A Survey on Content-Based Retrieval for Multimedia Databases

Atsuo Yoshitaka, *Member*, *IEEE*, and Tadao Ichikawa, *Fellow, IEEE*

Abstract—Conventional database systems are designed for managing textual and numerical data, and retrieving such data is often based on simple comparisons of text/numerical values. However, this simple method of retrieval is no longer adequate for the multimedia data, since the digitized representation of images, video, or data itself does not convey the reality of these media items. In addition, composite data consisting of heterogeneous types of data also associates with the semantic content acquired by a user's recognition. Therefore, content-based retrieval for multimedia data is realized taking such intrinsic features of multimedia data into account. Implementation of the content-based retrieval facility is not based on a single fundamental, but is closely related to an underlying data model, a priori knowledge of the area of interest, and the scheme for representing queries. This paper surveys recent studies on content-based retrieval for multimedia databases from the point of view of three fundamental issues. Throughout the discussion, we assume databases that manage only nontextual/numerical data, such as image or video, are also in the category of multimedia databases.

Index Terms—Multimedia databases, content-based retrieval, spatio-temporal relation, query-by-example, knowledge.

1 Introduction

Owing to the recent progress of hardware, managing large amounts of image, video, audio data, or a combination of them, has become ordinary. Growing needs of retrieving the contents of such data is a natural conclusion of the requirement for database systems. However, conventional database systems, most of which are based on the relational data model, are often pointed out as not providing enough facilities for managing and retrieving the contents of multimedia data for the following reasons:

1) First, object-oriented models as well as relational data models lack facilities for the management of spatio-temporal relations. This, in some sense, relates to the second reason described above. Audio and video data essentially imply a temporal aspect. It means that temporal relations (i.e., synchronization) between pieces of video data can be an element that needs to be managed by a database system. When we consider the case where text data that is superimposed onto video data is stored separately from the video data, spatial relations as well as temporal relations need to be managed to define the relation between them. In the case of image data, especially geographical databases that store geographical maps together with entities such as buildings, electric cables, sewers, and water pipes as separate entities, the data representing the entities should be stored with their spatial relationships. As observed in the above examples, the ability of managing spatio-temporal relations is one of the important features for multimedia database systems.
2) Second, the recognition and/or interpretation process of contents in multimedia data are often inevitable in retrieval, since representation of image, video, or audio is one thing and contents perceived is another. In order to evaluate the contents that are associated with the semantics of the data being retrieved, a database management system requires knowledge for interpreting raw data into the contents implied. Knowledge-assisted retrieval is also studied in the area of textual databases [40], [52]. However, it plays a more important role for multimedia database retrieval because even a single media data has many faces of meaning/contents, one of which should actually be referred to in query evaluation depending on a context of retrieval.
3) The last thing concerns query representation. Since a relational database assumes text and numerical data as its domain, the retrieval of records is basically composed of relational algebra or descriptions of simple comparison of attribute values with query conditions in the form of alphanumeric representation. In contrast to this, in multimedia database retrieval, textual (numerical) expression of query conditions are not always appropriate since the type of contents derivable from multimedia data is diverse. Query-by-example (QBE), where the form of representation is closer to that of data to be retrieved, would be a better solution in content-based retrieval because it expresses query condition more naturally than words.

As summarized above, conventional database systems do not provide sufficient flexibility in managing data, because of the inability to manage spatio-temporal relations, to recognize contents in multimedia data which relates to recognition of semantics of the contents of media data, and to allow various types of query representation based on

- *A. Yoshitaka and T. Ichikawa are with the Faculty of Engineering, Hiroshima University, Kagamiyama 1-4-1, Higashi-Hiroshima, Hiroshima 739-8527, Japan. E-mail: {yoshi, ichikawa}@huis.hiroshima-u.ac.jp.*

Manuscript received 16 June 1997; revised 29 Sept. 1998.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 108309.

1041-4347/99/$10.00 © 1999 IEEE

QBE for enabling intuitive representation of query condition. These three deficiencies relate to all of the components of database system. That is, the first issue depends on the data model or indexing, the second on DBMS construction, and the last relates to the user interface. Therefore, none of them can be neglected for discussing all possible cases of content-based retrieval.

In this paper, we survey recent studies related to content-based retrieval for multimedia databases, and show directions for the solution of these issues. First, we discuss recent studies from the point of views of the above-mentioned three aspects. After that we will show that all of these three aspects should be considered for covering widely spread bandwidth of multimedia content retrieval.

The organization of this paper is as follows: In Section 2, we describe data model and indexing issues related to content-based retrieval. QBE for multimedia database is discussed in Section 3. After that, knowledge-assisted content-based retrieval is discussed in Section 4, then we discuss the type of contents covered by above-mentioned three fundamentals and related issues in Section 5. Concluding remarks are given in Section 6.

2 CBR ON SPATIO-TEMPORAL RELATION

2.1 Overview

As mentioned in the previous section, the relational data model does not cover all features required for multimedia database retrieval. As a data model that provides a system with better facilities for the management of multimedia data, the object-oriented data model [2], [43] has been proposed. The idea of object-orientation is to encapsulate data with a set of operations that are applicable to the data. This framework provides a system with operational transparency. That is, the user does not need to be careful about the heterogeneity of operations caused by different type of data for the purpose of manipulating data but rather the need to send the same message to the different types of data for semantically identical operations. A composite object is considered to be an object that consists of other objects. It enables one to define the part_of relationship among objects which takes an arbitrary structure.

In addition to the structural complexity and operational transparency of multimedia data, spatial and/or temporal dimension is inherent in image, video, and audio data. However, the core of the object-oriented data model does not contain a facility for managing spatio-temporal relationships. In the following subsections, we describe studies on the management of spatial and/or temporal relations.

2.2 Spatial Relations

Managing spatial relation is one of the mandatory feature in many multimedia applications. One of the straightforward way of managing the spatial relation between components of information is to represent it by rectangular coordinates. The spatial position of a component object is represented by coordinates and the relation between components are calculated mathematically as in [19], [32]. Multimedia documents which consists of images, charts, and graphics as well as text is another example that requires the management of spatial relation for layout information [12].

In other applications such as GIS (geographical information system), the representation and indexing of abstract spatial relations is studied. A 2D string [8], [11] is an indexing technique for representing a spatial relation between the components of a picture; 2D strings represent abstract position of components, which consists of horizontal and vertical order of components. In addition, it represents several levels of a coarse-strict relation, where the strictness of direction differs from one level to another. Liu and Sun [32] permits coarse representation of spatial relation of objects as well as strict representation in rectangular coordinates. Meanwhile, other representations of spatial relation are also studied.

In [38], [39], a set of binary relations such as ‘left of,’ ‘right of,’ ‘in front of,’ ‘behind,’ ‘above,’ ‘below,’ ‘inside,’ ‘outside,’ and ‘overlaps’ is defined as primitive relations for representing spatial relation of pictorial components. This approach as well as 2D string representation is suitable for coarse evaluation of spatial relation. The advantage of this approach is, of course, that it can ignore subtle difference of relations that need not to be evaluated. However, since the relation is distance-independent, interval-oriented contents are out of their scope.

2.3 Temporal Relations

Recent studies related to representation and management of temporal relation assume video related applications such as VOD or video databases. There are two main approaches for representing temporal relations between multimedia objects: One is a point-based representation, and the other is an interval-based representation. The point-based representation represents the position of objects by points on timeline, whereas the interval-based representation represents the relatedness of objects by means of the intervals of their occurrences. Most of the studies that manage temporal relation between component video objects are based on the interval-based model.

In OVID [36], video objects are defined as a sequence of video objects represented by intervals, and a video object may consists of several sequences of continuous video frames. The modeling provides two operations, i.e., merge and overlap, for manipulating video objects. Textual annotation that denotes the contents of the video object as part of the definition of each video object is also merged/overlapped. Querying a video object in OVID enables both frame-based and interval-based specification.

