Pattern Recognition 42 (2009) 267 – 282

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

State-of-the-art on spatio-temporal information-based video retrieval

W. Ren[a], S. Singh[b,*], M. Singh[b], Y.S. Zhu[a]

[a]The Key Laboratory of Integrated Microsystems, Shenzhen Graduate School, Peking University, China
[b]Research School of Informatics, University of Loughborough, Loughborough LE11 3TU, UK

ARTICLE INFO

Article history:
Received 30 December 2007
Received in revised form 21 August 2008
Accepted 22 August 2008

Keywords:
Video retrieval
Semantic knowledge
Content-based analysis
Spatio-temporal information

ABSTRACT

Video retrieval is increasingly based on image content. A number of studies on video retrieval have used low-level pixel content related to statistical moments, shape, colour and texture. However, it is well recognised that such information is not enough for uniquely discriminating across different multimedia content. The use of semantic information, especially which derived from spatio-temporal analysis is of great value in multimedia annotation, archiving and retrieval. In this review paper, we detail how the use of spatiotemporal semantic knowledge is changing the way in which modern research the conducted. In this paper we review a number of studies and concepts related to such analysis, and draw important conclusions on where future research is headed.

1. Spatio-temporal information for video retrieval

Content-based video retrieval is a very important area of research and several practical systems have been developed over the last decade with the aim of improving retrieval performance and tested on large-scale databases such as TRECVID http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html. Video classification and retrieval problems can be hierarchically categorised with a taxonomy, an example of which is presented by Roach et al. [1]. A key characteristic of video data is its associated spatial and temporal information that delivers semantically coherent narrative. Temporally consecutive frames have explicit spatial constraints with object inheritance, spatial relationships and motion information from their previous frames. Temporal trajectories of spatial relations among objects are as important as temporal object trajectories to represent object activities and reveal semantic evolution of spatial properties over time. The holy grail of almost all content matching-based video retrieval systems is to improve precision and recall metrics both through the process of improved content representation and use of good quality similarity metrics [2], as well as using a range of relevance feedback architectures and algorithms to allow the system to learn with time what is and is not a good match [3–6].

Unfortunately, temporal and spatial characteristics have not been adequately addressed in most video retrieval systems despite their obvious importance. In such systems, retrieval techniques work on indexing video by treating video sequences as collections of still images, extracting relevant key-frames, and comparing their low-level features. Over the past years, the representation of spatio-temporal data has been extensively discussed. It has inspired the development of mathematical foundations to represent spatio-temporal logic (STL) and reasoning [7], spatio-temporal database models and query languages for the description and manipulation of spatio-temporal objects [8,9], the temporal extension of current spatial data models within GIS [10,11], and a new generation of spatio-temporal video retrieval systems [12]. Spatiotemporal information in video deals with the evolution of spatial objects that change over time. Spatio-temporal modelling in video retrieval is a crucial step for using semantic information on image object relationship to improve the quality of content-based video retrieval. Such information can be used to tag video content and used as the basis for similarity computation between query and database videos. The similarity metrics and matching approaches depend heavily on the representation of spatio-temporal information, e.g., motion feature, spatio-temporal relations, object trajectory, video transition, etc. However, how to effectively model and represent spatio-temporal information is not straightforward. A spatio-temporal model usually first partitions the video into physical meaningful units (shots). This is followed by modelling the spatial relationships among objects in each frame. A final step analyses the temporal evolution of spatial relationships among objects over temporal intervals in each shot as well as in the whole video sequence. More importantly, a spatio-temporal model should suggest a practical solution for effective indexing and comparison. In summary, a spatio-temporal model should provide for:

(a) Representation of the structural elements of video data such as frame, shot, and sequence at different levels of abstraction.

* Corresponding author.
E-mail address: s.singh@lboro.ac.uk (S. Singh).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.08.033

(b) Description of the spatial composition among video objects in each frame including directional and topological relations, and temporal composition among frames within shot and sequences.

Spatial and temporal compositions are two important aspects for the representation of a spatio-temporal model. There are two main approaches for modelling such information:

(a) An integrated approach where objects, their spatial relationships and events are considered as a 3D (three-dimensional) volume with time being the third axis. One can construct a volume of spatio-temporal data in which objects in consecutive images are stacked to form a third temporal dimension. In this approach, the video events can be represented by the analysis of this 3D space based on object trajectories, shape analysis and motion analysis. A sequence of frames $(f_1 f_2, \ldots f_N)$ is represented by a volume in (x,y,t) space, where (x,y) are the discrete spatial coordinates in each frame (f_i), and time (t) is a discrete temporal coordinate that specifies frame number. The key benefit of this representation is that objects' spatial and temporal continuity is explicitly and conjointly provided. Shape and position change of a video object over time (t) is considered in terms of translation, scaling, and rotation of the object. A semantic scene can be delivered as variances of visual appearance from sequence to sequence. This is sequence-to-sequence indexing model. Spatio-temporal information relating to object movement is identified by tracing the trajectories of objects in this 3D (x,y,t) space. The motion trajectories of objects are defined as a physical change in the geographic position of the objects in the video. The trajectories are derived from changing the location of particular points on the objects, or from tracking contours of the objects over time. The former is trajectory slice model, whereas the latter is called trajectory volume model. In this model, time (t) is a critical component. The representation of this model is highly time dependent. Therefore, using different time scales will impact on the representation of this model, and further impact on the final results of indexing and matching. For instance, when we try to match two actions under different time scales by shape comparison, the solution is not straightforward. This complexity is mainly due to the camera motion which induces a global motion in the video in addition to the object's motion during performance of an action. Additionally, it may be due to the action performed at a different speed or the object motion probably observed at different time instants with different temporal extents and under different viewpoint.
The representation of a video sequence as a volume in (x,y,t) space was first pioneered in Buxton and Buxton [13], in which a spatio-temporal gradient scheme is introduced for motion computation and inferring a static scene's depth information. Aldelson and Bergen [14] more explicitly proposed a motion sequence represented as a single pattern in x–y–t 3D space. Since then, the spatio-temporal volume has been predominantly studied in image processing. Bolles et al. [15] first investigated slices of the spatio-temporal volume to recover geometrically static scene structure from motion. Later they exploit spatio-temporal volume for object tracking [16]. Following this idea, other researchers have studied spatio-temporal helix [17], temporal slice analysis [18], oriented energy measurements [19], etc., and applied these concepts to spatio-temporal analysis of video sequences. We give details on this in Section 4.

(b) A separate modelling of spatial relationships (based on spatial logic relations) between object pairs, from temporal modelling based on how these relationships might vary, change in camera position or object movements, position of change in scenes (cut), change in illumination, colour, texture and shape across frames, etc. The information gathered is now fused together either by concatenating spatial and temporal vectors, or through a weighted combination. One option is to keep the information separate once extracted and the SQL type query can be applied—the video that matches the query on the majority of the spatio-temporal features is chosen as the best match. An example is SEMCOG system [20], which represents spatial constraints among objects by using 2D (two-dimensional) string and describes temporal action by using Allen's [21] 13 temporal logic relationships along with distance constraints. Queries use a semantic language—CSQL and VCSQL, which is similar to the standard SQL. These two types of information fusion models can deal with very complicated cases of video retrieval. However, the former is not addressed properly, whereas the latter does not support a comparison by using similarity metrics.

In this paper, we review the state-of-the-art spatial and temporal models with the aim of using these for image and video retrieval. This paper is organised as follows. In the rest of Section 1 we give an overall brief review of spatiotemporal models for video retrieval. Section 2 reviews spatial modelling of video and image data. In Section 3, we discuss research on temporal modelling. Finally, in Section 4, we present a brief review on spatiotemporal information fusion.

2. Spatial information modelling in multimedia retrieval

2.1. Spatial representation

Spatial information can be formulated with the following two methodologies:

- The first approach is to use weak spatial constraints and capture spatial local information to represent low-level texture features. Examples include Gabor wavelets [22], local histograms [23], co-occurrence matrices [24], colour correlograms [25], composite region templates (CRTs) [26], etc.
- The second approach is to represent global qualitative spatial relations that support high-level semantic textual queries. Examples include symbolic projections [27,28], spatial logic [29], θ–R representations [30], etc.

We are more interested in the second type of spatial representation. Spatial qualitative relations between objects are very important for video and image retrieval to support effectively high-level spatial queries. An overview of the major qualitative spatial representation and reasoning techniques is available in Cohn [31]. In the following three sections we discuss three major representation models: (a) 2D strings and its variants (Section 2.1.1); (b) spatial logic (Section 2.1.2) and (c) other models (Section 2.1.3). A number of these models have been inspired by the initial work of Allen [21].

Allen [21] introduced an interval-based temporal logic, which considered objects/events along a 1D (one-dimensional) time axis as a set of temporal intervals based on comparative relations. This differs from point-based approaches, prevalent at that time in the logic and reasoning literature. Allen [21] defined 13 mutually exclusive relations which hold between two intervals: {*before*, *meets*, *overlaps*, *during*, *starts*, *finishes*, and their inverse relations, and *equal*}. Allen's 13 relations can be expressed in terms of at most three order operators ($<$, $>$, $=$). The elegance and simplicity of Allen's temporal interval algebra has inspired several further developments both in temporal and spatial reasoning. It has been formalised as topological relations in 1D spatial domain. It promotes development of symbol projection for spatial image indexing. Lee and Hsu [32,33], for example, represented 13 types of topological relations in 2D-C string, shown in Fig. 1, using the principles of Allen's temporal

