社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE
PRMU2002-30(2002-6)

映像処理評価用映像データベースについて

馬場口 登[1)] 栄藤 稔[2)] 佐藤 真一[3)] 安達 淳[3)] 阿久津 明人[4)] 有木 康雄[5)]
越後 富夫[6)] 柴田 正啓[7)] 全 炳東[8)] 中村 裕一[9)] 美濃 導彦[10)] 松山 隆司[10)]
1）大阪大学 2）NTT ドコモ 3）国立情報学研究所 4）NTT 5）龍谷大学
6）日本 IBM 7）NHK 8）千葉大学 9）筑波大学 10）京都大学
E-mail: 1）babaguchi@sanken.osaka-u.ac.jp

あらまし 電子情報通信学会パターン認識・メディア理解研究会の下で検討，作成した映像処理評価用映像データベース（VDB：Video Data Base）について述べる．このデータベースは編集効果（シーン切替），カメラワーク，テロップの出現，音声品質という点においてテレビ放送に匹敵する品質の素材映像をもち，ニュース，ドラマ，ドキュメンタリー，情報番組（料理，観光）などのジャンルの映像からなる．また，ショット境界やシナリオ情報を MPEG7 形式のメタデータとして付与している．各種の映像処理アルゴリズムを比較評価するためのベンチマークデータとして利用されることが期待される．
キーワード 映像データベース，映像処理，メタデータ，MPEG7

Video Database for Evaluating Video Processing

Noboru BABAGUCHI [1)] Minoru ETOH[2)] Shinichi SATOH[3)] Jun ADACHI[3)]
Akihito AKUTSU[4)] Yasuo ARIKI[5)] Tomio ECHIGO[6)] Masahiro SHIBATA[7)]
Heitou ZEN[8)] Yuichi NAKAMURA[9)] Michihiko MINOH[10)] Takashi MATSUYAMA[10)]
1)Osaka University 2)NTT Docomo 3)National Institute of Informatics 4)NTT 5)Ryukoku University
6)IBM Japan 7)NHK 8)Chiba University 9) University of Tsukuba 10) Kyoto University

Abstract This report presents a video database (VDB) for evaluating video processing which has been considered and developed by VDB-Working Group affiliated with the Pattern Recognition and Media Understanding (PRMU)-Technical Group. The quality of the VDB contents is equivalent to that of actual TV broadcasts and the contents are classified into several genres: news, drama, documentary, cooking and sight-seeing programs. In addition, metadata described by MPEG7 is provided in the VDB. We expect that it should be made full use of as benchmark data to evaluate a variety of video processing algorithms and systems.
Keyword video database, video processing, metadata, MPEG-7

1. はじめに

映像メディア（ここでは，TV 番組や映画のような画像と音声の同期的データを指す）のコンテンツ解析及び構造化，インデキシングなどは，パターン認識，メディア理解の新しい応用分野として，90 年代後半から活発な研究がなされており，PRMU 研究会でもここ数年，テーマセッションとして開催され，多くの発表が行われている．その結果，信号レベルからコンテンツレベルに至るまで様々な手法が提案され，それらを正解率などの尺度で評価している．

言うまでもなく，パターン認識や画像解析研究における重要な要素はアルゴリズムやシステムの性能評価である．しかしながら，データに言及することなく，それらの認識率や正解率を○○%であると主張することの危うさはほとんどの研究者が認識していることであろう.つまり,認識率や正解率はデータ依存であり，

データを変えれば数字（率）は容易に変動するのである．このために標準的なデータベースを設定し，その土俵で性能評価して相互の比較を行うというプロセスが必須となる．これまでに，ETL 文字データベース[1]は文字認識アルゴリズムの検証用の標準データとして大きな役割を果たし，文字認識研究のレベル向上に寄与してきた．近年では同様の目的で，画像理解評価用画像データベース[2]，顔画像データベース[3]，RWC マルチモーダルデータベース[4]，RWC ジェスチャデータベース[4]などが供出されている．