Gibbs et al. [17] takes a timed Petri net based representation of component objects. This study considers temporally sequential representation of component objects. Therefore, it cannot give a natural representation of a condition of temporally overlapped components. Meanwhile, [1] discussed interval-based temporal relations; *before*, *equal*, *meets*, *overlaps*, *during*, *starts*, and *finishes*. Little and Ghafoor [30] extended basic binary temporal relations discussed in [1] to *n*-ary temporal relations, and discussed reverse relations of them. Hopner [24] also follows Allen’s definition of basic temporal relations. In his approach, a document is defined with a tree structure, where a parent node contains the definition of temporal relations.

Interval-based temporal relations referred to in the above studies are close to editing/construction of a story of video frames and more perceptible to human compared with point-based representation. That is, the representation is desirable for retrieval of temporal relations of components in the area of video databases. However, interval-based temporal relations can be translated into point-based relations or vise versa. Comparison of an interval-based query with a point-based query is discussed in [42].

2.4 Spatio-Temporal Relations

Regarding the representation of spatial and temporal relations, there are two alternatives; representing them in a consistent way and representing them independently. Iino et al. [26] is an example of the former, where the temporal relation is primary and the spatial relation is secondary. Objects are structured with temporal relations between component objects as proposed in [1], and spatial composition operations such as *overlay*, *overlap*, *abut*, *crop*, and *scale* are defined as spatial relations. With this approach, the spatial composition operations are not consistent with the temporal relations, since the spatial composition operations are assumed to be applied to multimedia data presentation.

Day et al. [13] is another work done by the same research group as [26], where both spatial and temporal relations are based on a single set of interval-based primitive relations. An advantage of this approach is that content based retrieval on both spatial and temporal relation is realized in a unified manner. Theodoridis et al. [41] also discussed spatio-temporal indexing of multimedia objects in an integrated manner, where their approach assumes multimedia presentation. With regard to the indexing, two approaches have been proposed: One is separating the spatial index by a 2D R-tree from the temporal index by a 1D R-tree, and the other is integrating spatio-temporal indexing by 3D R-trees.

As introduced above, there are not many studies of the spatio-temporal modeling of multimedia data. Whether or not a model provides consistent representation for both spatial and temporal relation depends on the application. There is still room for further discussion.

3 Query Representation for Content-Based Retrieval

In querying conventional databases consisting of text and/or numerical data, a query condition is often represented in the form of text or a numerical value. This approach is always a proper way of specifying a query condition for multimedia databases. Suppose there is an image database that consists of paintings or graphics. In retrieving such data, a user may specify a query condition in the form of text, e.g., paintings with 'blue sky' or graphics that contain 'an oval.' However, keywords representing graphical features or image attributes such as color represent subtle differences in an intuitive way. In the case where the degree of blue or the angle of the long axis of the oval to be retrieved is also a part of the query condition, we think giving a color example or graphical example is better than words.

Query-by-example is considered to be a promising approach since it provides a user with an intuitive way of query representation and the form of expressing a query condition is close to that of the data to be evaluated. In this section, we classify QBE approaches to multimedia databases by means of target data that is either image, video, or audio.

3.1 QBE for Image Retrieval

'Query-by-example' [53] is another method of query specification, which allows a user to specify a query condition by giving examples.

A query condition specified in 'query-by-example' assumes representation similar to the feature of objects being retrieved, which means that the query condition is directly comparable with the objects. Prior to the comparison, either one or both of the condition and data may be preprocessed in order to emphasize the contents (i.e., remove noise or unnecessary elements) being compared.

Compared with representing a query condition by a keyword, there are several advantages. First, 'query-by-example' provides a user with an intuitive way of representing his/her constraint in mind since the representation of a query condition corresponds to features of the data. Second, 'query-by-example' for nontextual data is often a better way of representing a query condition in the sense that an example can represent subtle difference more easily than representing it by words or numerical values. This is simply because the form of representation is the same or at least close enough to the features of data to be retrieved.

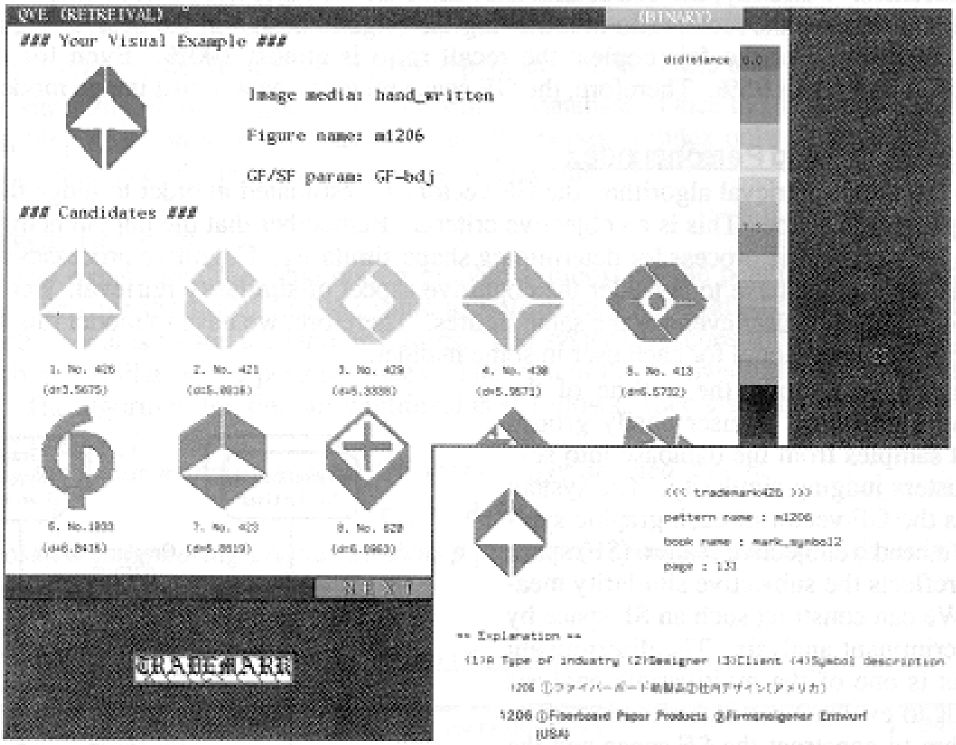
Most QBE studies concentrate on image retrieval compared with those for video or audio. QBE for image databases is further classified in terms of the feature to be specified as a query condition.

3.1.1 Shape

An image object is retrieved by evaluating the shape of objects. There are two forms of specification: One is to give a photo/graphic of the object in the database that contains the shape to be retrieved, another is to draw the shape by a user. The system described in [27] (see Fig. 1) is designed for the retrieval of graphical trademarks of companies. A query condition is specified by giving a hand written graphics or a registered trademark. Bimbo et al. [6] (see Fig. 2) demonstrated the retrieval of the object of an arbitrary shape for pictorial databases. This also accepts an example shape of the object in a painting by cut-and-paste as well as giving a hand-drawn shape.

3.1.2 Spatial Relation

QBE with spatial relation specifies one or more component objects as an example, where the evaluation of spatial relation among them are more important than color or shape. Geographical databases are the application typical of taking this type of QBE approach. Haarslev and Wessel [22], Meyer [34], and Egenhofer [14] discuss retrieval of geographical