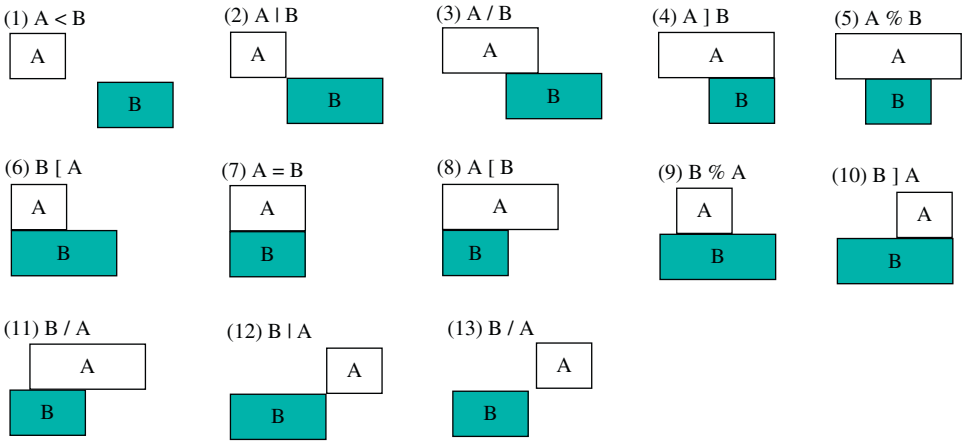

Fig. 1. The 13 types of spatial relations in one dimension for the 2D C-string.

interval logic. Chang and Jungert [34] proposed 2D string by projecting object centroids onto the two axes of a Cartesian coordinate system and deriving spatial relations from spatial order of objects similar to the temporal case along *x*- and *y*-axes. Two spatial relation strings are used to depict spatial configurations of object position for 3D scenes.

2.1.1. 2D string and its variants

Two-dimensional string provides a simple representation of spatial image properties in the form of two 1D strings and provide an efficient way to derive spatial queries from symbol projection image by arranging visual icons.

The 2D string associated with iconic representation is used to index the spatial layout of images, and the retrieval query is executed by string-substring matching [35]. To construct a 2D string, the image must be first segmented into disjoint regions or objects and an associated logic representation "symbolic image" is built. The positions of objects in a symbolic image represented by their centroids, projected on both *x* and *y* axes, and object labels are also projected on the two axes. Relationships among objects are specified along the *x* and *y* directions resulting in two 1D strings. The 2D string is the most common data structure that is used for representing positional relations. The 2D string uses simple "left/right", and "below/above" relationships between objects to represent the semantic or structural image content. The spatial relationship between two objects is denoted by one of the following symbols, $\{=, <, :\}$, where the symbol "$=$", "$<$", and ":" denote the spatial relation "located at the same position", "either left/right or below/above", "falling in the same grid square", respectively. A query such as "find all images having a tree to the left and to the bottom of a house" can be represented by the string (tree < house, tree < house). The problem of content-based retrieval of images becomes one of 2D string subsequence matching. To match the 2D string, it is first simplified into a reduced 2D string with the '$<$' operator only. Then subsequence exact matching between two reduced 2D strings is applied by exhaustively enumerating and storing all combinations of spatial relations of images for a query image and images in the database.

The 2D strings are mostly applied where images contain objects which are mutually disjoint and have rather simple shapes. The spatial relationships between overlapping objects cannot be clearly identified by the projection of object centroid, and therefore, the 2D strings as a representation of symbolic images are insufficient to provide the expressive power to describe an image of arbitrary complexity. To deal with such situations, various extensions of the original representations of the 2D strings have been proposed, such as 2D E-strings [36], 2D G-strings [37], 2D C-string [32,33], 2D B-strings [38], 2D C^+-strings [39], "expanded 2D strings" [40,41], and 2D RS-strings [42].

Two-dimensional G-string [37] still adopts the 2D string point-based representation. In the 2D G-string, there are three spatial operators $\{=, <, |\}$. Different to the 2D string, the 2D G-string introduces an operator '|' instead of ':' 2D string operator to denote "adjoin" or "edge to edge relation". To solve the problem of object overlapping, the 2D G-string uses a cutting mechanism to split objects into simple, non-overlapped sub-objects to specify spatial relationships between overlapped objects. Once this has been done, object structures are not longer simple and unified. With increasing number of sub-object pieces after cut, matching complexity will increase as well. As a result of this, the concept of 2D C-string was put forward by Lee and Hsu [33].

In the 2D C-string representation, Lee and Hsu [32,33] reduced the number of cuts required by introducing new spatial operators to represent images with arbitrary complexity. In the 2D C-string spatial relationships of overlapping objects are expressed as of three types: "disjoint", "same position" and "connection from edge to edge" by using bounding projection line. The string expressions of the 2D C-string are shorter than that of 2D G-string. Since the 2D C-string adopts a reduced cutting mechanism to split overlapped objects and a mechanism of interval-based projection, it does not reduce image indexing complexity much, in comparison with the 2D G-string. More importantly, no efficient techniques for indexing the 2D C-strings and the 2D G-strings have been developed. According to the bounding interval-based projection of the objects, in the 2D C-string, there are 13 types of spatial relations between two 1D intervals (see Fig. 1), seven spatial operators with symbols ($<$, $=$, |, %, [,], /) (see Table 1). The key advantage of 2D C-string is that it has very strong expressive power. It can represent 169 types of spatial relationships between two minimum bounding rectangles (MBRs) in 2D space [32]. These 169 relations consist of five types of relationships: disjoint (48), joint (40), part-overlap (50), contain (16), and belong (16). Except for the problems of complexity and indexing, the 2D C-string has a substantial problem in that it easily includes noise in its spatial relationship representation, since it tries to retain details of the object boundary. Thus, the 2D C-string does not produce very effective spatial representation.

