画像検索のための MPEG-7 標準最新動向
- Video Signature と CDVS -

Latest Activities in MPEG-7 Standards for Image Searching

- Video Signature and CDVS -

岩元浩太
Kota IWAMOTO

NEC 情報・メディアプロセッシング研究所
k-iwamoto@ay.jp.nec.com

概要： 本稿では，MPEG-7 標準（ISO/IEC 15938）における画像・映像特徴量の規格化に関する最新動向として，2010 年に規格発行された Video Signature と，現在規格策定中である CDVS（Compact Descriptors for Visual Search）について紹介する．これまで，MPEG-7 では類似検索などに汎用的に使える色・テクスチャ・形状などのビジュアル特徴量を規格化してきたが，ここ数年ではより特定の用途に特化した専用特徴量の規格化を進めている．Video Signature は，映像コンテンツを一意に識別することができる「指紋」特徴量を規格化している．各種改変・編集が加わっても映像の複製（コピー）を検知することができるため，コンテンツの不正流通検知や使用履歴調査などの応用に活用できる．一方で CDVS では，画像・映像コンテンツ自体を記述する特徴量ではなく，画像内に映る実世界のオブジェクトを検索するための特徴量の規格化を，2014 年の規格発行を目指し，進めている．撮影角度や照明条件などの撮影環境に頑健でコンパクトな特徴量を規格化することで，実世界オブジェクトの検索サービスを実現する共通ツールとして，その完成が期待されている．

1. はじめに

MPEG-7（ISO/IEC 15938）は，画像・映像を含むメディアコンテンツの内容標記のためのメタデータ体系を規定する標準であり，2002 年に規格の第 1 版が発行された．メタデータ体系を規格化することで，メディアコンテンツを扱うアーカイブ・サービス・デバイス間で，相互に検索・活用可能なインターオペラビリティを確保し，コンテンツをめぐる市場創出やエコシステムの形成を目的としている．MPEG-7 標準の中核は，コンテンツを検索可能にするための各種メタデータの共通フォーマットを規定するツール群であり，メタデータの種別によって異なるパートで規格化を進めている．Part-3 Visual[1]は映像・画像特徴量，Part-4 Audio[2]は音響特徴量，のローレベル信号特徴量の規格化を担当している．Part-5 MDS[3]は書誌情報（一般的なメタデータ）の記述方式の規格化を担当している．

その中で Part-3 Visual は最もアクティブに標準化活動が進められており，用途・目的に合わせた様々な画像・映像特徴量が規格化され，活用用途の拡大と技術進展も伴い，第 1 版発行後も追補（Amendment）によりツール群の拡張が行われてきた（表 1）．当初は，色・テクスチャ・形状・動きなどの汎用的に幅広く使えるビジュアル特徴量のツール群を規格化した．これらは，画像・映像の類似検索を主用途として，幅広いアプリケーションで活用できる基本的な特徴量である．これらの規格では，インターオペラビリティを確保する最低限の要素として，特徴量の記述フォーマット(ビットストリームシンタックス)を標準必須とし，特徴量の抽出方法と照合方法は，推奨方式は記載するものの，標準必須外としている（図 1）．

一方で，ここ数年では，より特定に用途に特化した専用の特徴量の規格化が進められてきた．Image Signature [4]（Part-3 Amd.3, 2009 年追補）と Video Signature [5]（Part-3 Amd.4, 2010 年追補）は，画像・映像コンテンツを一意に特定するための「指紋」特徴量を規格化している．コンテンツの複製（コピー）を高精度に検知することができるため，インターネット上に

表 1: MPEG-7 画像・映像特徴量の規格化の概要.