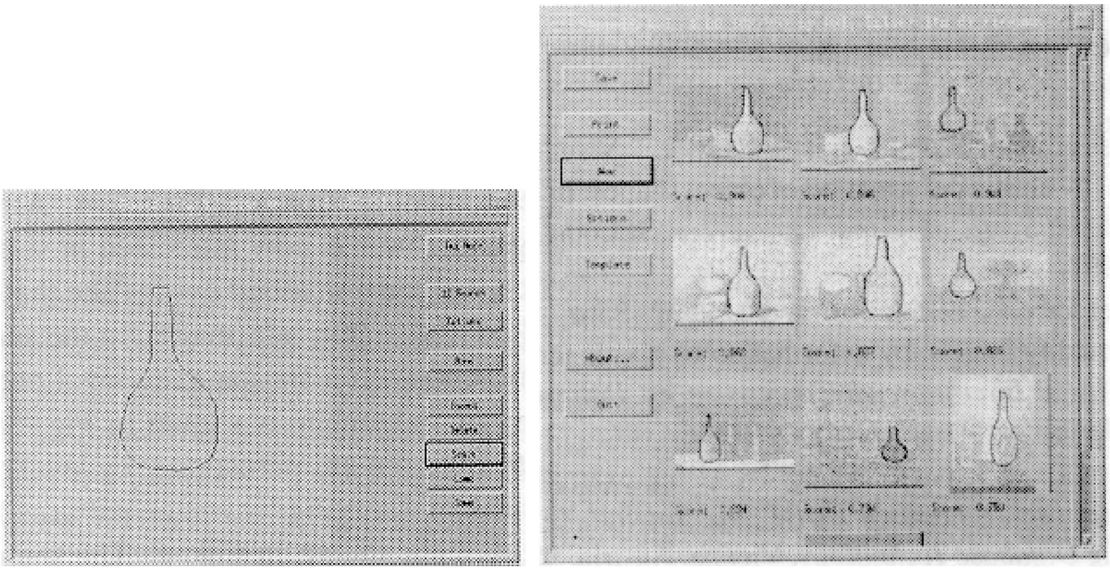


```
QVE (RETREIVAL)                                   (BINARY)
### Your Visual Example ###
                         Image media: hand_written
                         Figure name: m1206
                         GF/SF param: GF-bdj
### Candidates ###
1. No. 426     2. No. 421     3. No. 429     4. No. 430     5. No. 413
(d=3.5675)     (d=5.0016)     (d=5.3338)     (d=5.5871)     (d=6.5732)
6. No.1003     7. No. 423     8. No. 620
(d=6.8416)     (d=6.8619)     (d=6.0963)

<<< trademark426 >>>
pattern name : m1206
book name : mark_symbol2
page : 131

** Explanation **
(1)A Type of industry (2)Designer (3)Client (4)Symbol description
1206 ①ファイバーボード紙製品②社内デザイン(アメリカ)
1206 ①Fiberboard Paper Products ②Firmendesigner Entwurf
     (USA)
```

Fig. 1. An example of query-by-example (QBE) with a sketch [27].



(a)

(b)

Fig. 2. Another example of QBE with a sketch [6]: (a) an example drawing as the query condition; (b) the result of the retrieval.

databases where geographical objects such as buildings, rivers, and so on are depicted as an example. An example which consists of spatially placed objects is regarded as a query condition on the spatial relations of objects, and topological/spatial relations among them are evaluated. The topological relations considered in [14] are *disjoint*, *meet*, *overlap*, *contains*, *covers*, *inside*, *covered-by*, and *equal*. They can be integrated with interval-based temporal relations as discussed in [13].

3.1.3 Color

Images are retrieved by specifying colors and their spatial distribution in the image. This way of specification is often applied for the retrieval of paintings by principal colors with spatial relation. Examples of systems capable of processing this type of query are [10], [47], [18]. Gong et al. [18] allows us to specify an example image (digitized photo) as the example of colors with spatial distribution, where a representative plane is divided into nine (3 × 3) subplanes.

QBE by color is often accompanied by the specification of spatial relations of color components specified. Therefore, strictly speaking, the scheme of query specification in these studies is categorized into that of specifying the combination of spatial relation and color. In query evaluation, a color difference as well as the spatial distribution is valued in order to show the degree of satisfying specified conditions.

3.1.4 Texture

Texture is specified in order to retrieve a specific pattern appearing in an image. This is used for retrieving an object that shows a certain texture on its surface. This form of specification is applied for retrieving an object, one of whose property is the texture. QBIC [15] makes the query-by-texture feasible as well as three ways of specification stated above.

3.2 QBE for Video Retrieval

3.2.1 Motion of Object

QBE for image data becomes applicable to video databases as well, if we can regard a piece of video data as a set of images whose temporal interrelation is not interested. Specifying a sketch for retrieving a frame in video streams cannot always be proper, since this method of query specification loses information specific to video data, i.e., spatio-temporal relation/contents. QBE with motion examples is an approach to retrieve intrinsic features of video data, i.e., the motion of objects appearing in video.

The authors have implemented a system which makes retrieval of video data possible by specifying the motion of an object observed in video data by giving an example [50] (see Fig. 3). An example motion of an object is specified by making a mouse move, and then a trajectory and velocity are sampled in accordance with the movement. In addition, changing the size of an object is specified by drawing rectangles along with the timeline.

Another example of QBE with motion example is presented in [9]. This work also allows a user to specify trajectory, duration, and scaling as well as the basic feature of image such as color, texture, and shape. The proposed framework allows us to specify an example consisting of multiple objects. However, specifying a number of objects as an example is too complicated.

3.2.2 Spatio-Temporal Relations

Bimbo et al. [4] (see Fig. 4) presented an interesting method for expressing spatial relations of objects in an image. The system demonstrates the retrieval of a 2D image in terms of giving spatial relations of an object in a 3D space. Though the world projected into photos is a 2D representation, it was originally a 3D space. The method proposed in the literature provides a user with a way to represent spatial relations of objects in a 3D space with a data glove, which is recognized as an example of spatial relations among objects and translated into a 2D representation. Allowing a 3D representation of an example which is close to human's way of memorization is considered to be an effective way of representing an example even for a 2D image retrieval.

The QBE with motion example introduced above is concerned with the motion of an object by three elements, namely trajectory, velocity, and size of the object. This is another approach for the retrieval of spatio-temporal contents [5]. With this approach, an example is regarded as a sequence of spatial relations. A user specifies positions of objects on a screen from which the system extracts spatial relations at a certain point of time. The user defines two or more sets of spatial relations sequentially in accordance with the time, representing spatio-temporal relations of the objects.

This approach is not aimed at the retrieval of the detailed motion of an object, but is aimed at spatio-temporal correlation of multiple objects, since the motion of an object is specified on discrete time.

3.3 QBE for Audio Retrieval

QBE experiences for audio data are much less than those for images or video databases. One of the reasons is that there are difficulties in recognizing or extracting the contents of an audio object. However, some studies have presented QBE applications to audio data.

Ghias et al. [16] is one of the novel approach of QBE for audio data, specifically, the musical data of songs. A person humming is entered to the system as an example of a musical phrase. The system retrieves the melody of songs which match the given example. Humming as the way of representing a query condition for music retrieval is quite intuitive to a general user. Specifying melodies in other forms of representation such as a score is not easy.