Two-dimensional B-string [43] attempts to reduce the complexity of the 2D C-string. It is an interval-based representation. The 2D B-string considers any object as a compact MBR without using a cutting mechanism. In 2D B-strings, objects are represented by start- and end-bounding projections, which list all object

Table 1
The definition of spatial operators in the 2D C-string

Notation	Condition	Meaning
A < B	end(A) < begin(B)	A disjoins B
A = B	begin(A) = begin(B), end(A) = end(B)	A is the same as B
A|B	end(A) = begin(B)	A is edge to edge with B
A%B	begin(A) < begin(B), end(A) > end(B)	A contains B and they have not the same bound
A[B	begin(A) = begin(B), end(A) > end(B)	A contains B and they have the same begin-bound
A]B	begin(A) < begin(B), end(A) = end(B)	A contains B and they have the same end-bound
A/B	Begin(A) < begin(B) < end(A) < end(B)	A is partly overlapping with B

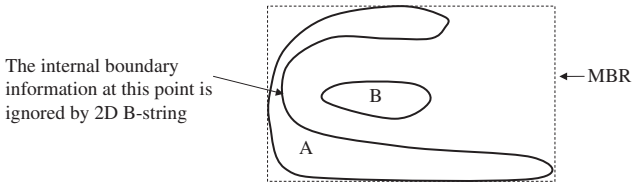

Fig. 2. Spatial configuration of objects A and B.

start-bounding positions in ascending order and followed by listing the end-bounding position ranked in ascending order along the *x*- and *y*-axis, respectively. There is only one spatial operator " = " used in the 2D B-string. It is used to specify that the projection of objects have the same bounding projection line or that start or end boundaries are projected at the same location. The advantage of the 2D B-string is that its expressions are simple. However, the problem with the 2D B-string is that it ignores the internal boundary details and has insufficient expressive power to describe topological relationships between objects with complex shapes, such as objects 'A' and 'B' as shown in Fig. 2. In Fig. 2, object 'A' and 'B' are not even partially intersecting. Unfortunately, it is interpreted by the B-string that the object 'A' is completely included in the object 'B'. Similar to the 2D C-string, the B-string does not have enough spatial constraints to eliminate noise effects because of using interval-based projection.

A number of applications for indexing and retrieving images and videos based on 2D strings and its variants have been developed.

For image retrieval SaFe system, Smith and Chang [44] provide an example application for searching and comparing images by using the 2D string. Multiple regions can be queried by absolute or relative locations of their spatial layout. For instance, a spatial query is specified by constructing symbolic images. SaFe also introduced a simple extension of the 2D string for projection and approximate rotation invariance around the image centre point with 0°, 45° and 90° image rotation and region projection. The 45° rotated 2D string was extracted by the projecting objects onto the diagonal of the image. Image rotation by 90° was provided by swapping the *x* and *y* projections.

Further extensions of the 2D string and its variants used for representing 3D scenes of video sequences can be found in Costagliola et al. [45]; Chang et al. [35]; Bimbo et al. [46], 3D-list [47], 3D C-string [48], 2D C-trees [49], and 3D string [47].

Hsu et al. [49] proposed 2D C-trees, which improve upon 2D C-string [33] for video retrieval. A video sequence could then be represented and indexed by a temporal set or an ordered set of 2D C-trees. Video retrieval was treated as the problem of video sequence matching by computing the minimum cost of matched frames. Each frame was constructed as two representative 2D C-trees along the *x*- and *y*-axis. The 2D C-tree was organised with spatial operators associated with the links of the tree. Each node with a label, or symbol name, represented an object in the image. The links connecting two nodes were signed with the relation operators. A set-node was a multi-label node consisting of objects that had the same begin-bounding and end-bounding.

Liu and Chen [47] extended the 2D string [28] to a 3D string concept and proposed a data structure to represent video. The knowledge structure of 3D string used the projections of video objects to represent spatial and temporal relations between them. The basic idea was to project the video objects onto the *x*-, *y*-, and time-axis to form three strings representing the relative positions of the projections in the *x*-, *y*-, and time-axis, respectively. A video object is represented by its centroids and starting frame number. Two operators "$/_n$" and " ≡ " are introduced in the 3D string representation. The operator "$/_n$" denotes that the two objects are adjacent and have the distance *n* between them, whereas " ≡ " denotes the same position. However, similar to 2D string, they only use position relations. Also the 3D string ignores topologic relations between the video objects.

Lee et al. [48] extended 3D-list [47] and 2D C^+-string [39] approach and proposed a 3D C-string data structure to represent video by using the projections of objects to represent their spatial and temporal relations in a video. Moreover, it can keep track of the motion and size changes of the objects in a video.

2.1.2. Spatial logic and systematic derivation of spatial relations

There are various extensions to the 2D string concept discussed earlier to deal with the situation of overlapping objects with complex shapes based on direct point- or interval-based projection comparison between objects.

Qualitative description of object relationships is important not only for computer vision but also for other cognitive tasks, such as GIS, databases, and other physical and engineering applications. The question here is what distinctions are necessary to qualitatively describe object relationships. In the following, a brief review of some important spatial models using spatial logic is given.