ここで映像メディア研究に目を移しても，評価やその検証の難しさという課題は依然として残っている．例えば，基本問題の一つであるショット切替の検出にしても，対象依存性が強く，映像性質の異なるジャンルの映像には，うまく動作しない（論文に掲載されている精度がとてもでない）事例も見受けられる．また一方で，論文投稿したときにしばしば指摘される従来手法との比較という点でも，標準データがなければ比較は容易でない．

現状は，研究者各々が自前のデータとして日々放送されているテレビ映像やビデオ映像を切り出して独自に性能評価を行っている．映像メディア研究自体まだ歴史が浅く，研究分野の立ち上げ期ということもあって，手法の提案が優先され，他手法との相対評価は軽視される傾向があったかもしれない．当然の帰結として，評価軸があいまいになり，手法相互の性能が客観的に評価できないことになる．映像メディア研究も充実期に向かうことを考えると，研究分野の健全な発展のために，各種アルゴリズムに正当な評価を下しうるための標準データベース（DB）を作成し，評価を積み重ねることが何より重要と言えよう．

しかしながら，映像 DB の作成には特有の問題がある．それは著作権や肖像権などの問題である．放送された TV 番組や映画，ビデオなどの素材映像を自由に使えないという点である．特に，論文書きを仕事の一つとする我々につらいのは，論文に映像やその一部である静止画を気兼ねなく掲載できないことである．以前ある国際会議でプロフットボールの映像を使ったところ，「許可を取っているのか？」と聞かれたことがあり，実のところ筆者（馬場口）も権利関係に無知であったことを白状せねばならない．欧米でも大学に在籍する研究者は，割と無頓着に学術目的なら無許可使用ＯＫとする向きもあるが，触法していることに変わりはない．

このような状況から，種々の映像解析アルゴリズムを適正かつ公平に比較するための映像 DB（VDB）を作成することを目標に，パターン認識・メディア理解（PRMU）研究会傘下の WG（本稿の著者グループがそのメンバーである）として，1999 年秋から 2002 年の春まで活動を続け，今回ようやく配布する運びとなった．以下，本稿では２～５章で，VDB 作成の経緯，VDB の目的と要点，VDB の構成，利用申し込み手続きを各々述べ，６章にまとめと今後の方向を示す．

2. 経緯

VDB 作成のために，PRMU 研究専門委員会の下に，評価用映像メディア DB 作成に関する検討部会（VDB-WG，主査：馬場口，副査：栄藤・佐藤）を発足させ，1999 年の 10 月に活動を開始した．メンバーには美濃委員を始めとする映像メディア関連の有力研究者に入って頂き，DB 作成経験の豊富な京大の松山教授にアドバイザとして加わって頂いた．

VDB-WG ではまず，既存の DB を多方面から検討すると共に，VDB 作成の手段を協議した．そこで問題となったのは，素材映像として既存のものを収集するのか，あるいは新規に自主制作するのかという点であった．当初は，学術目的という御旗の下に，放送局からも供出頂けると甘く考えていたが，種々の権利関係が輻輳しており，既存の放送素材を DB として収集するのは極めて困難であることが交渉するにつれ分かってきた．一方で，NTT ドコモからは栄藤委員のご尽力により，素材を貸与頂けることとなったが，素材映像を収集するという点では苦戦を強いられていた．

そのような折に，新情報処理開発機構 RWCP の岡隆一氏から共同で素材映像を制作するという話が偶然，持ち上がった．ご存知のように RWCP は，研究の推進と同時に，数多くのメディア研究用 DB[4]を作成している．その一環として，VDB の素材映像も作成しようということであった．この話は我々にとって正に僥倖であり，これを機に VDB-WG の主方針も自主制作に転換した．

素材映像の制作は，2001 年度の前半に TBS・東通・イーストおよびパルスステーションに依頼した．いずれも TV 用の素材制作の経験が豊富な会社で，シナリオライター・俳優・監督・カメラマン・編集者・音声技術者などスタッフのいずれもがプロである．この自主制作によって複数本の素材映像がようやく集まった訳である．