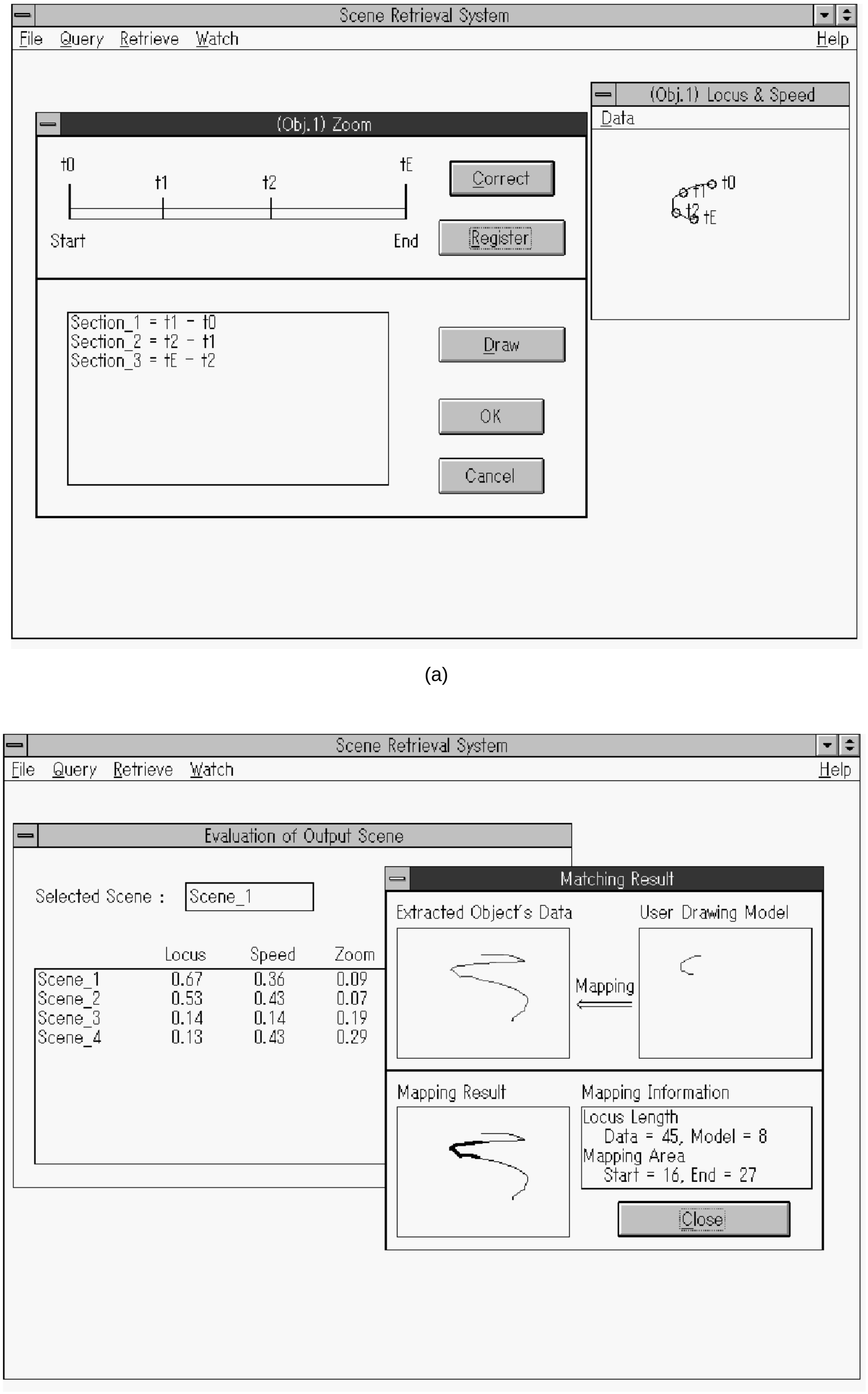
Wold et al. [45] discusses content-based retrieval of generic audio data, by means of QBE as well as the retrieval by attributes attached to sound data such as keyword, sampling rate, or date of creation. It takes the approach of extracting basic elements called an *analysis feature vector*, which consists of duration, pitch, amplitude, brightness, and bandwidth. Melih and Gonzalez [33] describes an on-going study of content-based retrieval of audio, which mainly discusses an audio signal processing method of extracting features for evaluating peculiarity of individual sound.

Content-based retrieval of audio is an area less mature than content-based retrieval studies for image or video. This comes from the difficulty of extracting features of audio which clearly shows the peculiarity of individual sound source.

4 Knowledge-Assisted Content-Based Retrieval

QBE discussed in the last section is one of the promising scheme for representing query conditions for multimedia database retrieval in a natural and an intuitive way. Since a query condition in QBE is the representation of an example that a user wish to retrieve, semantics of the data are not analyzed and processed by a database system during the process of query evaluation.

However, there are also cases where a database is queried by specifying semantic contents. Here, we call this type of query specification *query-by-subject*, which allows the user to specify a subjective description of a query



Section_1 = t1 - t0
Section_2 = t2 - t1
Section_3 = tE - t2

(a)

Selected Scene : Scene_1

	Locus	Speed	Zoom
Scene_1	0.67	0.36	0.09
Scene_2	0.53	0.43	0.07
Scene_3	0.14	0.14	0.19
Scene_4	0.13	0.43	0.29

Locus Length
Data = 45, Model = 8
Mapping Area
Start = 16, End = 27

(b)

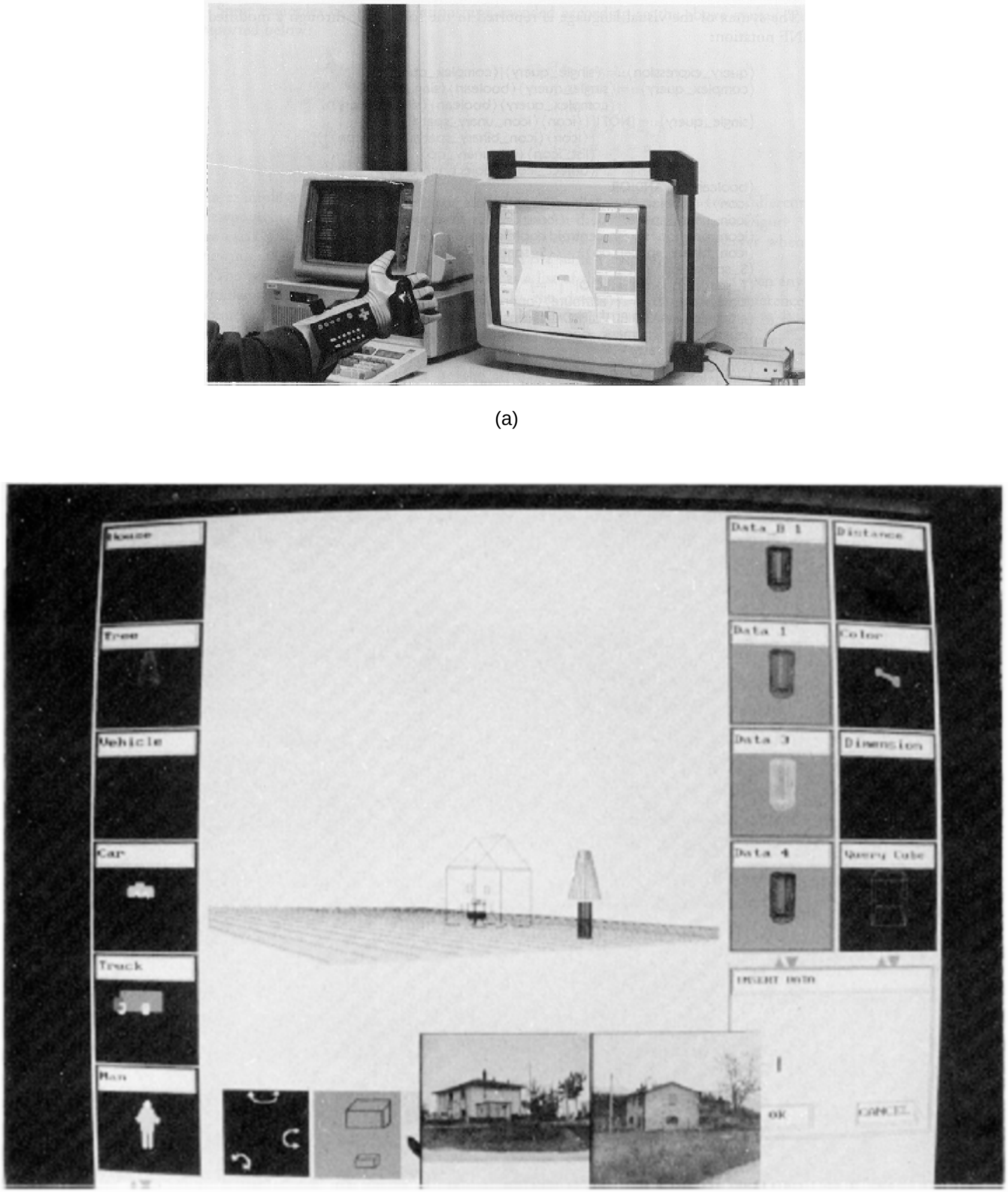
Fig. 3. An example of QBE with motion example [50]: (a) specifying a trajectory, velocity, and the size of a moving object; (b) matching evaluation.

condition. In such cases, knowledge is required to capture the semantic contents of multimedia data as well as to interpret the query.

In this section, we concentrate on discussing knowledge-assisted methods for extracting and managing contents in processing 'query-by-subject.'

4.1 CBR by Descriptive Knowledge

In 'query-by-subject,' a keyword representing a semantic content is specified. The semantic content implied in multimedia data is extracted from raw data in order to evaluate a query. A simple way of managing the semantic content



(a)

(b)

Fig. 4. An example of QBE for spatial relations [4]: (a) system overview; (b) specified query (center) and the result (bottom).

of multimedia data is to annotate an image, a video or an audio data with text. Retrieving the contents of an image by referring to the annotation is described in [28], [23], [20], [21]. Klinger and Pizano [28] assumes multimedia geographical information, where geographical entities in a map such as cities and roads are defined together with a textual description representing names of the cities or the roads, based on a kind of ER model.

