2004－AVM－47（7）
2004／12／9

被写体の位置情報を用いたコンテンツ分類

伊藤 学[1] 小池 真由美[2] 池田 佳代[3] 日高 宗一郎[4] 青木 輝勝[1]

[1]東京大学 先端科学技術研究センター 〒153-8904 東京都目黒区駒場 4-6-1
[2]エスパリエ 〒175-0094 東京都板橋区成増 3-20-16
[3]エクセリードテクノロジー 〒167-0054 東京都杉並区松庵 3-20-11
[4]国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: [1]{ito, aoki}@mpeg.rcast.u-tokyo.ac.jp [2]koike@espalier.co.jp
[3]kayo@excellead.jp [4]hidaka@nii.ac.jp

あらまし 近年，ハイスペックPCや大容量HDDが出現し，デジタルコンテンツが爆発的な増加を続けている中，大量に蓄積されたコンテンツを効率よく探すための検索技術が注目を集めてきた．しかしながら，効率的な検索手法は未だ十分とはいえず，この要因として考えられるのは，コンテンツそれぞれが持つ多義性に他ならない．本報告では，検索効率の向上を大目標とし，ある程度の候補画像を絞り込むための分類として，多くの自然画像に対し，人間の経験値より共通に認識される被写体の位置情報や，カメラの高さ及び角度といったパラメータを組み合わせた分類法と，特徴量表現について述べる．

キーワード 画像分類，画像検索，奥行き情報，検索クエリー

Content Classification Using Photographic Subject's Position

Manabu ITO[1] Mayumi KOIKE[2] Kayo IKEDA[3] Soichiro HIDAKA[4] and Terumasa AOKI[1]

[1]The University of Tokyo RCAST 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8904 Japan
[2]Espalier 3-20-16 Narimasu, Itabashi-ku, Tokyo, 175-0094 Japan
[3]Excelled Technology 3-20-11 Syouan, Suginami-ku, Tokyo, 167-0054 Japan
[4]National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
E-mail: [1]{ito, aoki}@mpeg.rcast.u-tokyo.ac.jp [2]koike@espalier.co.jp
[3]kayo@excellead.jp [4]hidaka@nii.ac.jp

Abstract In recent years, the high spec. PC and large scale HDD appear, and digital contents are continuing the explosive increase. The reference technology for under these circumstances looking for the contents accumulated in large quantities efficiently attracts attention. However, it cannot be said that the efficient reference technique is still enough. The polysemy which each contents have is thought as this factor. This report describes the classification for narrowing down a candidate picture by considering improvement in reference efficiency as the with-kindly-tolerance mark. It is the taxonomy, which combined a parameter called the height and angle of a camera with a photographic subject's position information recognized in common than man's experience value. Moreover, the amount expression of the features using the parameter is also described..

Keyword Image Classification, Image/Video Retrieval, Depth Information. Retrieval Query

1. はじめに

高速ネットワークの常時接続や，ハイスペック PCの普及により，デジタル化された画像の送受信がいつでもどこでも可能となった．しかしながら，自分の意図する画像を効率よく検索する技術は未だ十分とはいえない現状がある．

従来画像検索の研究においては，動画像中のどのような特徴量を用いて画像検索を行うか，その特徴量を動画像中からどのようにして抽出するか，あるいはあらかじめ検索目標となる画像が用意されており，どの

ようにしてその目標画像が動画像中のいずれにあるかを特定するか,などの観点から研究が進められてきた.しかしながら,画像検索の研究が数多くなされているにも関わらず,十分実用に耐えうる検索技術が出現していないのは,画像が本質的に持つ意味の多義性のため,キーワード付けや内容把握などの実現が簡単ではないということが,要因であると考えられる.

一般的に,画像検索の際に重要となるポイントは3つあると考える.1つは自分のイメージする画像を探し出すための検索クエリー創出,2つ目はデータベースなどに蓄えられた画像に対し検査要求に対応させるための特徴量抽出,最後は両者のマッチングである.

人間は画像に対してキーワードのような言語的な認識だけでなく,意味的な認識と感覚的な認識とを組み合わせて記憶しており,また文字検索と異なり,検索の前に検索対象がはっきりしていないことも少なくない.つまり,画像検索の難しさは,検索目標であるにもかかわらずその目標画像に対する記憶があいまいであり,目標画像のイメージを正確に検索クエリーとして生成できないことに起因しているとも言える.したがって,高精度な画像検索の実現のためには,ユーザの検索要求をどのように入力し,システムがその入力をどのように解釈・処理し,結果をユーザに返すかというユーザインターフェースの観点からの積極的な検討が必要不可欠であり,特にユーザのクエリー生成をどのように支援するかは画像検索技術における大きな課題であると言える.