ツール名	内容・機能	用途・アプリ	パート	発行年
Dominant Color	色特徴	汎用用途（類似検索など）	Part-3	2002
Scalable Color				
Color Layout				
Color Temperature				
Homogeneous Texture	テクスチャ特徴			
Edge Histogram				
Region Shape, Contour Shape	形状特徴			
Contour Shape				
Motion Activity	動き特徴			
Parametric Motion				
Advanced Face Recognition (AFR)	顔特徴	顔検索	Part-3 Amd.1	2004
Image Signature Tools	画像の指紋特徴	画像の複製(コピー)検知	Part-3 Amd.3	2009
Video Signature Tools	映像の指紋特徴	映像の複製(コピー)検知	Part-3 Amd.4	2010
Compact Descriptors for Visual Search (CDVS)	オブジェクト検索用特徴	実世界オブジェクトの検索	Part-13	2014 予定

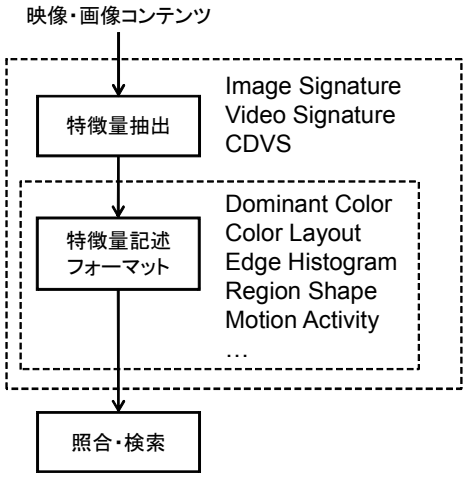

図 1: 規格化（標準必須）のスコープ.

拡散する不正コピー・流通の検知に用いることができる. さらに, 新設の Part-13 として現在規格策定が進んでいる CDVS（Compact Descriptor for Visual Search）[6]では, これまでの特徴量のように画像・映像コンテンツ自体を記述する特徴量ではなく, 画像・映像内に映る実世界のオブジェクトを検索するための特徴量の規格化を目指している. 主に, モバイル端末向けの実世界オブジェクトの検索サービスをターゲットとしている. これらの新たな MPEG-7 画像・映像特徴量は, 目的を特化することで, その用途においては圧倒的な性能を発揮できるように設計されている. また性能を発揮するために, 特徴量の記述フォーマットだけではなく, 特徴量の抽出方法も標準必須の要素として厳密に規定している（図 1）.

本稿では, MPEG-7 での画像・映像特徴量の規格化の最新動向として, 発行済みの最新規格である Video Signature と, 現在規格策定中である CDVS について, 規格内容の詳細について紹介する.

2. Video Signature

2.1 背景・目的

これまでの色・テクスチャ・形状・動きなどのビジュアル特徴量は, それぞれの特徴が表す信号特性に基づいてコンテンツの「類似性」を判断することができた. それに対して, Image Signature（Part-3 Amd.3: Image Signature Tools）と Video Signature（Part-3 Amd.4: Video Signature Tools）は, コンテンツの「同一性」を判断するための一意識別可能な「指紋」特徴量を規定する. すなわち, コンテンツに対してユニークな ID を付与する仕組みを提供する. これにより, 画像・映

像コンテンツの複製(コピー)を検知することが可能になる.

Image/Video Signature は, インターネットの普及に伴い爆発的に流通・拡散が進む画像・映像コンテンツを, 統一的な仕組みで識別・管理したいニーズに応えるものとして規格化された. 特に, コンテンツの不正コピー・不正流通がコンテンツ産業の根幹を揺るがす社会問題として拡大してきたため, これらの不正コピー・不正流通されたコンテンツを発見できる共通ツールとしての期待が大きい. その他にも Image/Video Signature はコンテンツの使用実績調査や, コンテンツ間のリンク生成などにも活用できる.

2.2 技術要件

画像コンテンツと映像コンテンツはその特性や使用方法が異なり,「指紋」特徴量として技術要件も若干異なるため, Image Signature と Video Signature という個別の標準として規格化された. Video Signature 特徴量は, 2.1 節に記載した目的に対応できるように, 以下の技術要件をクリアするように設計されている.