Examples that provide content-based retrieval for a video database through textual annotation are [44], [31], [36]. In these studies, a textual description representing semantic contents is assumed to be defined for an image or a video data by a human. These studies focus on ‘retrieval-by-content’ which cannot be extracted from an image or a video data through image processing. Such kinds of information include, for example, the name of a road in a map, the name of a person appearing in a news video. In these systems, content-based retrieval for image/video data is internally replaced by a keyword retrieval for annotations. One of the advantages of this method is that it can easily be implemented. Another advantage is that misevaluation of contents will hardly occur, which is often one of the issues of content-based retrieval extracting contents directly from raw data such as video or audio. However, this approach is not practical especially in large multimedia databases.

4.2 CBR by Derivation Knowledge

Another method of implementing 'query-by-subject/object' is to provide the system with a rule base or a knowledge base. The knowledge-based approach for text databases has a long history than that for multimedia databases. In that area, the main interest is not only in generating cooperative answers [46] but also in retrieving semantic contents implied in the text [29], [7]. In the area of CBR for multimedia databases, [49], [35] study the content-based retrieval of images using knowledge that interprets semantic contents into image representations. The content to be retrieved concerns the meaning of an image that is represented by a keyword. Under the assumption that objects appearing in the image are already known, semantic contents that relate to the attribute value of the objects are also in the target of retrieval [3].

In [49], *Domain Knowledge* defines the method for extracting semantic features from multimedia data such as, images, rules to transform a certain operator into content-dependent calculus, and rules for transforming query conditions into an internal representation whose type is the same as the extracted semantic features. A definition of *pseudo attribute* associates a query condition with a domain knowledge describing the contents to be retrieved(Fig. 5). In [35], a keyword is defined with a description that denotes image features such as regions of colors and their location. A state transition model defines a hierarchical relation between primitive color regions and semantic contents represented by the spatial composition of color regions. With these approaches, images are retrieved by a content based on image features specific to the content.

Hsu et al. [25] is also an example applying knowledge-based CBR to a medical image. Knowledge is referred to for the evaluation of shapes and spatial relations of objects (in the application, it is tumor), and image semantics. Knowledge for interpreting contents on spatial relations or semantic contents is constructed by Type Abstraction Hierarchy (TAH). TAH defines a general level of concepts to detailed level with sets of attribute values.

In these studies, semantic contents are represented by knowledge directly, in the sense that the knowledge associated with feature values and/or spatial relations are the fundamental property of the contents. In contrast to this knowledge-based CBR, there is another approach of defining knowledge for the subject of interest indirectly [51].

In [51], the semantic expression of the contents of a scene, such as the scene of conversation or the scene of tension, are associated with camera framing and/or editing technique commonly applied by film directors or editors, which is the so called "film grammar." Note that camera framing and film editing does not directly represent semantic contents such as conversation or scene on tenterhooks, nor tries to extract from the database system the faces of person for evaluating a conversation scene. That is, such semantic contents themselves are not extracted from video data, but the editorial technique is extracted to evaluate the semantic contents. Prior to the extraction of scene features, the process of cut detection [37] or extraction of scene unit [48] is mandatory for automated parsing of video data.

4.3 Issues on Knowledge-Based CBR

In this section, we will summarize issues related to the knowledge-based retrieval and 'query-by-subject.'

As introduced above, one way of implementing 'query-by-subject' requires preparation of annotations in association with image, video and/or audio data. As explained, this method is often adopted when it is very hard to extract/recognize the target contents from image, video, or audio data. Therefore, the annotation is often created by humans.

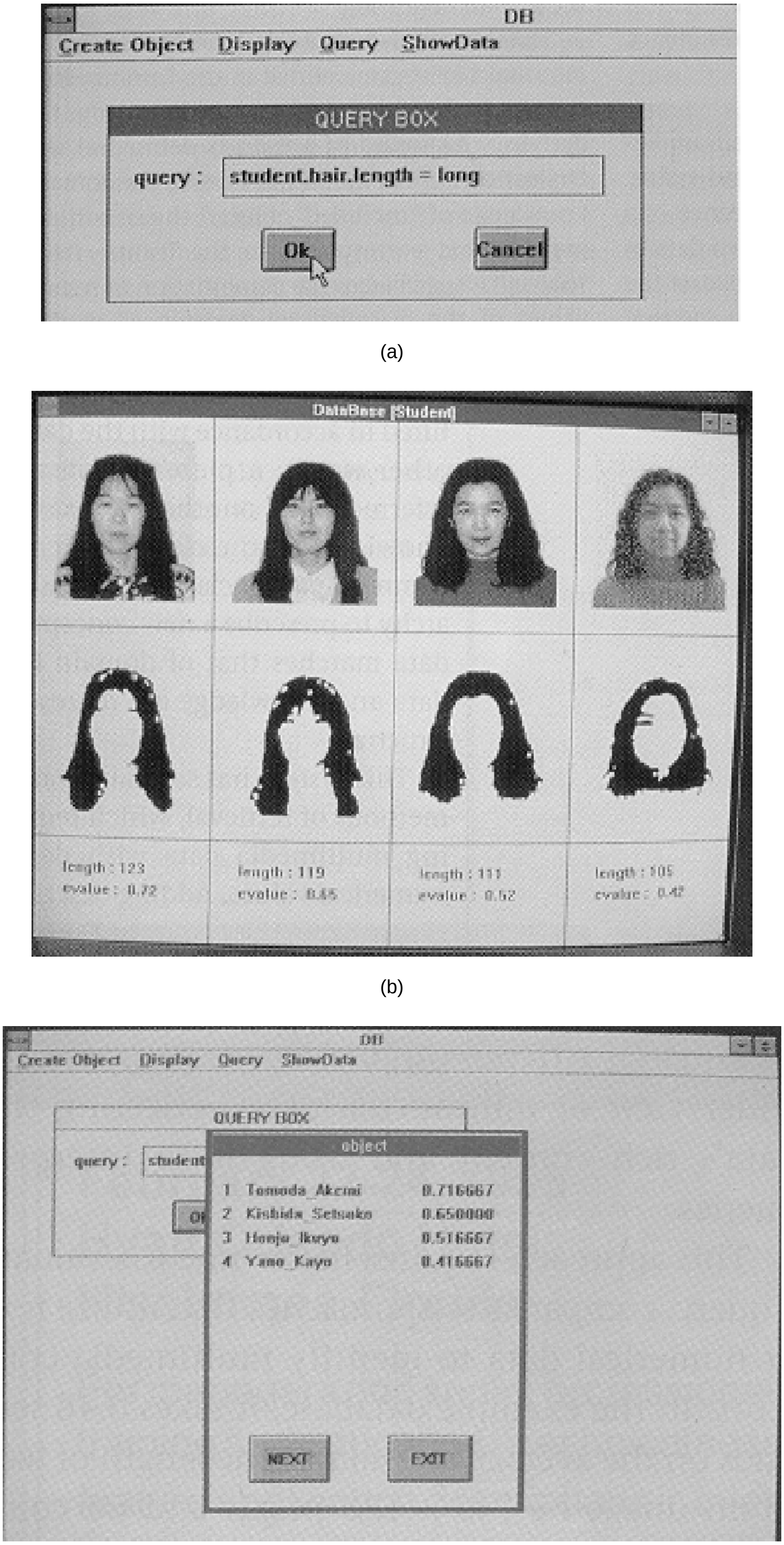
There are two main problems with this approach. One is that this approach is not practical especially in large databases, as long as annotations are made by human. Under an assumption where raw data accompanied by annotations are frequently updated, nor is this approach practical even in a relatively small databases. Another problem will arise when an annotation associated with the target data is represented by the degree of an attribute. In such a case, keeping the consistency of annotations is quite a difficult problem due to the characteristics of annotations defined by human. In addition, such a consistency, i.e., consistent criteria for judgment, needs to be well managed throughout database evolution.