Allen's temporal interval algebra expresses topological relations in 1D space. A variety of studies on qualitative spatial reasoning have exploited and extended Allen's [21] temporal reasoning logic. Examples include Refs. [29,50], Pullar and Egenhofer (1988) and Guesgen (1989). These studies extend concepts of Allen's temporal interval algebra to spatial logic in order to represent geometric ordering relationships between the projections of the objects in 1D or 2D spaces.

Egenhofer and Franzosa [29] specify eight fundamental topological relations between two objects: {*disjoint*, *contains*, *inside*, *meet*, *equal*, *covers*, *covered-by*}. These relations are based on the discussion of intersections over the boundary and interior point sets between two objects. Egenhofer and AI-Taha [50] emphasise that the internal relationships are important to represent the closeness of topological relations after studying gradual changes to topological relations over time. Vazirgiannis et al. [51] extended the topological relation definition of Ref. [50] by giving spatial relations a more complete and systematic description with directional relationships, topological relationships, and of the distance characteristics. Similar to Vazirgiannis et al. [51], Cohn [31] considered spatial relations of topology, orientation, shape, size and distance. Furthermore, Papadias et al. [171] investigated topological relationships within the

context of MBRs. Li et al. [52] classified directional relations into the following three categories: strict directional relations (north, south, west, and east), mixed directional relations (Northeast, southeast, northwest and southwest), and positional relations (above, below, left, and right).

At least two important factors determine the relative position of objects in 2D space: the topological relations and directional relations between objects. *Topological* relations address how the boundaries of two objects relate, express topological extension of two object boundary, and describe neighbourhood and incidence (e.g., overlap, disjoint). It is an interval-based representation. *Directional relations* address where the objects are placed relative to one another, express relative orientation of two objects and describe spatial order along the x and y directions. It is a point-based representation based on the comparison of two object centroids. In comparable projection space, *distance* relations address qualitative or quantitative distances between objects, and express relative space range between objects (e.g., far, near). However, if videos or images in databases are collected from different sources and are captured under different resolutions or at different viewpoints, quantitative value of distance between two objects does not have a physical meaning.

Spatial logic has explored definition of spatial relations from mathematical theories of points, intervals, and sets. Its development led to the mathematical foundation used to interpret and recognise spatial configurations for cognitive science. However, spatial logic does not provide a practical solution for effectively indexing and storing these spatial relations.

The spatial models given by spatial logic have been applied in the database and knowledge-based spatial domain. For example, Liang et al. [53] formulated video object spatial relations via 13 types of topological relations and directional relations called R-strings, including {*disjoins*, *touches*, *intercepts*, *starts*, *is-inside*, *finishes*, *is-equal-to*, *is-finished-by*, *contains*, *is-started-by*, *is-intercepted-by*, *is-touched-by*, and *is-disjoined-by*}. User queries are transformed into a structured video query language (Video SQL). STARS video indexing system [54] also computed spatial relationships based on MBRs. STARS represented 12 directional and eight topological relations. Directional relations consist of {*north*, *south*, *west*, *east*, *northeast*, *southeast*, *northwest*, *southwest*, *above*, *below*, *left*, and *right*}, while topological relations include {*covers*, *covered by*, *inside*, *contains*, *equal*, *overlap*, *touch*, and *disjoint*}.

Knowledge-based spatial models have been extended to capture more detailed spatial description for medical image retrieval and face recognition, where image information is expressed based on spatial relationships between regions of interest. For example, the spatial relationship for a lesion that is near another object can be captured using the distance of the centroids of the two contours on the x- and y-axis, the angle of coverage (the angle for viewing a contour from the centroid of another contour), and the ratio of object area to classify the spatial relationship. Hsu et al. [55] proposed a knowledge-based spatial model (KSIM) for medical image matching. KSIM takes into account qualitative topological relations. For example, “disjoined” is further sub-categorised as “far-away” and “nearby”; “joined” is further categorised as “bordering”, “invading”, and “circumjacent” and had a corresponding set of parameter measures.

2.1.3. Other spatial models for image retrieval

Apart from the 2D string and its variants, some other spatial models are also used in computer vision. These models have gained less popularity but none the less offer a novel and fresh perspective for solving the problem of image and video retrieval, such as 2D projection interval relationships (2D-PIR) [56], the nine direction low triangular (9DLT) matrix [27], qualitative pairwise salient region patch [57], CRTs [26], attributed relational graphs (ARGs) [58–61], θ–R representations (also called spatial orientation graph) [30,62], and mixed directional relations-based quadtree representation [63].

Some studies compare spatial relationships in a weak manner based on spatial partition of the image without image segmentation, such as weighted five fuzzy regions [23], 2D tagging 5×5 grid scheme in CAETI Internet multimedia library [64] and the entities and attributes model [65,66].

Most approaches compare spatial relationships using segmented images. In this case, regions (or objects) are taken as basic units of analysis. MBRs are used to represent the extracted regions for efficient computation and comparisons. Thus, in query comparisons, the similarity between colour regions in the images can be measured based on user-specified query rectangles. Examples include ImageSearch System [67], Hsu et al. [68], Netra [69], QBIC [70], and VisualSEEk [71].