- 頑健性: 複製(コピー)で各種改変・編集(テロップ重畳, 符号化圧縮など)が加わっても, 頑健にコンテンツの同一性が判定できること.
- 識別能力: 大量映像中でも, 異なるコンテンツは十分異なると判定できること. すなわち, 誤判定が極小であること.
- 高速性: ネット規模の大量映像に対応できるように, 特徴量の抽出・照合速度が高速であること.
- コンパクト性: 特徴量のサイズが小さいこと.
- 部分照合機能: 映像の一部区間が切り出されても, 区間を特定できること.

2.3 Video Signature 規格

Video Signature の規格(Part-3 Amd.4)[5]では, 特徴量の抽出方法と, 特徴量データの記述フォーマットを規定する. Video Signature の特徴量データは, フレーム単位の特徴を記述する Frame Signature, Frame Confidence と Word, そして連続する複数フレーム(90 フレーム, 45 フレーム間隔)から構成される区間単位の特徴を記述する BagOfWords, という要素から構成される. それぞれの構成要素の内容を表 2 に, 特徴量データの構造を図 2 にまとめる. また, 特徴量抽出のフローを図 3 に示す. 以下, それぞれの詳細を説明する[7].

表 2: Video Signature の構成要素.

単位	要素名	内容
フレーム	Frame Signature	絵柄の構造を表す 380 次元特徴量. 76 バイト/フレーム.
	Frame Confidence	Frame Signature の信頼度. 1 バイト/フレーム.
	Word	Frame Signature の集約特徴量. 5 バイト/フレーム.
区間	BagOfWords	連続フレーム区間(90 フレーム, 45 フレーム間隔)を表すバイナリヒストグラム特徴. 152 バイト/区間.

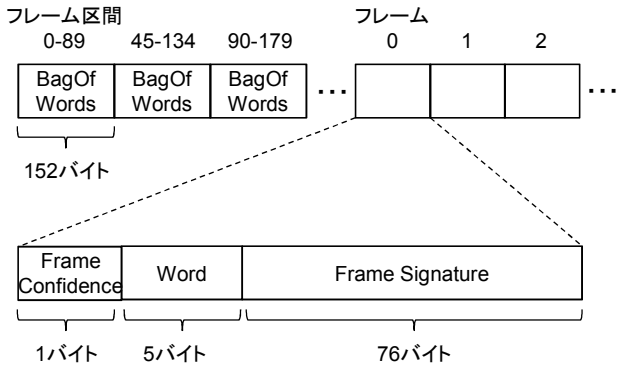

図 2: Video Signature 特徴量のデータ構造.

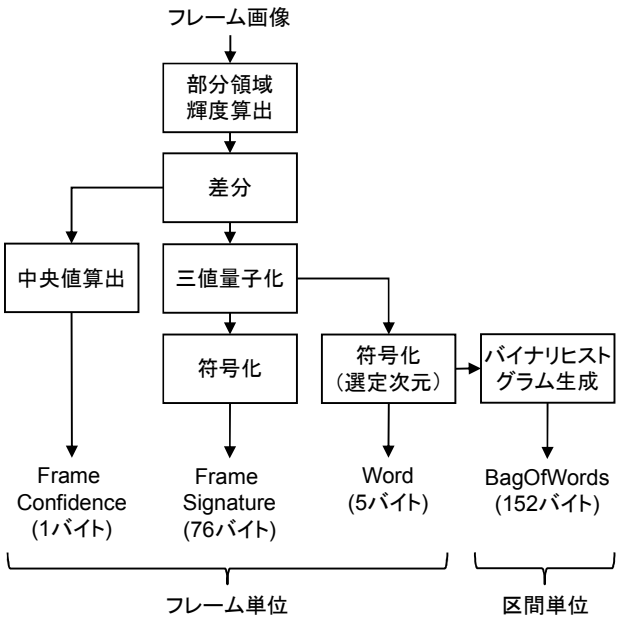

図 3: Video Signature の抽出方法.