一方,ユーザからの検索要求に対し該当する候補画像を絞り込むためには,画像の特徴量を抽出し検索対象とさせるメタデータ生成が重要な技術となる.近年,マルチメディアコンテンツの検索用析データとして注目を集めている国際標準に MPEG-7 (ISO/IEC 15938 Multimedia Content Description Interface)[1]がある.この標準において記述するメタデータとしては,大きく分けて2種類ある.一つはローレベルなメタデータ(Visual, Audio),もう一つはハイレベルなメタデータ(MDS: Multimedia Description Schemes)である.前者は,画像や音声に関する特徴量を,PCなどを用い機械的なデータを抽出するもので,画像の場合,輪郭線などの形状や,色配置などの色情報などがあげられる.一方後者は,コンテンツの内容(タイトル,内容説明など)を手入力により記述するものである[2][3].後者を用いてメタデータ化を行う場合,撮影された情報や内容を書くことになるが,それらの情報を詳しく知っている者は,撮影者や編集者となる.この場合,詳しく知っているがゆえに,キーワードやタイトルなどの記述は被写体そのものの固有名詞を用いやすい.しかしながら,検索するユーザはそれらの固有名詞を知らない人がほとんどである.では,逆に万人が共通に認識できるキーワード(例えば,山,建物など)を付与した場合どうなるであろうか,検索のおける候補画像は膨大となり,データベースが大きくなればなるほど探し出すのは困難となる.メタデータ化は,ユーザから送られる検索クエリーに対し,万能性を追求すればするほど,候補映像として該当するが,それらが膨大になってしまうという現状がある.

ユーザの検索要求に対してどの程度の候補画像(正解画像)を返すかは,検索クエリーとして送られてくる文字情報や数値情報と,画像の特徴量として抽出される情報とのマッチング精度が重要となってくる.文字情報の場合,前記したように画像のもつ多義性のため,同じ被写体に対しても多くの表現(例えば,家,住宅,ホーム,建物,建造物など)があり,語彙階層を定義しなければならないなど,多くの課題がある.また,数値的な情報(色,明るさ,輪郭など)を用いた場合,人間の記憶の曖昧さを補うために,冗長性の範囲(閾値)を広く設定する必要があるが,その分候補画像が膨大となってしまうと言う弱点がある.

画像検索において,自分のイメージした画像を正確にクエリーとして生成するインターフェースと,画像が持つ特徴を的確に抽出するメタデータ化が最も重要であり,この両者が同じ思想の元に生成されれば,マッチングの負荷か軽減するとともに,検索効率向上に大きく寄与できると考える.

冒頭でも述べたようにデジタルコンテンツは爆発的な増加を今日も続けている.これら増加するコンテンツに対し,多くの検索技術が存在するがどれも万能ではない.そこで,本報告では現在存在する画像の検索効率向上に貢献することを目的として,画像の見え方の共通性に着目し,人間の経験値より共通的に認識される情報として,被写体の奥行き情報や,カメラ位置及び仰角といった3つのパラメータを用いた画像分類(フィルタリング)法について述べる.本方式を用いることで,従来からある検索技術の前段階として,ある程度,類似画像としてまとめることができると考える.また,自動的なメタデータ抽出の可能性やマッチング補正について考察する.

2. 画像検索の従来研究と問題点,分類の必要性

画像検索の方法としては大別すると,(A)テキスト入力型,(B)略画入力型,(C)オブジェクト選択型,(D)画像探索型に分けることができる.

上記分類(A)は,テキスト語句によるもので,任意名詞句や制限された名詞句,また印象語や感性語,動作語などをユーザが入力し検索を行うものである[4][5].