In the above discussion we discussed the different spatial representation models that can be applied for image and video analysis. The basic idea is to encode spatio-temporal information in a format that can be compared across two different image sequences. The quality of match will depend on two key factors: (i) quality of spatio-temporal information encoded; and (ii) the similarity matching approach used. In the following section we address similarity matching approaches which have a major impact on the quality of video retrieval.

2.1.4. Similarity metrics and spatial indexing

A number of similarity metrics are used for comparing two spatial relationships. Some representative examples of similarity metrics and associated spatial indexing approaches with a reference to studies include:

- Chang et al. [28] and Chang and Lee [27] used *subsequence exact matching* for 2D-string, and Smith and Chang [71] in VisualSEEk used 2D-string indexing.
- Messmer [59] used *edit distance* for ARGs, and Hsu et al. [49] using edit distance for 2D C-trees.
- Petrakis et al. [60] in ImageMap system used *R-tree matching* for ARGs.
- Smith and Li [26] used co-occurrence matrix comparison for counting the difference in colour frequencies of sequential spatial regions.
- Nabil et al. [56] used *Euclidean distance* for 2D Projection Interval Relationships (2D-PIR).
- Gudivada and Raghavan [30] and Li and Ozsu [54] used *Cosine metric* for θ–R representations.
- Stricker and Dimai [23] used weighted distance for partitioned region grids.
- Lipson [57] used *template matching* pairwise region patches to define spatial scene layout.
- Yu and Wolf [64] used LOST language for *query matching* for CAETI Internet multimedia library.
- Rodrfguez and Jarur [72] used a *genetic algorithm* for searching spatial configurations.

Although a number of studies have worked towards spatial representation using qualitative spatial logic in the literature, most of them only use SQL-like or other query languages to perform matching. A few studies have employed qualitative similarity metrics of computer vision or pattern recognition for similarity comparison. An example can be found in Hsu et al. [55] who suggest a knowledge-based spatial model for medical images representation and retrieval. Unfortunately, their spatial model only supported SQL-like query language matching.

2.2. Temporal relations

Temporal relations are explicit or inferred coherence relations that are used to depict events or states ordered with respect to time. These relations can be expressed by using beginning and ending bounds of intervals, time sequence (before, after), or the simultaneous relations. Allen [21] proposes temporal interval algebra for representing and reasoning about temporal relations between events represented as intervals. Allen's work [21] on temporal intervals lays the foundation for further research concerned with time intervals. For example, Freksa [73] presents a generalisation of Allen's temporal interval reasoning approach based on semi-intervals (beginnings or endings events). He introduces an important notion of "conceptual neighbourhood" for qualitative temporal relations. Events corresponding to neighbouring relations "can be directly transformed into one another" by continuous "deformation" operations (i.e., moving in time, shortening and lengthening due to duration varying). An application example can be found in Hamameh et al. [74]. These neighbourhoods not only lead to increase temporal reasoning inferencing efficiency but also to prompt systematic development and effective representation of topological relations in spatial logic.

Similar to Freksa [73], Vazirgiannis and Hatzopoulos [75] extend Allen's interval algebra by defining a set of operators to implement Allen's temporal relations logic representation. This set of operators can be used to present the semantics of a synchronisation mechanism.

Further developments of temporal reasoning for manipulating multimedia data can be found in Little and Ghafoor [76] and Hjelsvold et al. [77]. Ref. [76] apply Allen's temporal interval algebra to represent time-dependent multimedia data. They introduce two interval-based conceptual temporal hierarchy models for capturing these timing relationships and managing them as part of a database. They define *n*-ary and reverse temporal relations along with their temporal constraints, whereas Weiss et al. [78] propose a system implementing video algebra operations for video access and management such as composing, searching, navigation, and playing back. They define 12 video algebra operations {*concatenation*, *union*, *intersection*, *difference*, *parallel*, *parallel-end*, *conditional*, *loop*, *stretch*, *limit*, *transition*, and *contains*} to describe video temporal relationships between video segments. In addition, Hjelsvold et al. [77] also consider video as time dependent data. Video production is thought of in terms of three interval-based operations {*intersection*, *union*, *and difference*}. They suggest that indexing videos should be based on these interval-based temporal relationships. Videos can be queried with different time scales by mapping video objects between different time coordinate systems.

Typically, nowadays temporal relations commonly adopted are still using Allen's [21] original definition which contains 13 relations {*before*, *during*, *overlaps*, *starts*, *ends*, *equal* and their inverses (does not apply to *equal*)}. These 13 temporal relations contain considerable discriminatory power. However, it is not trivial to capture temporal characteristic of a video and determine temporal relations in an implicit manner for effective retrieval analysis. In the OVID system, Oomoto and Tanaka [79] introduce a temporal model using algebraic operators to define video objects relationships. Similar to OVID, video database systems such as the VideoSTAR [80], the VIQS [81], SEMCOG [20] and the common video object tree (CVOT) [82] are also focused on temporal intervals. All query the temporal properties of video data and locate video objects or video segments based on their temporal relationships.