2.3.1 Frame Signature の抽出

Frame Signature は，映像の各フレームから抽出される，絵柄の構造を表す特徴量であり，フレームの様々な部分領域間の輝度の大小関係を表す 380 次元の特徴ベクトルである．380 次元の各要素について輝度を比較する部分領域のペアが規定されており，図 4 に示すように多様なスケール・形状・位置から構成されている．このパターンの多様性により，特徴量に高い識別能力と頑健性をもたらしている．また画像の周辺よりも中央領域を重視し，フレームの中央のほうがより密に部分領域がサンプリングされている．各部分領域ペアの輝度差分を三値{0, 1, 2}に量子化する．次元 i の部分領域ペアのそれぞれの平均輝度値を $v1_i$ と $v2_i$，輝度差分を $d_i=v1_i-v2_i$ とすると，三値 x_i は次式で算出される．

$$x_i = \begin{cases} 2 & (\text{if} \quad d_i > th) \\ 1 & (\text{if} \quad |d_i| \le th) \\ 0 & (\text{if} \quad d_i < -th) \end{cases} \qquad (1)$$

ここで閾値 th は，フレームごとに輝度差分の分布を考慮して，量子化値の出現頻度が均等（1/3 ずつ）になるように適応的に決定される．この量子化により，輝度変化に対する頑健性を確保すると同時に，特徴量の識別能力を最大化することができる．

こうして求められた 380 次元の三値特徴ベクトル $\mathbf{x} = \{x_1, x_2, \cdots, x_{380}\}$ を，5 次元ずつまとめて 1 バイトに符号化し，76 バイトのコンパクトなデータに圧縮する．符号化値 b_j (j=1,…,76)は，次式で計算される．

$$b_j = 81 \times x_{5j-4} + 27 \times x_{5j-3} + 9 \times x_{5j-2} + 3 \times x_{5j-1} + x_{5j} \qquad (2)$$

この符号化方法により，三値要素を各 2 ビットずつ独立して符号化するよりも，20%符号量を削減することができる．

2.3.2 Frame Confidence の抽出

Frame Confidence(信頼度)は画像の複雑さを表す数値で，Frame Signature の有効度を示す．Frame Confidence の算出には，Frame Signature の抽出で用いた部分領域間の輝度差分の絶対値$|d_i|$から，その中央値を求め，それを 1 バイト(0-255)で記述する．平坦で特徴のない絵柄のフレームは Frame Confidence が低い値となる．Frame Confidence は Frame Signature の照合で併用し，平坦なフレームによる誤検出を排除するのに用いられる．

2.3.3 Word と BagOfWords の抽出

Word は，各フレームから抽出される三値 380 次元

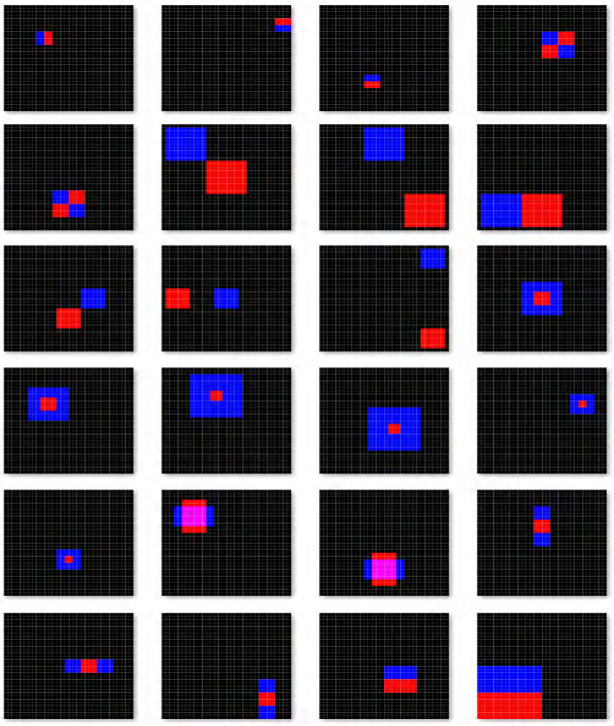

図 4: 部分領域ペアの例．

（=76 バイト）の Frame Signature を，三値 5 次元（=1 バイト）×5 つに集約した集約特徴量である．380 次元のうち，識別に特に有効な既定された 5 次元×5 つ（=25 次元）を，式(2)により符号化して生成する．

