス（3 階層）を代表すると判定されたキー画像、上位 12 画像を図 8 に示す。この例では、トピッククラスは 36 トピックから構成され、関係する同一映像ショット（キー画像）は 57 種類であった。キー画像としては、帰国した拉致被害者らの画像を中心に、被害者の親族、政府、北朝鮮核施設などに絡む画像が得られた。比較のため、同じトピッククラスに属する各トピックのスタジオショットを除く最初のショットの画像の例（8 画像）を図 9 に示す。この場合、拉致被害者の画像なども得られているが、画像からトピックの内容を推定するのは困難である。

同一映像ショットから得られたこれらのキー画像は、各トピックがこれらの画像があらわす話題に関連する、あるいは派生するということを画像的に示している。この二つの例をみる限りでは、単純に各トピックのショットから画像を抜き出す場合に比べ、トピックの内容を示すのに有効な画像が得られていることがわかる。

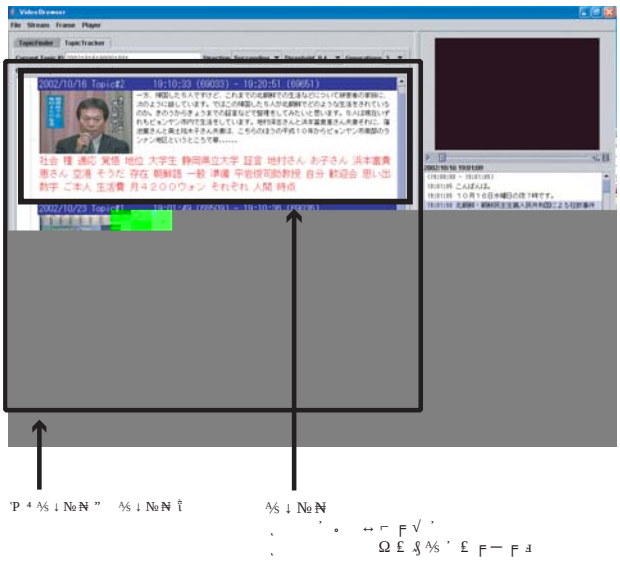


図 5　トピック閲覧インタフェース

図 6　同時多発テロに関するあるトピッククラスのキー画像

図 7　同時多発テロに関するあるトピッククラスの画像の例

図 8　拉致被害者に関するあるトピッククラスのキー画像

図 9　拉致被害者に関するあるトピッククラスの画像の例

5. おわりに

ニュース映像アーカイブへの効果的なアクセスを実現するため、画像と文字の両視点から意味構造の解析を行い、ニュース映像アーカイブの可視化を目指し、トピックに対するキー画像の取得について検討を行った。トピック、及びスレッドの抽出は文字情報を、キー画像の取得は画像情報を解析、構造化することで実現した。

「同時多発テロ」、「拉致被害者」に関するトピックに対する予備実験により、キー画像によるトピック内容の提示の可能性を示し、トピックを提示する場合の画像の選択について一つの効果を確認することができた。

今後は、画像と文字情報をより強く統合させた形で行う知的な構造化や、取得された同一映像ショットが、重要な映像なのか単なる定型的な映像なのかなどの判別、トピックの統合や分岐などトピック構造をより深く考慮した形での画像の選択、提示などについて検討していく予定である。

文　献

[1] H.D. Wactlar, M.G. Christel, Y.Gong, and A.G. Hauptmann, “Lessons learned from building a Terabyte digital video library,” *IEEE Computer*, vol.32, no.2, pp.66–73, 1999.

[2] M.G.Christel, A.G.Hauptmann, H.D.Wactler, and T.D.Ng., “Collages as dynamic summaries for news video,” *Proc. 10th ACM Intl. Conf. on Multimedia*, pp.561–569, 2002.

[3] A.Merlino, D.Morey, and M.Maybury, “Broadcast news navigation using story segmentation,” *Proc. 5th ACM Intl. Conf. on Multimedia*, pp.381–391, 1997.

[4] 山岸 史典, 佐藤 真一, “同一映像断片探索に基づくニュース映像ブラウザの実装,” 電子情報通信学会技報報告, PRMU2003, 2003.

[5] 井手 一郎, 孟 洋, 片山 紀生, 佐藤 真一, “大規模ニュース映像コーパスの意味構造解析,” 電子情報通信学会技報報告, PRMU2003-97, 2003.

[6] 黒橋 禎夫, 長尾 真, “日本語形態素解析システム JUMAN Version 3.61,” 京都大学大学院情報科学研究科, 1999.

[7] I.Ide, R.Hamada, S.Sakai, and H.Tanaka, “Semantic analysis of television news captions referring to suffixes,” *Proc. 4th Intl. Workshop on Information Retrieval with Asian Languages*, pp.37–42, 1999.

[8] 山岸 史典, 佐藤 真一, 浜田 喬, 坂内 正夫, “大規模放送映像アーカイブにおける映像断片照合の提案と高速化,” 電子情報通信学会技報報告, PRMU2002-166, 2002.

[9] 片山 紀生, 孟 洋, 佐藤 真一, “映像インデクシング研究のための大規模映像アーカイブシステムの試作,” 情報処理学会研究報告, 02-DBS-127, pp.17–23, 2002.