A representative example of building a temporal video management system and retrieval from videos by using interval temporal relations is the CVOT [82]. The CVOT scheme builds a tree based on the common salient objects in a set of clips. In CVOT, the leaf nodes are ordered from left to right by their time intervals. An interval node represents a set of common salient objects, which appear in all of its child nodes. The only node that can have an empty common salient object set is the root node or a node of the clip with an empty salient object set. Every node (including internal, leaf, and root node) has a time interval and a set of salient objects which appear during this time interval. Traversing the tree from the leaf nodes to the root, shrinks the cardinality of the common object set. In CVOT, however, all videos are modelled based on user viewpoints with manual frame annotation. No image understanding is performed in the CVOT scheme.

These systems are based on knowledge-based formalisms for event specification of video data. The problem with such systems is that manually annotating temporal intervals is time-consuming and may cause semantic heterogeneity due to inconsistent perceptions by different users.

2.3. STL representation

There are several approaches using STL that model and represent video data. Day et al. [83] perform spatio-temporal indexing based on generalised *n*-ary relations and corresponding interval constraints [76]. This framework allows data modelling and semantic abstraction of video data. Video objects are manually annotated. The specified spatial and temporal relationships between objects are represented using STL. Multi-level indexing and searching mechanism analyses information at various levels of granularity.

Another attempt is to express scenes and sequences in a formal logic language supported by symbolic projection and STL, Bimbo et al. [7]. In their study, temporal characteristics of video data are captured with temporal logic in 1D time axis to represent the spatial constraints of the objects in a scene. This approach uses boolean connectives and temporal ordering relationships. In their model, a video scene could be searched and browsed from three levels to assert and interpret the spatial relationships between object pairs on *x*-axis projections. The first level is to distinguish "before", "after" and "overlapping" conditions. The second level includes adjacency conditions and overlapping with either complete inclusion or partial intersection. Finally, the third level distinguishes all of the 13 possible distinct mutual positions between two objects. However, the key shortcoming of this approach is that the language used does not support qualitative comparison; the same is true for the models proposed by Bimbo et al. [7] and Vazirgiannis et al. [51].

Finally it is worth mentioning the work of Vazirgiannis et al. [51]. Their model integrates the following three important concepts to define the spatio-temporal composition and explore the specification of spatio-temporal relationships for video data:

(a) temporal logic operators [75] based on Allen's temporal interval algebra [21],
(b) topological relationships [29],
(c) directional relationships [84].

This work provides a framework for spatio-temporal query language of video databases, where spatial and temporal information can be independently extracted.

Pissinou et al. [85] give a more detailed specification of query representation based on spatio-temporal composition of objects in video sequences. They project 3D scenes on three 2D planes. In each plane, a video object is represented by MBRs. For the spatial part, each object-pair in each plane is labelled using one of the 169 directional and topological relationships, the projection value of closest distance and centroid distance in the horizontal and vertical axes. For temporal analysis, the operators proposed by Little and Ghafoor [76] are used to represent the binary temporal relationship between two temporal intervals. The main weakness of this model is its high

computational complexity. These spatio-temporal representations supported by logic language or operators are only suitable for video search and retrieval using an SQL type language.

Salembier et al. [86] presented a set of description schemes (DS) dealing with video programs, users and devices. The physical video structure was described by the temporal organisation of the sequences (segments), the spatial organisation of images (regions) as well as the spatio-temporal structure of the video (regions with motion). The semantic description is built around objects and events. Finally, the physical and semantic descriptions are related by a set of links defining where or when instances of specific semantic notions could be found.

The above retrieval works focus on spatio-temporal representation for browsing video databases or as query language. However, they do not provide a solution on how to index and compare the similarity of two videos. Only a few researchers have explored the use of spato-temporal logic for video indexing as discussed below.

Hsu et al. [49] extend 2D string [28] iconic approach to video data indexing and propose 2D C-trees to represent the spatial content within individual frames. A video sequence can then be represented by a temporal ordering set of 2D C-trees. The video frame sequence matching (VFSM) problem is solved by computing the minimum editing distance of 2D C-trees. A tree-matching algorithm is used in deciding the editing distance.

Ren and Singh [87] proposed R-String representation to formulate position relation and topological relations of objects into binary string. Image frame sequences were described through discrete sequential feature point sets in hyperspaces. They apply two different approaches to address video indexing problem: firstly, by finding minimum cost flow in a bipartite network, and secondly by searching nearest neighbours in sequential feature vectors.

In the above discussion, we have detailed the state-of-the-art research in the area of spatio-temporal representation. We presume that image objects and their positions are known, and this forms the basis of a representation that can be matched. However, this is not the only form of spatio-temporal representation. Image object behaviour across image frames based on motion and trajectory analysis can also be used for matching. With another approach, video sequences can be initially described by a sequence of feature values and these can be used to compute the similarity between image sequences. In Section 3 we discuss these methods of spatio-temporal information representation and matching.