BagOfWords は，連続する 90 フレーム区間（45 フレーム間隔）を表す特徴であり，それらのフレームの Word をヒストグラム化したものである．具体的には，90 フレーム区間において，5 つの Word それぞれに対して，取り得る値 0～242（3^5-1）をビンとしたバイナリヒストグラムを生成する．各バイナリヒストグラムが 243 ビットのため，5 つで 1215 ビット(=152 バイト)となる．なお，Word と BagOfWords は共に，元の Frame Signature があれば完全に再現できる．

Word と BagOfWords は，照合の第一段階で用いることで，効果的なフィルタリングを行うことができ，超高速な照合を可能にする．

2.3.4 照合方法（標準必須外）

Video Signature 特徴量の照合方法は標準必須ではないが，Part-8 のテクニカルレポート[8]に推奨の照合方式が記載されている．ここに記載された照合方式では，2 つの映像の間の Video Signature 特徴量の照合を以下の 3 ステップで行っている．

(1) BagOfWords を用いて照合区間を絞り込み．
BagOfWords の照合には Jaccard 距離を用いる．

(2) Frame Signature を用いたフレーム系列の 1 対 1 照合により一致区間(始端と終端)を特定. Frame Signature の照合には L1 距離を用いる.

(3) Frame Confidence の低い一致区間を誤検出として除去.

2.4 性能評価

MPEG では, 2.2 節で述べた技術要件に沿って, 規格化した Video Signature の性能評価を行っている. 具体的には, 3 分の元映像から切り出された短い映像クリップ(2 秒, 5 秒, 10 秒)に対して各種改変を加えた改変映像クリップを作成し, これから元映像とその切り出し位置を識別できるかを評価した. 加えた改変はテロップ重畳, カメラ撮影, 符号化圧縮, 解像度縮小, アナログ録画, 明度変換, モノクロ変換, IP 変換, フレームレート変換の計 9 種類である(図 5). なお評価には, 映画・ホームビデオ・バラエティ番組などの多様なジャンルの映像データ 100 時間以上を使用している.

まずは, 全く無関係なクリップ同士を照合させ, 誤検出率が 5ppm(100 万分の 5)という極めて低い率を実現する照合閾値を決定する. 決定された照合閾値を用いて改変映像クリップと元映像を照合させ, 正しく識別でき, かつ切り出し位置を 1 秒以内のズレで特定できる正当率(識別率)を評価した. 図 6 に 2 秒の改変映像クリップに対する識別率の結果を示す. Video Signature を他の映像特徴量である Difference Block Luminance [9]と Ordinal Measure [10]と比較している. Video Signature は, あらゆる改変に対して安定した性能を示しており, 平均して 96%の識別率を達成した. 他の特徴量[9][10]と比較して, 全ての改変に対して精度改善が認められるが, 特にテロップ重畳(+39%)とカメラ撮影(+62%)で顕著である.

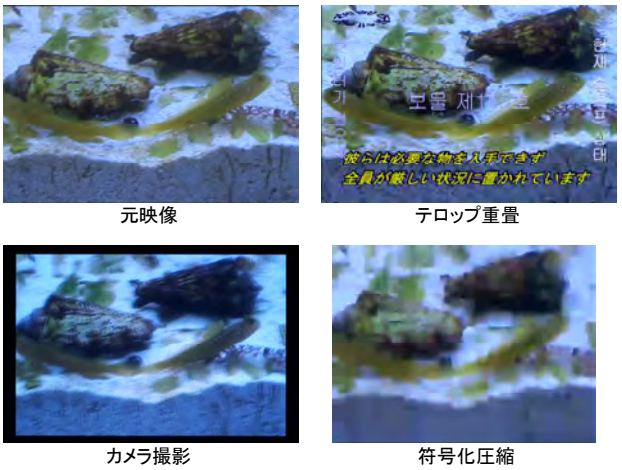

図 5: 改変映像クリップの例.