2001 年度の後半からは，評価用 DB には不可欠なグランドトゥルースに相当するメタデータに関して集中的に議論した．今後の映像流通環境を考慮して，MPEG-7[5]でメタデータを与えることとした．柴田委員，栄藤委員，越後委員など MPEG-7 に精通した委員が多く，MPEG-7 の採用はすぐに決定されたが，何を記述するかに関しては，アプリケーションに依存することもあって，なかなか議論がまとまらなかった．結

表1 VDBコンテンツ一覧

ラベル	コンテンツ名	制作	著作権	長さ	音声	テロップ	メタデータ	サイズ：Mpeg1/Mpeg2
News-1	ニュース19	TBS他	TBS	15分	日/英	有り	カット/シナリオ	147MB/1.65GB
News-2	ニュース19	TBS他	TBS	15分	日/英	有り	カット/シナリオ	147MB/1.66GB
Drama	ピエロの涙	パルス	RWCP	20分	日	無し	カット/シナリオ	203MB/2.29GB
Documentary	音色～日本の夏を彩る伝統の技	パルス	RWCP	15分	日	有り	カット/シナリオ	153MB/1.73GB
Cooking	ランクアップCooking	パルス	RWCP	15分	日	有り	カット/シナリオ	154MB/1.73GB
Travel	女王様のブランチ	TBS	DoCoMo	40分	日/英	有り	カット	2.72GB(vod形式)

局，ショットの境界情報に加えて，コンテンツ記述についてはシナリオに記載されている情報を中心に記述する方針を採った．メタデータの作成には，作業グループ（柴田，中村，佐藤，馬場口の各委員）で詳細を検討し，アルバイトに実務作業を依頼した．

3. VDB の目的と要点

VDB の目的は，複数のジャンルからなる映像属性の異なる映像ストリームを標準的なベンチマークデータとして提供することにより，映像処理の各種アルゴリズムの相互比較の便宜を図り，映像メディア関連研究のレベル向上の一助とすることである．

VDB の要点は以下の通りである．

- 動画像，音声（日本語と英語），テキスト（音声のトランスクリプト），図形（テロップ文字やオーバーレイ）からなるマルチメディアデータである．
- ニュース・ドラマなど種々のジャンルを含む．
- コンテンツがストーリー性を持ち，十数分の映像長を持つ．
- 動画像（映像）について，カメラワーク（パン，ズームなど），カメラ運動（ドリー，クレーンなど），編集効果（カット，ディゾルブ，ワイプなどのショット切替）を含み，現実のテレビ放送の素材素材と限りなく近い映像品質を持つ．
- 音声について，会話音声，効果音，BGM などに関して現実のテレビ放送と限りなく近い音声品質を持つ．
- グランドトゥルースに相当するデータを備える．
- 著作権や肖像権の問題はクリアされていて，素材映像の一部を論文発表や口頭発表に利用可能である．

4. VDB の構成

4.1 素材映像

VDB に収容されている素材映像を表１に示す．何れの映像もそれ自体で一つの番組となっており，例えばニュースでは，複数のニューストピックに対し，キャスター，資料映像，フリップ，テロップなどが順に出現する構成となっており，天気予報のシーンも含む．また，２つのニュースでのトピック間で関連をもたせた部分（「News-1」で行方不明になり，「News-2」で発見される）もある．ドラマ，ドキュメンタリー，料理では，効果音や音楽も挿入されている．一方，映像属性の面では，ショット切替検出の弊害になりうるフラッシュシーンも数多く含めてあり，昼や夜のシーンなど変化に富んだ構成となっている．表２に各映像のショット切替操作の内訳を示す．表中，DVE は画面がズーム，スパイラル，めくれなどを伴って切り替るものを指す．尚，「Travel」については最初の１５分（chaptor1）に対するデータである．

表２ ショット切替操作の内訳

	カット	ディゾルブ	フェード	ワイプ	DVE
News-1	94	0	2	2	2
News-2	99	0	2	2	6
Drama	128	0	2	1	0
Documentary	140	1	3	1	1
Cooking	101	0	13	8	14
Travel	102	14	23	1	1