一方,上記分類(B)は,スケッチを描くことでクエリ

ーとするものである[6][7]. これらのスケッチによる検索は，ユーザの略画の描き方にも精度が依存するが，一般的に厳密な検索に用いることは困難である．

上記分類(C)オブジェクト選択型とは，例えば簡単な図形を矩形領域中に配置し，検索目標のイメージを構成したり，また検索目標のアイコンを配置，変形しユーザの目標画像を示したりするものである．

上記分類(D)画像探索型とは，あらかじめ検索したい画像を所有している場合に限って利用できる技術である．このような方式は検索目標の画像が厳密に用意できる点では，文章中からの単語の検索と同様高い検索精度が期待できるクエリー入力方式である．

しかしながら，これらの方式は，一般的な画像検索としてユーザが検索したいと考えている画像を参照画像として用意しなければならないことや，画像構成が限られてしまうなどの問題があり，さらに，これらユーザの直観的な入力に対して，支援できるメタデータ生成は困難を極める．さらに，検索クエリーと画像とのマッチングについては，クエリーとして与えられる言葉や数値データに対してどの程度冗長性のある候補画像をヒットさせるのかなど，多くの課題が残されている．

一方，ある程度の候補画像を絞り込む方法として，フィルタリングがある．これは，画像のもつ特徴量(色や形状など)を機械的に抽出し，それを用いて選別する方法や，日付やイベント毎に選別する方法などがある．しかしながら，検索の際，機械的な選別の場合は，人間の選別イメージと機械とが同期が取れるか否か，日付などによる場合，人間の記憶がそれらの情報を覚えているのかどうか疑問が残る．また，“風景”，“建物”さらには“人物”など，被写体に対して意味的に解釈した選別などは，人それぞれの主観に左右されてしまうため，万能ではない．一方，動画に関しては．ジャンルなどによる選別がある．しかし，ジャンルとしては主にテレビ番組に対するものとしての TV-Anytime Forum[8]や ARIB[9]などが有名であるが，一般画像に対してこれらのジャンルを用いることは，番組ではないことと，たとえ用いたとしても対応するジャンルはごくわずかしかないなど，ジャンルを用いた絞り込みは現実的でない．しかしながら，爆発的に増加する画像に対し，検索効率向上を図るためには，ある程度のフィルタリングを用い，分類しておくことが重要であることは容易に想像できる．

3. 画像の分類と特徴表現

人間が今まで撮影したことのある画像や，1 度見たことのある画像を探し出す際，頭の中でその画像をイメージする．「遠くに山があって，その手前に川が流れていて・・・・.」など考えるであろう．また，今まで1度も見たことの無い画像を検索する際も，これまで見たことのある画像と類似させてイメージするであろう．本研究のポイントは，これら万人が画お図に対してイメージする共通項を探し出し，それを用いた特徴表現を行い分類することにある．

人間それぞれのイメージから，共通項を見つけ出すことは，ばらつきの無いクエリー生成と，メタデータ生成が可能となり，画像検索精度の向上に大きく貢献できるはずである．

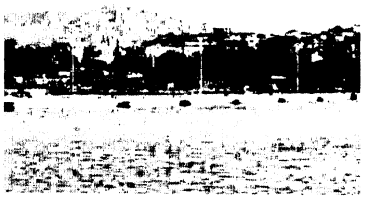

図 1　風景画像 1

図 1 に，一般的な風景画像を示す．これらからイメージされる特徴として，キーワードであれば「風景」「海（海岸）」「町並み」など思いつくであろう．しかしながら，これらのキーワードは前記したように，あまりにも高い語彙階層にあり，このままクエリーとすると，候補画像は膨大になる．

では，他に画像を見て分かる情報として考えると，

- かなり遠くに山がある．
- その手前に建物ある．
- カメラの高さ（撮影位置）は，ほぼ地上（水平線上）にある．
- カメラは水平（仰角は水平）に向いている．

などは容易に判断できるであろう．

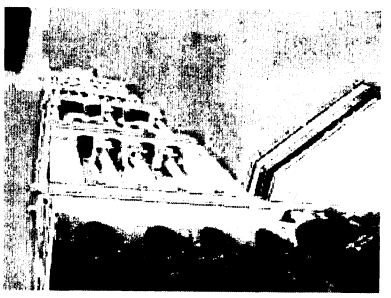

図 2　風景画像 2

では，図 2 の画像ではどうであろうか．

- 近くに建物がある．

・ カメラの高さは地上であろう．
・ カメラの仰角は水平よりプラスに大きい．
と感じるであろう．

図１，２共に建物が写っているがその見え方は異なる，この要因は，建物とカメラまでの距離とカメラの仰角が異なるためである．

これらの画像情報を整理して特長表現として用いることはできないか検討する．特徴表現するためには，万人が画像を見て共通の認識を持つ必要がある．図１および２で示した見え方に対して仮説を立てる．人間が画像を見て感じる（思う），共通的に認識可能な内容として，