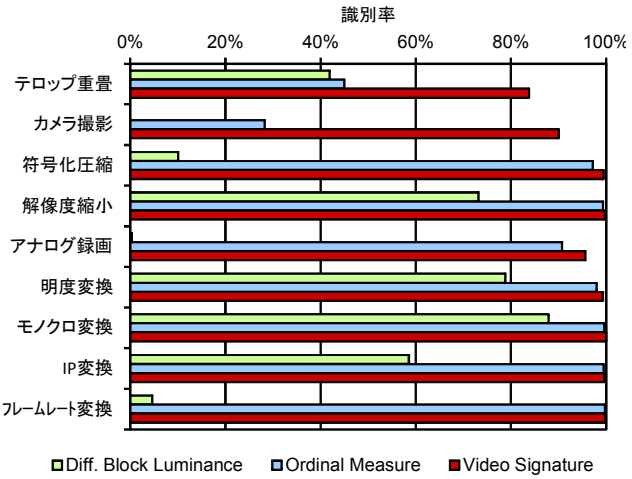

図 6: 性能評価結果(改変映像クリップ=2 秒).

3. CDVS

3.1 背景・目的

CDVS(Compact Descriptor for Visual Search)標準化では, これまでの画像・映像コンテンツ自体を記述する特徴量ではなく, 画像・映像内に映る商品・印刷物・建物などの実世界のオブジェクト(textured rigid objects)を検索するための特徴量の規格化を進めている.

画像・映像内のオブジェクトを検索する技術は, 近年コンピュータビジョンの分野で盛んに研究開発が行われている. 1998 年に開発された SIFT 特徴量[11]の出現により, カメラの撮影角度・サイズの変化に対して極めて頑健にオブジェクトを検索することが可能になった. 2000 年代に入ると, SIFT を精度や速度の面で改良した派生版の局所特徴量を多数提案され, 実世界オブジェクトの検索技術の研究開発が活発化した. 2000 年代後半には Google Goggles や SnapTell など, こうした局所特徴量技術を活用したモバイル端末(スマートフォン)向けの実世界オブジェクトの検索サービスがスタートしている. このように実世界オブジェクト検索の実用化は徐々に進んでいるが, 各サービスが独自技術で実装されているため, 大きな広がりは見せていない. CDVS は, このような背景のもと, オブジェクト検索を実現する共通特徴量を規格化することで, 多数のデータベース・サービス・端末間での相互運用を可能にすることを目的に, 2010 年に規格化プロジェクトがスタートした. 2011 年 7 月に技術募

集（CfP: Call for Proposals）[6]が発行され，それ以降技術選定を進めている．なお，MPEG-7 Part-3 Visual ではなく，新設の MPEG-7 Part-13 CDVS として規格化を進めることになった．

CDVS の主なターゲットアプリは，図 7 に示すようなモバイル端末による実世界オブジェクトの検索サービスである．このシステムでは，まずモバイル端末内で撮影された画像から特徴量を抽出する．抽出した特徴量を，データベースを保有する検索サーバに 3G 回線などのネットワーク経由で送信する．検索サーバでは，特徴量を受信すると検索を行い，検索結果をモバイル端末に返信する．この際に，レスポンスタイムを短くするためには，モバイル端末からサーバに送信する特徴量のサイズを小さくする必要がある．リアルタイムレスポンスを実現するために，JPEG 画像や SIFT 特徴量の 1/10 以下のサイズを目指して規格化を進めている．

3.2 技術要件

CDVS 特徴量は，3.1 節に記載した目的に対応できるように，以下の技術要件をクリアするように技術選定を進めている．

- 頑健性: 撮影角度・サイズ，照明条件，オクルージョンなどの撮影環境に対して頑健なこと．
- コンパクト性: 検索サービスの低遅延応答が可能なように，特徴量サイズが小さいこと．
- スケーラビリティ: アプリケーションや帯域状況などに応じて特徴量サイズを自由に変更できること．
- 抽出高速性: モバイル端末でも高速かつ省メモリで特徴量抽出ができること．
- 照合高速性: 大規模データベースでも検索が高速なこと．
- 位置特定機能: 認識したオブジェクトの画像中の位置を正確に特定可能なこと．