映像の記録形式は MPEG2 と MPEG1 の両方を使用する．図１に MPEG 符号化形式の諸元を示す．

MPEG2 program stream
video: NTSC 720x480
  bitrate: 15Mbps
  aspect: 4:3
  profile: MP@ML
audio: 48kHz sampling, stereo
  bitrate: 224kbps

MPEG1 system stream
video: NTSC 352x240
  bitrate: 1.2Mbps
  aspect: 4:3
audio: 44.1kHz sampling, stereo
  bitrate: 128kbps

図１ 映像の符号化形式

4.2 メタデータ

映像メタデータは，映像解析研究のためのグランドトゥルースとして利用されることを念頭において作成されている．「映像解析研究のためのグランドトゥルース」として要求される記述の内容については，さまざまな方針がありうる．ここでは，メタデータ作成者の主観を極力排除し，客観的に作成することによりグランドトゥルースとしての「揺らぎ」を最小限におさえるような戦略を取った．

メタデータ作成プロセスは映像制作プロセスとはまったく独立としており，映像制作者から渡されたシナリオや映像プロダクションの際のメモなど，ドキュメントとして与えられたものの他は，最終成果物の映像のみが与えられ，メタデータ作成プロセスではこれらの情報から「客観的」に得られる情報のみを記述している．すなわち，シーンの説明など映像の意味内容に関わるような記述は，シナリオやプロダクションメモなどに記述されていない限り，メタデータには取り込まれていない．一方，ショット，スピーチの内容，テロップの内容など，比較的客観的に観測可能な記述を盛り込んでいる．

映像メタデータは，MPEG-7 標準に準拠したフォーマットで提供される．実際の記述では，DescriptionMetadata 等による番組名，制作者，ジャンル，日付等に関する記述に引き続き，全体を 4 つの TemporalDecomposition に分割し，それぞれを独立した階層として記述している．これらの 4 つは，それぞれショット，トピック(ニュースの場合)，キャプション，スピーチの階層に対応している．

以下ニュースを例にそれぞれの階層の記述について述べる．

```
<AudioVisualSegment id="structure-###">
 <StructuralUnit href="urn:ricoh:mmVISION:... ">
   <Name xml:lang="en">shot</Name>
  </StructuralUnit>
 <TextAnnotation type="classification">
  <FreeTextAnnotation>\class</FreeTextAnnotation>
 </TextAnnotation>
 <MediaTime>
  <MediaRelTimePoint mediaTimeBase="/Mpeg7/... ">
  \start</MediaRelTimePoint>
 <MediaIncrDuration mediaTimeUnit=
  "P0DT0H0M0S1N30F">
  \duration</MediaIncrDuration>
 </MediaTime>
</AudioVisualSegment>
```

図 2　ショットを表す AudioVisualSegment の例

(i)ショット階層に関する記述

ショット階層では，カットやワイプ等によるシーン切り替えにより分割されたショットに関する情報を記述している．各ショットは，ショット階層を表す TemporalDecomposition 内の複数の AudioVisualSegment として記述され，それぞれの所属する階層（=shot），分類（ニュースの場合，スタジオ内，街頭インタビュー，中継のレポータ，CG などのショット分類），開始時間（シーン切り替え時点に一致），継続時間が記述されている．図 2 にショットを表す AudioVisualSegment の例を示す．

実際には，図中の **\class** にはショット分類を表す文字列が記述され，**\start** には開始時間，**\duration** には継続時間がそれぞれ記述されることになる．ショット階層を表す TemporalDecomposition には，ショットの数だけ上記のような AudioVisualSegment が連続することになる．

(ii)トピック階層に関する記述（ニュースの場合）

トピック階層は，ニュース映像におけるニューストピックに関する情報を記述している．実際の記述はショット階層に関する記述にならって表現しており，トピック階層を表す TemporalDecomposition 内の複数の AudioVisualSegment として記述され，それぞれの所属する階層（=topic），分類（ニュースの場合，政治，国際，経済，文化などのトピック分類），開始時間（原則としてショット境界と一致），継続時間が記述されている．