（仮説１）被写体の距離感や位置関係
（仮説２）被写体に対するカメラの高さ
（仮説３）被写体に対するカメラの仰角

をあげる．これらの仮説が証明できれば，パラメータとして用いることで，その組合せで画像の特徴表現ができることになる．

4. 認識実験

表１に画像認識シートを，図 3 にその概要を示す．本シートは　画像に対して被験者が認識する（感じる）特徴の共通性を判断するもので，パラメータとしては，“被写体との距離”，“カメラの高さ”及び“カメラの仰角がある．

表１　画像認識シート

パラメータ	パラメータの値		
被写体との距離	0~10m	10m~1km	1km~∞
カメラの位置	0~2m	2~100m	100m~∞
カメラの仰角	＋	±0	－

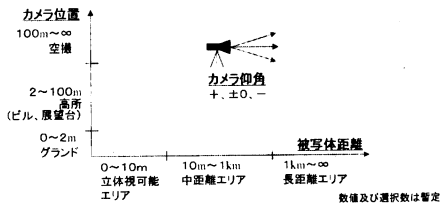

図３　各パラメータ値の概要

表１において，“被写体との距離”とは画像に写っている被写体がどれくらいの距離に認識するかを示すもので，全ての値にマークが入ることもある．ここで，シートの数値（被写体との距離）は筆者が暫定的に定めた数値ではあるが，被写体となるものの絶対的な大きさと，撮影した場合どの程度写るのかを想定し，各パラメータ値に対応する被写体は，あらかじめ想定してある．例を示すと，被写体との距離 0～10m では，人物，自転車などの大きさ，10m～1km では，家やビルなどの建造物，樹木などの自然物．さらに 1km～∞では，島，山などの地形に関するものなどがある．逆に言えば，人物が 1km 先に写っていたとしても，被写体の大きさと距離から計算するとほとんど写らないということである．“カメラの位置”及び“カメラの仰角”とは，画像より認識する位置と仰角であり，3 つより 1 つを選ぶことになる．実験にあたり５人の被験者に 10 枚の画像を提示した．また，被写体とは，人工物や自然の物体であり，空や海などは含まないとする．被験者による認識の例を図 4a~c に示す．

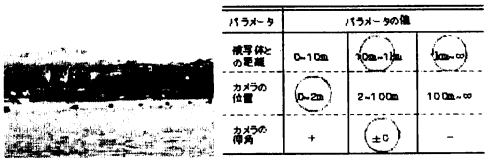

図 4a 提示画像と認識 a

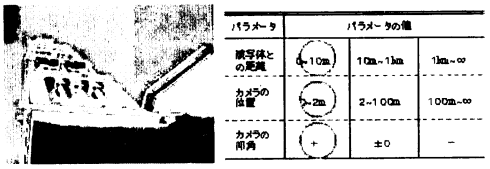

図 4b 提示画像と認識 b

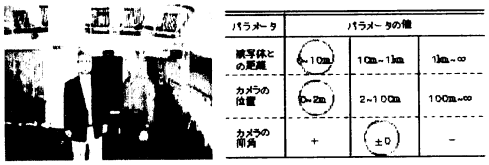

図 4c 提示画像と認識 c

本実験では，被験者が５人という少数ではあったが，５人ともほぼ同様の認識をしている．このことは，暫定的ではあるが，これらのパラメータによる認識が有効であることを示している．

5. パラメータについての考察

実験において，被写体の距離，カメラの高さや仰角について共通的な認識ができることが証明された．このジャンル分けは，画像の持つ意味的なものではなく，人間の経験値により見え方の認識を判断する客観的な方法を用いているため，人それぞれのばらつきがあまり無いと考える．このことは，クエリーを生成するユ

ーザと，画像の特徴としてメタデータを付与する記述者間で，共通的な認識を持つことであり，より正確なマッチングが可能と考える．一般画像であれば，接写や空撮なども含め，全てこのカテゴリ分けで整理することができる．もちろん，このカテゴリ分けのみでは意図する画像を検索することが不可能であるが，第2項で紹介した現在の検索技術における，候補画像の絞込みのサポートとして利用できると考える．