Another approach is to provide a system with a knowledge/rule. Knowledge is used for feature extraction from raw data, content matching, query analysis and translation, and so on. This approach is more practical than the former even in large databases. One of the issues in knowledge-based CBR relates to knowledge for content evaluation and query translation, which is referred to, to interpret a query condition into semantically equivalent expressions and evaluates the expressions with features extracted from raw data. Keeping the knowledge base semantically consistent with the database schema is a difficult and important issue. A solution to maintaining consistency between the database and the knowledge base is to make them semantically dependent on each other by integrating them together with rules that prescribe semantic association of one with the other. This also includes the issue of DB schema evolution caused by KB evolution and vise versa.

5 Taxonomy of Contents and Coverage of Contents by Basic Frameworks

In this paper, we consider a multimedia database in a broad sense: We regard a database, which at least stores either one of image, video, or audio data, as a multimedia database. Following this standpoint, we include image databases or video databases, for example, into the category of multimedia databases.

In order to show how the three basic frameworks discussed in Sections 2, 3, and 4 relate to content-based retrieval, we first show the taxonomy of contents based on two aspects, i.e., homo-/heterogeneity of evaluated media and the type of contents evaluated.



query : student.hair.length = long

(a)

length : 123
evalue : 0.72

length : 119
evalue : 0.65

length : 111
evalue : 0.52

length : 105
evalue : 0.42

(b)

```
1  Tomoda_Akemi       0.716667
2  Kishida_Setsuko    0.650000
3  Honjo_Ikuyo        0.516667
4  Yano_Kayo          0.416667
```

(c)

Fig. 5. Knowledge-assisted retrieval from image [49]: (a) query specification (the attribute 'hair.length' is not defined in database schema); (b) intermediate result of extracting the region of 'hair'; (c) presenting the result of the query.

5.1 Taxonomy of Contents

5.1.1 Homo-/Heterogeneity of Component Media

This classification concerns homo-/heterogeneity of data forming the content. As stated above, we include databases that manage at least one of nontextual/numerical data such as image, video, or audio data into the category of multimedia database. This is based on an observation that in many cases retrieval referring only to image or video (without any reference to an audio track) is often discussed in the area of multimedia database retrieval. Therefore we first classify contents into two classes on the basis of homo-/heterogeneity of the data forming the content.

a) **Single media contents**. This type of content is represented by a homogeneous type of data: Contents classified into this category are retrieved by referring to homogeneous types of data. It means that even retrieval of multimedia data (in a narrow sense of multimedia data implying a database consisting of heterogeneous types of data) is regarded as the retrieval of a single media content if only homogeneous types of data is referred to in the query evaluation.

An example of this type of content is an object that is contained in an image data. This type of content includes those constructed by two or more pieces of homogeneous type of data. Most of the studies of content-based retrieval for multimedia databases focus on the retrieval of this type of contents.

A media object, such as image, video, and audio data, may have a number of attributes. For example, color image data has several attributes such as width, height, RGB values for representing the image itself. Contents of video data that are represented by features of a certain image at a certain time relate to two attributes, namely, frame identifier and image condition like shape or color. Therefore, this type of content is further classified into two classes based on the attribute values associated with the content. Here, let $C(p_1, p_2, ..., p_n)$ denotes a function that derives the content that is referred to in query evaluation, where $p_i(i = 1, ..., \mathrm{n})$ corresponds to one of the attributes owned by the media object. Let $\tau_\mathrm{m}(p_i)$ denote the media type of an object which has p_i as a part of its attributes, and let $\tau_\mathrm{a}(p_i)$ denote the attribute type of p_i.

1) *single media, single attribute content*. This type of content is represented by one of the attributes of a single media object. That is, this content is represented as $C(p_1)$.
2) *single media, multiple attribute content*. This type of content is constructed by two or more distinct attributes of a single media object. That is,

$$C(p_1, p_2, .., p_n) : \tau_\mathrm{m}(p_1) = \ldots = \tau_\mathrm{m}(p_n),$$
$$\exists \mathrm{i}, \tau_\mathrm{m}(p_1) \neq \tau_\mathrm{m}(p_i)\ (\mathrm{i} \neq 1).$$

b) **Multimedia content**. Contents which are represented by two or more heterogeneous types of data are classified as multimedia contents. We think the retrieval of this type of contents is a narrow sense of content-based retrieval for multimedia databases. In the query processing of this type of contents, two or more heterogeneous types of data are referred to, to evaluate a query condition. That is,

$$C(p_1, p_2, .., p_n) : \exists \mathrm{i}, \tau_\mathrm{m}(p_1) \neq \tau_\mathrm{m}(p_n)\ (\mathrm{i} \neq 1)$$

This type of content implies that a content can be constructed from two or more values of heterogeneous types of data. This type of content is not classified further as in the case of single media contents.

5.1.2 Source of Contents

This taxonomy is largely related to the data models by which a multimedia database is constructed. In the relational data model, only a part_of relationship is assumed between attributes, both of which are defined in the same relation.

However, as discussed in Section 2, not only attribute values themselves but also spatial and/or temporal relations are defined along with a hierarchy. There are several models proposed for the representation of such relations (see Section 2). In such multimedia databases, content lies in the value of a component object (i.e., image, video, or audio data), relation between the components, or the meaning derived from value or relation by knowledge.

From the point of view of the source of contents which are evaluated in query processing, three types of contents are enumerated.

a) **Contents of value**. Contents of value are the contents that lie in the value of media object. Queries for this type of contents are evaluated by comparing raw data of object with a query condition. If needed, raw data may be filtered or cut to extract a subset of the raw data. This is an ontology based content.

b) **Contents of relationship**. This type of content is not in values but in relationship between primitive multimedia objects, which is defined under spatio-temporal data model. The relation consists of spatial/temporal relations and intervals or coordinates associated with the relations representing spatial and/or temporal correlation. Several studies have given representations of spatial and/or temporal relations, as seen in [11], [13], [30], [24], [38], [39]. The frameworks of these studies which concern the synchronization of objects have been influenced by the work in [1].

c) **Contents of derived semantics**. Contents of derived semantics are not the contents which directly correspond to values or relations of media data, but those obtained by semantic interpretation or derivation from raw data or spatio-temporal relations. The content is epistemology-based; content of this type is evaluated by introducing the process of human's perception of concept or the recognition of objects.

5.2 Contents Covered by Three Fundamental Frameworks

In this section, we clarify the coverage of contents in the retrieval implemented by the three fundamental frameworks discussed in the previous sections, namely, spatio-temporal

TABLE 1
COVERAGE OF CONTENTS

<table>
<tr><th colspan="2" rowspan="2"></th><th colspan="3">Source of content</th></tr>
<tr><th>value</th><th>relation</th><th>semantics</th></tr>
<tr><td rowspan="9">component data type</td><td rowspan="6">single media</td><td>[Kato, 91]*
[Bimbo, 96]*</td><td>[Little, 93a]+
[Allen, 83]+
[Toman, 96]+
[Sistla, 94]+
[Sistla, 95]+</td><td>[Yoshitaka, 97]#</td></tr>
<tr><td colspan="2">[Oomoto, 93]+
[Little, 93b]+
[Corridoni, 96]*
[Flickner, 95]*
[Chang, 97]*</td><td></td></tr>
<tr><td></td><td>[Guivada, 96]#
[Egenhofer, 97]*
[Haarslev, 97]*
[Costagliola, 91]+</td><td></td></tr>
<tr><td></td><td colspan="2">[Ono, 96]#</td></tr>
<tr><td>[Ghias, 95]]*
[Wold, 96]*</td><td></td><td></td></tr>
<tr><td>[Little, 93b]#(+)</td><td colspan="2">[Corridoni, 95]+(*), [Gupta, 91a,b]#(+), [Bach, 93]#</td></tr>
<tr><td rowspan="3">multi media</td><td>[Cardenas, 93]#</td><td colspan="2">[Meyer, 94]*, [Yoshitaka, 94]#</td></tr>
<tr><td colspan="2">[Egenhofer, 94]#
[Cruz, 97]+</td><td></td></tr>
<tr><td></td><td colspan="2">[Hsu, 96]#</td></tr>
</table>

Main contribution: +:Data modeling/indexing *:QBE #:Knowledge-assisted retrieval

data model/indexing, query-by-example, and knowledge-based processing. The coverage of contents studied in the literature introduced so far is categorized based on two points of views described in Section 5.1.