(iii)キャプション階層に関する記述

キャプション階層では，いわゆるテロップに関する情報を記述している．キャプション階層もショット階層に関する記述にならって表現しており，キャプション階層を表す TemporalDecomposition 内の複数の AudioVisualSegment として記述され，それぞれの所属する階層（=caption），分類，テロップの内容，開始時間，継続時間が記述されている．分類は，ショット階層と同様，type 属性が classification の TextAnnotation として記述しているが，その内容としては，客観的に判断できる記述とするため，テロップの出現位置（上部，下部左側，等）のみを記述している．テロップの内容は，type 属性が text の TextAnnotation として記述している．

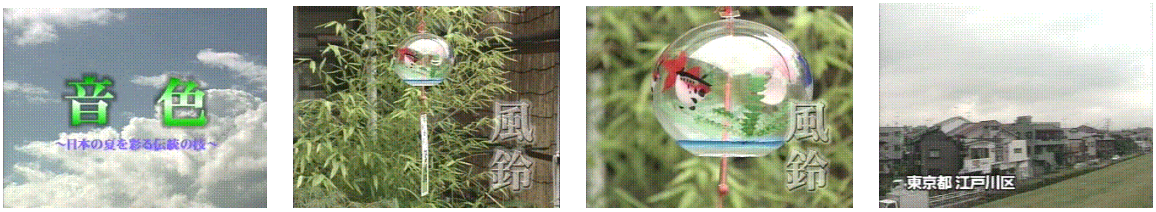
(iv)スピーチ階層に関する記述

スピーチ階層では，出演者の発話内容に関する情報を記述している．スピーチ階層も他の階層と同様の記述にならって表現しており，スピーチ階層を表す TemporalDecomposition 内の複数の AudioVisualSegment として記述され，それぞれの所属する階層（=speech），話者名および性別，発話内容，開始時間，継続時間が

研究用素材 『伝統工芸編』 音色～日本の夏を彩る伝統の技～

Time	Cut	解説･セリフ
00:00 00:01	(1) 青空 T 音色 ～日本の夏を彩る伝統の技～	NA 音･色。 日本の夏を彩る伝統の技
00:09 00:11	(2) 風鈴、風に揺れる T 風鈴	NA 蒸し暑い夏。
00:15	(3) 風鈴下より PAN	心地好い音色で涼を運んできてくれる風鈴。 しかし 300 年の歴史を持つ「江戸風鈴」を作る工房が 残り後1軒になってしまっているのです。
00:21	(4) 河原→住宅PAN T 東京都江戸川区	NA 東京、江戸川区。 江戸川の川伝いに広がる住宅街の中に、
00:27	(5) 篠原風鈴本舗 入り口 T 江戸風鈴 篠原風鈴本舗	その1軒はありました。 篠原風鈴本舗、

図３ シナリオの一部（ドキュメンタリー）



（1） （2） （3） （4）

図４ 図３の画像フレームの例

記述されている．話者名および性別については，シナリオなどから明確である場合，および映像などから明らかである場合に限り，type 属性がそれぞれ speaker_name および speaker_gender の TextAnnotation として記述している．type 属性が text の TextAnnotation として記述している．発話内容は，type 属性が text の TextAnnotation として記述しており，シナリオに記述されているものにとどまらず，聞き取れる限りすべての発話を記述している．

4.3 シナリオ

素材映像の内，ドラマ，ドキュメンタリー，料理についてはシナリオも公開する．その一部を図３に示す．さらに図３のシナリオの第 1～第４ショット（シナリオでは cut と記載）の画像フレームを図４に示す．

5. 利用申し込み手続き

素材映像およびメタデータは，利用を申請する研究者からの申し込みに基づき配布される．詳細については

映像データベース Web ページ

http://research.nii.ac.jp/VDB/

をご参照頂きたい．

具体的な手続きは，上記 URL から「研究用素材映像コンテンツ使用許諾に関する覚書」を取得し，必要事項を記入の上捺印した覚書を二通作成し，一通を指定の申し込み先に郵送することにより利用申し込みを行う．申し込みを受けて，DVD 数枚に収められた素材映像が返送される．メタデータおよび付随するドキュメントについては，上記 URL より最新版を取得することができる．申し込みに当たっては，実費程度を徴収する予定である．