ここで，“被写体の距離”のパラメータについて全ての組合せを表2に，また，“カメラの高さ”と“カメラの仰角”のパラメータの全ての組合せを表3に示す．

表2　被写体と距離についての組合せ

	近距離 約 0~10m	中距離 10m~1km	遠距離 1km~∞	オブジェクトの存在
1	○	○	○	全て存在
2	○	○	×	近，中距離が存在
3	○	×	○	近，遠距離が存在
4	○	×	×	近距離のみ存在
5	×	○	○	中，遠距離が存在
6	×	○	×	中距離のみ存在
7	×	×	○	遠距離のみ存在
8	×	×	×	全て存在しない

表3　カメラの仰角と高さとの組合せ

	カメラ 仰角[注1]	カメラ 高さ[注2]	該当する画像
I	A	a	地上から上を見上げた画像
II	A	b	展望台などから上を見上げた画像
III	A	c	超高層ビルや空中で見上げた画像
IV	B	a	最も一般的な画像（記念写真など）
V	B	b	展望台などでの記念撮影
VI	B	c	地平線の空撮など
VII	C	a	花などの接写
VIII	C	b	展望台や建物から見る風景など
IX	C	c	町並みなどの空撮

注1 A：＋，B：±0，C：－

注2 a：0~2m，b：2~100m，c：100m~∞

表4　本方式による風景画像の選別

カメラの仰角と高さ（仰角 高さ）

仰角 A:＋, B:±0, C:－ 高さa:0~2m, b:2~100m, c:100m~∞

被写体の有無（○ ○ ○）＝（近 中 遠）	I (A, a)	II (A, b)	III (A, c)	IV (B, a)	V (B, b)	VI (B, c)	VII (C, a)	VIII (C, b)	IX (C, c)
1 (○○○)				15	6			2	1
2 (○○×)	15			30	22	1	1	6	1
3 (○×○)				10	2	1			
4 (○××)	22			65	2		32	6	
5 (×○○)				14					
6 (×○×)	4			38				4	
7 (××○)									
8 (×××)				22パターンに選別					

本稿のパラメータの組合せでは 72 通りの選別が可能となる．表4に筆者が個人で所有する風景画像300枚に対して，本方式を用いた選別を示す．一般的な風景画像に対して，本方式を用いることでばらつきはあるものの22パターンに選別できている．

6. メタデータ化とマッチングについての考察

これまで検討してきたパラメータの組合せを用いて検索クエリーとした場合，重要となるのは画像に対するメタデータ化である．これからメタデータを生成する場合，大きく別けて2つの方法がある．1つは既に撮影され蓄積されている画像に対するものと，もう1つはこれから撮影する新規の画像に対するものである．前者の場合，自動的なメタデータ化は非常に困難である．被写体の距離情報については，画像理解の範囲であり，空気遠近による色の変化や，輪郭線の形状などを考慮し何が写っているのかを判断すべく多くの研究[10]～[13]がなされているものの，カメラの仰角や高さなどはある程度手動での作業が必要なるであろう．

一方後者の場合は，少々乱暴ではあるがある程度の自動化が考えられる．カメラの仰角や高さなどは，高度計や水平器など，被写体までの距離はステレオ画像技術やレーザーによる測定も考えられる．

ここで1つの疑問点が生じる．被写体までの絶対的な距離が機械的に測定された場合，その撮影された画像を人間が見た時，絶対的な距離を感じるであろうか．もし，人間が違う距離感を認識するのであれば，本研究で検討しているパラメータによる分類はそのまま適用できない．そこで，被験者に対し画像のみを見て，被写体がどの程度の距離に煮えるのか実験を行った．実験に用いた被写体は車（被写体A）と7階建ての建造物（被写体B）といた，異なった大きさの物を用意した．実験に用いた画像として，図5Aに被写体Aを，図5Bに被写体Bを示す．

図5A　被写体A

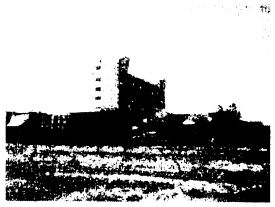

図5B　被写体B

図に示す2つの被写体に対し，それぞれ異なる距離から撮影した6枚の画像を被験者20人に提示し，それほどの距離に見えるのか実験を行った．それぞれの実験結果表5A，5Bに示す．結果として，被写体までの距離が5～10mや30～40mといった近距離よりも，遠距離になるにつれて実際と写真から感じる認識は，大きな差が生じている．