As shown in Table 1, as far as we surveyed, most of studies concern the contents of a relation classified as single media contents. On the contrary, less studies are conducted for the retrieval of semantic contents classified as multimedia contents. This table also shows that all three fundamental frameworks are mandatory for covering the whole range of multimedia database retrieval.

6 CONCLUSION

In this paper, we surveyed content-based retrieval for multimedia databases. Here, we referred to multimedia databases in a broad sense; we included retrieval of contents associated with a single type of nontextual data as a part of multimedia data retrieval. As discussed in this literature, there are two principal ways for the representation of queries, namely, 'query-by-subject/object' and 'query-by-example.'

'Query-by-example' allows the user to specify a query condition in an intuitive way, i.e., it is easy to express a query condition in a natural way. In QBE, a query condition for nontextual data is represented, for example, in the form of a rough sketch, a rough painting with colors, or a motion example of trajectory and/or velocity. Such representations express the query condition for nontextual data better than keywords, since it is often difficult to express slight differences of shape, color, or spatio-temporal relation with keywords. QBE works well for content-based retrieval in the case where contents are formed in terms of a single data type. However, the QBE approach is not adequate when two or more heterogeneous types of data form the content. Rather 'Query-by-subject/object' is appreciated for such cases, where a keyword can well represent the semantic content.

Currently, many more studies have been done in relation to content-based retrieval that refers to a single non-textual data. However, we think content-based retrieval studies for *multimedia* databases should pay more attention to the multimedia content that is associated with heterogeneous types of data. Extracting implicit contents from semantically related heterogeneous types of data is advantageous in some aspects. One reason is that clues from two or more pieces of heterogeneous data which are semantically related with each other give us more implicit content, which cannot be extracted only from one of them. Another reason is that evaluating contents extracted from two or more pieces of data together may give us the results with more certainty, since current image/video processing or audio processing techniques do not always give a result with enough accuracy.

With reference to implementation of content-based retrieval facility, there are several issues. In the process of extracting components that is associated with a content, raw data processing is inevitable. This processing is one of the most time-consuming part in content-based retrieval. Improving the performance of the raw data processing therefore improves the overall performance of the system. The underlying data model of the system plays an important role in content-based retrieval as well, since the ability of content-based retrieval on spatio-temporal relations is determined by the model. Related to this, there are still open problems on content-based retrieval of spatio-temporal contents. These problems should be managed from three points of view, namely, performance, accuracy of results, and the user interface.

ACKNOWLEDGMENTS

The authors acknowledge Associate Professor Masahito Hirakawa and the members of the Information Systems Laboratory at Hiroshima University for their discussions and valuable comments.

REFERENCES

[1] J.F. Allen, "Maintaining Knowledge About Temporal Intervals," *Comm. ACM*, vol. 26, no. 11, pp. 832-843, 1983.

[2] M. Atkinson, F. Bancilhon, D. DeWitt, K. Dittrich, D. Maier, and S. Zdonik, "The Object-Oriented Database System Manifesto," *Proc. First Int'l Conf. Deductive and Object-Oriented Databases*, pp. 40-57, 1989.

[3] J.R. Bach, S. Paul, and R. Jain, "A Visual Information Management System for the Interactive Retrieval of Faces," *IEEE Trans. Knowledge and Data Eng.*, vol. 5, no. 4, pp. 619-628, 1993.

[4] A.D. Bimbo, M. Campanai, and P. Nesi, "3D Visual Query Language for Image Databases," *J. Visual Languages and Computing*, vol. 3, no. 3, pp. 257-271, 1992.

[5] A.D. Bimbo, E. Vicario, and D. Zingoni, "Symbolic Description and Visual Querying of Image Sequences Using Spatio-Temporal Logic," *IEEE Trans. Knowledge and Data Eng.*, vol. 7, no. 4, pp. 609-622, 1995.

[6] A.D. Bimbo, P. Pala, and S. Santini, "Image Retrieval by Elastic Matching of Shapes and Image Patterns," *Proc. IEEE Int'l Conf. Multimedia Computing and Systems*, pp. 215-218, June 1996.

[7] A. Celentano, M.G. Fugini, and S. Pozzi, "Knowledge-Based Rtrieval of Office Documents," *Proc. 13th Int'l Conf. Research and Development in Information Retrieval*, pp. 241-254, Sept. 1990.

[8] S.K. Chang, "Iconic Indexing By 2D String," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 4, pp. 413-428, 1984.

[9] S.F. Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong, "VideoQ: An Automated Content Based Video Search System Using Visual Cues," *Proc. ACM Multimedia*, pp. 313-324, 1997.

[10] J.M. Corridoni, A.D. Bimbo, S. De Magistris, and E. Vicario, "A Visual Language for Color-Based Painting Retrieval," *Proc. Int'l Symp. Visual Languages*, pp. 68-75, 1996.

[11] G. Costagliola, M. Tucci, and S.K. Chang, "Representing and Retrieving Symbolic Pictures by Spatial Relations," *Visual Database Systems*, vol. II, E. Knuth and L.M. Wegner, eds., Elsevier, pp. 55-65, 1991.

[12] I.F. Cruz and W.T. Lucas, "A Visual Approach to Multimedia Querying and Presentation," *Proc. ACM Multimedia*, pp. 109-120, 1997.

[13] Y.F. Day, S. Dagtas, M. Iino, A. Khokhar, and A. Ghafoor, "Spatio-Temporal Modeling of Video Data for On-line Object-Oriented Query Processing," *Proc. Int'l Conf. Multimedia Computing and Systems*, pp. 98-105, May 1995.

[14] M.J. Egenhofer, "Query Processing in Spatial-Query-by-Sketch," *J. Visual Languages and Computing*, vol. 8, no. 4, pp. 403-424, 1997.

[15] M. Flickner et al., "Query by Image and Video Content: The QBIC System," *Computer*, vol. 28, no. 9, pp. 23-32, 1995.

[16] A. Ghias, J. Logan, and D. Chamberlin, "Query by Humming," *Proc. ACM Multimedia*, pp. 231-236, 1995.

[17] S. Gibbs, L. Dami, and D. Tsichritzis, "An Object-Oriented Framework for Multimedia Composition and Synchronization," *Proc. Multimedia—First Eurographic Workshop Systems, Interaction, and Applications*, W.T. Hewtt et al., eds., pp. 101-111, Stockholm, Springer-Verlag, 1992.

[18] Y. Gong, H. Zhang, H.C. Chuan, and M. Sakauchi, "An Image Database System with Content Capturing and Fast Image Indexing Abilities," *Proc. Int'l Conf. Multimedia Computing and Systems*, pp. 121-130, May 1994.

[19] V.N. Gudivada and G.S. Jung, "An Algorithm for Content-Based retrieval in Multimedia Databases," *Proc. Int'l Conf. Multimedia Computing and Systems*, pp. 193-200, 1996.

[20] A. Gupta, T. Weymouth, and R. Jain, "Semantic Queries in Image Databases," *Visual Database Systems*, vol. II, E. Knuth and L.M. Wegner, eds., Elsevier, pp. 204-218, 1991.

[21] A. Gupta, T. Weymouth, and R. Jain, "Semantic Queries With Pictures: The VIMSYS Model," *Proc. 17th Int'l Conf. Very Large Data Bases*, pp. 69-79, Sept. 1991.

[22] V. Haarslev and M. Wessel, "Querying GIS With Animated Spatial Sketches," *Proc. Int'l Symp. Visual Languages*, pp. 201-208, Sept. 1997.

[23] S.A. Hawamdeh, B.C. Ooi, R. Price, T.H. Tng, Y.H. Ang, and L. Hui, "Nearest Neighbour Searching in a Picture Archive System," *Proc. Int'l Conf. Multimedia Information Systems*, McGraw-Hill, pp. 17-33, 1991.

[24] P. Hopner, "Synchronizing the Presentation of Multimedia Objects—ODA Extensions," *Multimedia Systems, Interaction, and Application*, pp. 87-100, Springer-Verlag, 1992.

[25] C.C. Hsu, W.W. Chu, and R.K. Taira, "A Knowledge-Based Approach for Retrieving Images by Content," *IEEE Trans. Knowledge and Data Eng.*, vol. 8, no. 4, pp. 522-532, 1993.