6. むすびと今後の方向

既に述べたように，今回の VDB は，ベンチマークデータとして利用されることを主眼としている．例えば，実際のテレビ放映の映像群に対してモデル化したアルゴリズムを最終的に評価する際に利用してもらうという形態である．相互比較のための便宜を向上させるため，VDB のユーザから処理結果や解析結果をフィードバックしてもらうプロセスが重要と考えており，そのような情報を共有し得る仕掛けの導入が急務である．

その一案であるが，解析・処理結果を MPEG-7 のタグ形式（XML 形式）で提出してもらい，メタデータとして再公開するというサイクルを繰り返す，いわば雪だるま式メタデータ増強計画は如何であろうか．第一次配布のメタデータは客観的に決め得る最小限のセットと考えられ，順次拡張していけばさらに使いやすい DB になると想像できる．

また，VDB の作成を通じて，アプリケーションを考慮しつつ,映像コンテンツに対して，何をどのように記述するかを定めることの難しさを痛感した．この VDB を土台にしてそのような問題を議論する良い題材になることを期待している．

ところで，本 VDB に収容されている映像のジャンルは，ニュース，ドラマ，ドキュメンタリー，情報番組であるが，現在の TV 番組の主要ジャンルであるスポーツについては映像を収容できなかった．これは権利面での問題のため既存のものを収集できなかったことによる．スポーツ番組を自主制作ということは非現実的で何らかの方策を見出さねばならない．また，収容するストリーム数の拡充にも継続的に努力する必要がある．

尚，VDB を用いた映像処理関連のコンテスト（例えば，映像要約）を今後企画して，VDB の普及を促進することも視野に入れている．多くの PRMU・CV・マルチメディア・音声の研究者が VDB を利用して，映像メディア関連研究が益々発展することを願ってやまない．

謝辞

VDB の作成にあたり，数え切れぬほどの多くの方々のご支援を頂いた．まず，VDB-WG の活動にご理解，ご支援賜った PRMU 研究専門委員会・谷内田前委員長，横矢委員長を始めとする PRMU 専門委員会各位に深謝する．

電子情報通信学会情報・システムソサイエティ（ISS）からは 2000, 2001 年度に WG 活動の補助金を頂いた．国立情報学研究所には共同研究として議論の場を提供頂くと共に，配布作業にも多くのご助力を頂いた．

新情報処理開発機構（RWCP）の岡隆一，橋口博明の両氏には，素材映像の制作にあたり，多大なるご援助を賜った．RWCP のご援助がなければ，VDB は陽の目を見なかったものに違いない．また，NTT ドコモからは貴重な映像データを貸与頂いた．

大吉なぎさ・中尾聡（TBS），東通およびイーストの担当者，藤本せつこ・井上高宏（パルスステーション）の各位には素材映像の制作において大変お世話になった．

メタデータの作成には MovieTool β を一部利用した．MovieTool β は（株）リコーと（株）次世代情報放送システム研究所の研究協力で開発し，リコーにより無償提供されているツールである．メタデータ作成は，孟洋・井手一郎（国立情報学研究所），織田友恵（お茶の水女子大），伊津野英克（筑波大），小倉武紘・八尾智之（阪大）の各氏の献身的な努力による．更に，小倉・八尾の両氏は VDB ホームページ作成にも寄与した．

ここに記して以上の方々に深甚なる感謝の意を表する．

文　献

[1] http://www.etl.go.jp/~etlcdb/
[2] 画像理解評価用画像データベース，1999～2001.
[3] http://www.hoip.jp/web_catalog/top.html
[4] http://www.rwcp.or.jp/wswg/rwcdb/
[5] “Overview of the MPEG-7 Standard (version 6.0),” ISO/IEC JTC1/SC29/WG11 N4509, 2001.