表 5A　実験結果（被写体 A）

	1-1	1-2	1-3	1-4	1-5	1-6
平均値(m)	6.615	17.462	33.615	57.308	92.692	128.461
標準偏差	3.279	12.366	20.774	30.387	44.141	67.28
実際の距離(m)	5	10	20	40	60	80
平均との差(m)	1.615	7.462	13.615	17.308	32.692	48.461

表 5B　実験結果（被写体 B）

	2-1	2-2	2-3	2-4	2-5	2-6
平均値(m)	51.538	77.692	134.615	212.308	312.308	490.769
標準偏差	27.868	35.391	73.441	134.793	190.27	291.103
実際の距離(m)	30	40	70	110	190	280
平均との差(m)	21.538	37.692	64.615	102.308	122.308	210.769

実験の結果を被験者ごと見てみると，実際の距離より写真から感じる距離の方が，遠く距離を示した者がほとんどであったが，逆に近い距離を示したものも数人いた．距離情報を用いた検索インターフェースを構築する際，ユーザーの特性を考慮する必要があると考える．今後は，あらゆる被写体や距離，太陽光と夜間など，多くのパターンでの実験を行う予定である．

7. まとめと今後の課題

本報告では現在存在する画像の検索効率向上に貢献することを目的として，画像の見え方の共通性に着目し，人間の経験値より共通的に認識される情報として，被写体の奥行き情報や，カメラ位置及び仰角といった３つのパラメータを用いた画像分類(フィルタリング)法について述べた．本方式を用いることで，従来からある検索技術の前段階として，ある程度，類似画像としてまとめることができると考える．加えて，実際の被写体との距離と，写真から感じる距離感の相違について実験を行い，写真から感じる被写体の距離感は，実際のものよりも遠く感じることが結果として抽出された．今後の課題としては，被験者を増やしより正確なデータの収集や，パラメータの選択肢の増減による正確性の検証，メタデータ化とマッチングの検討を行う予定である．さらに，動画への応用も検討していく．

謝辞：本研究は総務省戦略的情報通信研究開発推進精度研究主体育成型研究開発平成 15 年度「簡単映像コンテンツ制作のための高度映像検索技術に関する研究（研究開発）」（代表者，青木輝勝）の一環として行われたものである．

文献

[1] ISO/IEC FDIS 15938-5 : “ Multimedia Content Description Interface - Part 5: Multimedia Description Schemes”, JTC1/SC29/WG11/ N4242 (Oct. 2001)

[2] 堀，“MPEG-7 の概要と役割”，情報処理学会シンポジウムシリーズ，2001，10，pp3-17（2001

[3] 柴田，“MPEG-7 MDS チュートリアル”，情報処理学会シンポジウムシリーズ，2001，10，pp27-43（2001）

[4] 宮森恒,粕谷英司,富永英義,"動作語を用いた問い合わせによる映像検索方式",信学論 D-II Vol.J80-D-II No.6, pp.1590-1599, June,1997

[5] 原田将治,伊藤幸宏,中谷広正,"感性語句を含む自然言語文による画像検索のための形状特徴空間の構築",情処学論 Vol.40, No.5, May.1999

[6] 黒川雅人,洪政国,"形状情報を用いた画像の類似検索システム",情処学論 Vol.32 No.6,June 1991

[7] 椋木雅之,美濃導彦,池田克夫,"対象物スケッチによる風景画像検索とインデックスの自動生成",信学論 D-II Vol.J79-D-II No.6, pp.1025-1033, June,1996

[8] TV-Anytime Forum, http://www.tv-anytime.org/

[9] ARIB, http://www.arib.or.jp/

[10] 美濃導彦,岡崎洋,坂井利之,"対象物の属性情報による検索法—風景画像中の山を例として—",情処学論 Vol.32, No.4, Apr,1991

[11] 高橋友一,島則之,岸野文朗,"位置情報を手がかりとする画像検索法",情処学論 Vol.31, No.11, Nov,1990

[12] 北川高嗣,中西崇文,清木康,"静止画メディアデータを対象としたメタデータ自動抽出方式の実現と意味的画像検索への適用",情処学論 Vol.43, No.SIG 12, Dec,2002

[13] 武者義則,広池敦,"画像群の意味的な可視化表現を用いた画像検索システム",映情メ学論 Vol.54, No.12, pp.1742-1747, 2000