3.3 性能評価フレームワーク

現在 CDVS 標準化は規格策定段階であり，定められた性能評価フレームワーク[12]に従ってコア実験を行い，技術選定と改善を進めている．性能評価フレームワークでは，図 8 に示すような印刷物，一般物体，建物などの実世界のオブジェクトを多種多様な撮影環境（撮影角度・サイズ，照明条件，オクルージョン）で撮影した画像群を使って評価を進めている．

性能評価は，(1)対照合実験と，(2)検索実験の 2 つの実験で進めている．対照合実験では，画像対を照合してその中に共通のオブジェクトがあるか否かを判定する．共通のオブジェクトを含まない画像対を用いて誤判定率 1%となる照合閾値を定め，その照合閾値を用いて共通のオブジェクトを含む画像対を照合させ，正判定率を評価する．また対照合実験では，位置特定の精度も評価する．検索実験は，100 万件規模のデータベースから，クエリ画像に含まれるオブジェクトを含む画像を検索する，という実験である．検索結果の精度を MAP（Mean Average Precision）で評価する．各実験について，1 画像あたりの特徴量サイズを 0.5KB，1KB，2KB，4KB，8KB，16KB の 6 種類に対して精度評価を行う．特徴量抽出時間，対照合

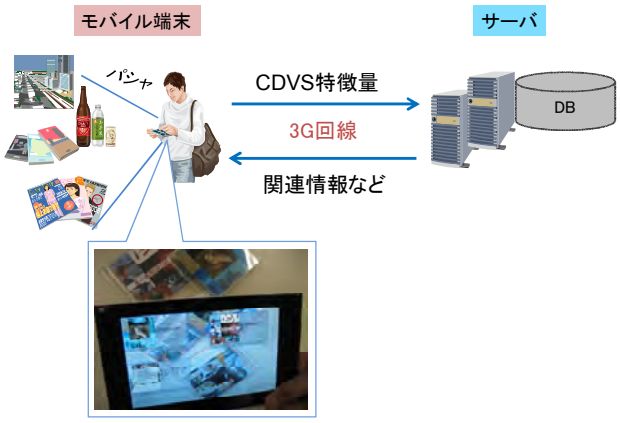

図 7: 実世界オブジェクト検索サービス．

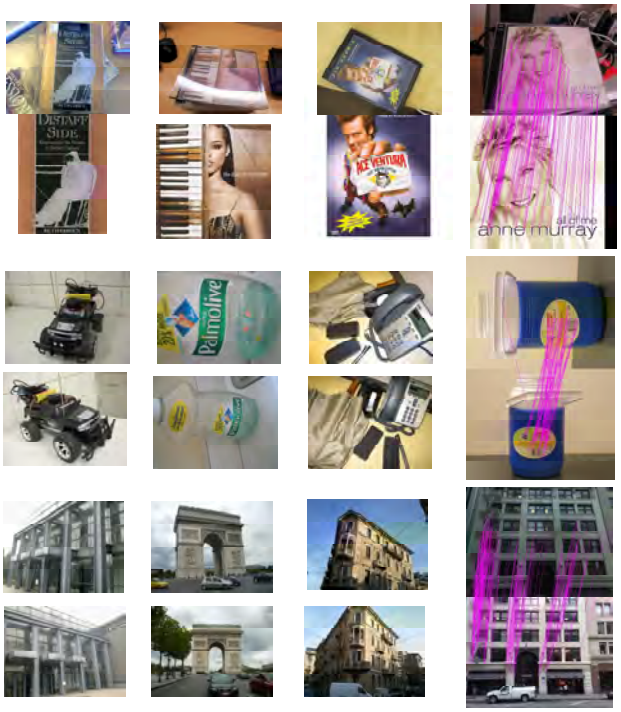

図 8: CDVS 標準化の性能評価で用いる画像例．

実験の照合時間，検索実験の検索時間に対して処理時間の制約を設け，その制約内で総合的に最も高精度な技術を標準化技術として採用する形で，技術選定を進めている．

3.4 現状と今後の予定

CDVS 標準化は，現状（2013 年 7 月現在）で WD（Working Draft）段階（第 4 版）であり，コア実験を通して技術選定と改善を現在進行中で進めている．現状の規格ドラフト案は，以下の構成要素で構成されている（現在も修正・改善中であり，変更の可能性あり）．