[26] M. Iino, Y.F. Day, and A. Ghafoor, "An Object-Oriented Model for Spatio-Temporal Synchronization of Multimedia Information," *Proc. Int'l Conf. Multimedia Computing and Systems*, pp. 110-119, May 1994.

[27] T. Kato, T. Kurita, H. Shimogaki, T. Mizutori, and K. Fujimura, "A Cognitive Approach to Visual Interaction," *Proc. Int'l Conf. Multimedia Information Systems*, pp. 109-120, McGraw-Hill, 1991.

[28] A. Klinger and A. Pizano, "Visual Structure and Databases," *Visual Database Systems*, T.L. Kunii, ed., pp. 3-25, Elsevier, 1989.

[29] E.B.W. Lieutenant and J.R. Driscoll, "Incorporating A Semantic Analysis into A Document Retrieval Strategy," *Proc. ACM/SIGIR Conf. Research and Development Information Retrieval*, pp. 270-279, Oct. 1991.

[30] T.D.C. Little and A. Ghafoor, "Interval-Based Conceptual Models for Time-Dependent Multimedia Data," *IEEE Trans. Knowledge and Data Eng.*, vol. 5, no. 4, pp. 551-563, 1993.

[31] T.D.C. Little, G. Ahanger, R.J. Folz, J.F. Gibbon, F.W. Reeve, D.H. Schelleng, and D. Venkatesh, "A Digital On-Demand Video Service Supporting Content-Based Queries," *Proc. ACM Multimedia*, pp. 427-436, 1993.

[32] Z.Q. Liu and J.P. Sun, "Structured Image Retrieval," *J. Visual Languages and Computing*, vol. 8, no. 3, pp. 333-357, 1997.

[33] K. Melih and R. Gonzalez, "Audio Retrieval Using Perceptually Based Structures," *Proc. Int'l Conf. Multimedia Computing and Systems*, pp. 338-347, 1998.

[34] B. Meyer, "Pictorial Deduction in Spatial Information Systems," *Proc. Int'l Symp. Visual Languages*, pp. 23-30, 1994.

[35] A. Ono, M. Amano, M. Hakaridani, T. Satou, and M. Sakauchi, "A Flexible Content-Based Image Retrieval System with Combined Scene Sescription Keyword," *Proc. Int'l Conf. Multimedia Computing and Systems*, pp. 201-208, 1996.

[36] E. Oomoto and K. Tanaka, "OVID: Design and Implementation of A Video-Object Database System," *IEEE Trans. Knowledge and Data Eng.*, vol. 5, no. 4, pp. 629-643, 1993.

[37] K. Otsuji and Y. Tonomura, "Projection Detecting Filter for Video Cut Detection," *Proc. ACM Multimedia*, pp. 251-257, 1993.

[38] A.P. Sistla, C. Yu, and R. Haddad, "Reasoning About Spatial Relationships in Picture Retrieval Systems," *Proc. Int'l Conf. Very Large Databases*, pp. 570-581, Sept. 1994.

[39] A.P. Sistla, C. Yu, C. Liu, and K. Liu, "Similarity Based Retrieval of Pictures Using Indices on Spatial Relationships," *Proc. Int'l Conf. Very Large Databases*, pp. 619-629, Sept. 1995.

[40] P.J. Smith, S.J. Shute, and D. Galdes, "In Search of Knowledge-Based Search Tactics," *Proc. 12th Int'l Conf. Research and Development in Information Retrieval*, pp. 3-10, 1989.

[41] Y. Theodoridis, M. Vazirgiannis, and T. Sellis, "Spatio-Temporal Indexing for Large Multimedia Applications," *Proc. Int'l Conf. Multimedia Computing and Systems*, pp. 441-448, 1996.

[42] D. Toman, "Point vs. Interval-Based Query Languages for Temporal Databases," *Proc. Fifth ACM SIGACT/MOD/ART Symp. Principles of Database Systems*, pp. 58-67, 1996.

[43] K. Tsuda, K. Yamamoto, M. Hirakawa, and T. Ichikawa, "MORE: An Object-Oriented Data Model with A Facility for Changing Object Structures," *IEEE Trans. Knowledge and Data Eng.*, vol. 3, no. 4, pp. 444-460, 1991.

[44] R. Weiss, A. Duda, and D.K. Gifford, “Content-Based Access to Algebraic Video,” *Proc. IEEE Int’l Conf. Multimedia Computing and Systems*, pp. 140-151, May 1994.

[45] E. Wold, T. Blum, D. Keislar, and J. Wheaton, “Content-Based Classification, Search, and Retrieval of Audio,” *IEEE MultiMedia*, vol. 3, no. 3, pp. 27-36, Fall 1996.

[46] X. Wu and T. Ichikawa, “KDA: A Knowledge-Based Database Assistant with A Query Guiding Facility,” *IEEE Trans. Knowledge and Data Eng.*, vol. 4, no. 5, pp. 443-453, 1994.

[47] H. Wynne, T.S. Chua, and H.K. Pung, “An Integrated Color-Spatial Approach to Content-Based Image Retrieval,” *Proc. ACM Multimedia*, pp. 305-313, 1995.

[48] M. Yeung, B.-L. Yeo, and B. Liu, “Extracting Story Units from Long Programs for Video Browsing and Navigation,” *Proc. Int’l Conf. Multimedia Computing and Systems*, pp. 296-305, 1996.

[49] A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa, “Knowledge-Assisted Content-Based Retrieval for Multimedia Databases,” *IEEE MultiMedia*, vol. 1, no. 4, pp. 12-21, 1994.

[50] A. Yoshitaka, Y. Hosoda, M. Yoshimitsu, M. Hirakawa, and T. Ichikawa, “VIOLONE:Video Retrieval By Motion Example,” *J. Visual Languages and Computing*, vol. 7, no. 4, pp. 423-443, 1996.

[51] A. Yoshitaka, T. Ishii, M. Hirakawa, and T. Ichikawa, “Content-Based Retrieval of Video Data by the Grammar of the Film,” *Proc. Int’l Symp. Visual Languages*, pp. 314-321, Sept. 1997.

[52] G.P. Zarri, “Conceptual Representation for Knowledge Bases and ‘Intelligent’ Information Retrieval Systems,” *Proc. 11th Int’l Conf. Research and Development in Information Retrieval*, pp. 551-565, 1988.

[53] M.M. Zloof, “QBE/OBE: A Language for Office and Business Automation,” *Computer*, vol. 14, no. 5, pp. 13-22, 1981.

Atsuo Yoshitaka graduated from Hiroshima University in March 1989, and received his masters and doctor of engineering degrees from Hiroshima University in March 1991 and March 1997, respectively. He is currently serving as a research associate in the Information Systems Laboratory at Hiroshima University in Japan. His research interests include content-based retrieval for multimedia databases, visual user interfaces for database retrieval, and medical information processing. He received a IEEE Computer Society Certificate of Appreciation in 1996. He is a member of the IEEE and the IEEE Computer Society.

Tadao Ichikawa graduated from Waseda University, also receiving his doctor of engineering degree from Waseda University. He is now a professor in the Information Systems Laboratory at Hiroshima University in Japan, where he conducts research on multimedia computing, real-world computing, and virtual-world interactions. For the IEEE Computer Society, he led organization of the IEEE Symposium on Visual Languages and the IEEE International Conference on Multimedia Computing and Systems in 1984 and 1994, respectively. He founded the Computer Society’s Technical Committee on Multimedia Computing and the Task Force on Youth Forum in Computer Science and Engineering (YUFORIC) in 1992 and 1995, respectively. He is a member of the Editorial Board of *IEEE Transactions on Knowledge and Data Engineering* and the *International Journal on Visual Languages and Computing*, and he was recently invited to serve as a member of the Steering Committee for a new IEEE transactions devoted to multimedia. He is currently a member of Board of Governors of the Technical Activities Board, the Membership Activities Board, the Conferences and Tutorials Board, and chair of the 1998 IEEE Fellow Evaluation Committee. He received the Meritorious Service Award, the Outstanding Contribution Award, and the Technical Achievement Award of the IEEE Computer Society in 1992, 1994, and 1998, respectively. He is a fellow of the IEEE.