- LoG（Laplacian of Gaussian）フィルタを用いた局所特徴点の検出方式
- 勾配ヒストグラム（HoG: Histogram of Gradients）を量子化し，コンパクトに圧縮された局所特徴量
- 高速検索用に，局所特徴量を統合して生成されたグローバル特徴量
- 効率的な走査方法と算術符号を用いた座標値符号化方式

現状方式は，SIFT 特徴量[11]と比較して圧倒的に小サイズながら，SIFT と同等以上の精度を実現している．

今後の予定では，2013 年 8 月に CD(Committee Draft)，2014 年 1 月に DIS（Draft of International Standard），2014 年 8 月 FDIS（Final Draft of International Standard）であり，順調にいけば 2014 年夏には規格が発行される．その後，参照ソフトウェア，適合性試験，テクニカルレポートなどの周辺規格も順次発行される予定である．

4. まとめ

本稿では，MPEG-7 での画像・映像特徴量の規格化の最新動向として，最新規格である Video Signature と，現在規格策定中の CDVS について報告した．Video Signature は，映像コンテンツを一意に識別することができる「指紋」特徴量を規格化しており，コンテンツの不正流通検知などの応用に活用できる．CDVS では，画像内に映る実世界のオブジェクトを検索するための特徴量の規格化を進めており（2014 年規格発行予定），モバイル端末を用いた実世界オブジェクトの検索サービスのための共通ツールとして活用できる．Video Signature も CDVS も，それぞれの用途に特化した専用特徴量として圧倒的な性能を有しており，今後の実用化・商用化が大きく期待される．

参考文献

[1] ISO/IEC 15938-3, “Information Technology - Multimedia content description interface - Part 3: Visual”.

[2] ISO/IEC 15938-4, “Information Technology - Multimedia content description interface - Part 4: Audio”.

[3] ISO/IEC 15938-5, “Information Technology - Multimedia content description interface - Part 5: Multimedia description scheme”.

[4] ISO/IEC 15938-3 AMD. 3, “Information Technology - Multimedia content description interface - Part 3: Visual, Amendment 3: Image signature tools”.

[5] ISO/IEC 15938-3 AMD. 4, “Information Technology - Multimedia content description interface - Part 3: Visual, Amendment 4: Video signature tools”.

[6] MPEG, "Call for Proposals for Compact Descriptors for Visual Search", ISO/IEC JTC1/SC29/WG11 N12201, July 2011.

[7] S. Paschalakis, K. Iwamoto, P. Brasnett, N. Sprljan, R. Oami, T. Nomura, A. Yamada, and M. Bober, “The MPEG-7 Video Signature Tools for Content Identification”, IEEE Trans. on Circuits and Systems for Video Technology, vol. 22, issue 7, pp.1050-1063, 2012.

[8] ISO/IEC TR 15938-8 AMD. 6, “Information Technology - Multimedia content description interface - Part 8: Extraction and use of MPEG-7 descriptions, Amendment 6: Extraction and matching of video signature tools”.

[9] J. Oostveen, T. Kalker, and J. Haitsma, “Feature extraction and a database strategy for video fingerprinting”, Proc. of 5th Int’l Conf. on Recent Advances in Visual Information Systems, pp.117-128, 2000.

[10] X.-S. Hua, X. Chen, and H.-J. Zhang, “Robust video signature based on ordinal measure”, Proc. of ICIP2004, 2004.

[11] D. Lowe, “Distinctive image features from scale-invariant keypoints”, IJCV, 60, 2 (2004), 2004.

[12] MPEG, "Evaluation Framework for Compact Descriptors for Visual Search", ISO/IEC JTC1/SC29/WG11 N12202, July 2011.