3. Visual appearance-based representation and recognition

Spatio-temporal motion-based recognition has the wide spectrum of applications in surveillance, automation, health and medical systems, etc., through perceptual identification of biometrics, activity recognition. For example, motion analysis can be used in sports and athletic training, e.g., analysing tennis strokes. For instance, the discrimination between different tennis strokes is investigated by Yamato et al. [88]. Motion-based recognition can be employed in obstacle avoidance of moving objects for robots, and satellite monitoring of weather disturbances. Motion plays an important role in the human visual system, such as recognising a distant walking person by his/her gait, cyclic motion, dance steps, analysing gait to find clinical abnormalities, distinguishing flying birds and airplanes, interpreting lipreading, and hand gestures. Motion perception helps us recognise different objects and their motion in a scene, infer high-level semantic events or actions, etc.

As described in Section 1, another type of spatio-temporal information representation or models is based on visual appearance that are analysed using pattern recognition techniques. The discussion on these is subdivided into motion trajectory-based method (Section 3.1) and sequence-to-sequence-based matching (Section 3.2).

3.1. Motion trajectory-based representation and recognition

Spatio-temporal recognition can be typically categorised as either trajectory-to-trajectory-based approach or sequence-to-sequence-based approach. The former attempts to estimate object motion trajectories to recognise moving objects' behaviour, activities, human gait, etc. The latter performs inference based on pixel changes frame-by-frame. There are four significant research trends for analysing object trajectory: slice of trajectory volume, trajectory volume, shape-from-motion (SfM), and multi-camera tracking with view invariance. These approaches are discussed in the following sections.

3.1.1. Slice of space–time trajectory volume

Spatio-temporal (x,t) slices from the image sequence volume (x,y,t) are very popular for video event description because they are relatively simple to extract, and their interpretation is obvious. Both the intrinsic properties of the objects represented by image regions with colour, shape and texture and their dynamics represented by the motion trajectories are used to describe events. Knowledge about an object and its motion can be used to construct models of object behaviour. Usually, moving objects absorb the most attention. Most research on motion trajectory recognition relies on spatio-temporal image processing. Examples such as Ricquebourg and Bouthemy [18] exploit (x,t) slices to interpret typical trajectory patterns associated with articulated motion such as human gait. They address the problem of tracking the apparent contours of a moving articulated structure for analysing human motion in video. They claim that this paradigm can lead to a simple trajectory recognition scheme that can be used for analysing human gait. An extensive treatment of this topic can be found in Jahne [89].

Motion trajectory matching based on spatio-temporal slice comparison is the most common method for spatio-temporal video analysis and retrieval. Of course, actual approaches are implemented in several different ways. A good survey can be found in Aggarwal and Cai [90]. The following Sections 3.1.1.1–3.1.1.8 describe individual recognition methods.

3.1.1.1. Curvature representation and matching. The trajectory of an object can be considered as a 2D curve on a plane. The problem of representing motion is translated into the problem of representing curves. Numerous approaches to represent curves have been developed. Polygonal approximations and spline approximations are the most commonly used techniques for this task.

Polygonal approaches are used to approximate the shape boundary using the polygonal line. These methods are based on the use of the minimal error as approximation criteria. One of the most popular methods in this group is the split and merge algorithm [91]. Splines are also very popularly used for the interpolation of functions and the approximation of curves [92]. Especially, B-splines have the advantage that the local function value change does not spread to the rest of the intervals.

Both approaches agree on the significance of high curvature points for visual perception. Extraction of critical points with high curvature as feature points for shape recognition has been previously investigated [93–97]. For example, Chen and Su [95] adopt a maximum curvature approximation to derive feature points. Little and Gu [96] use the trajectory path and speed curve for motion representation. The path curve records the position temporal information of the object and the speed curves records the magnitude of its velocity. The maximum curvature of feature points, angles between successive segments, and the relative lengths of adjacent segment are calculated. A warping method is adopted for matching this curve. Hierarchical search is used for angle length comparison. In Bashir and Khokhar's [93] study, spatio-temporal curvature is used to represent the trajectories. Dominant inflection points from

the curvature are simultaneously extracted at multiple levels of scale revealing the structure at varying levels of details. These inflection points, the maxima of curvature scale space (CSS), are then use for indexing and retrieval. Representation of the CSS is computed by convolving a path-based parametric representation of the curve with a Gaussian function, as the standard deviation of the Gaussian varies from a small to a large value. Object trajectory indexing and retrieval inspire shape matching for CSS image analysis. A ranked list of sorted trajectories with query trajectory is output.

Wai and Chen [98] also attempted to solve the problem of curve matching with sketch query of object trajectory. High curvature points are extracted as feature points. Sub-matching and approximate-matching is performed by aligning these feature points.

3.1.1.2. Model based and pdf-based matching. This approach aims to train and build a knowledge-based model to recognise motion trajectory through mapping the spatial information in each frame or the differences in spatial information between successive frame-pairs into a temporal sequence of features. It attempts to estimate a set of model parameters to minimise fitness errors or modelling costs from the video data and use them to recognise the activity. State space models have been widely used for prediction, estimation, and detection of discrete spatio-temporal data. One representative model is the hidden Markov model (HMM), which can be summarised as a hidden Markov chain with a finite set of output probability distributions [99,100]. Each state is connected by probabilities to other states or its own, and an observation is derived from each state. HMM uses the Baum–Welch (forward–backward) algorithm for maximum likelihood estimation of the model parameters. HMM has been used widely in speech recognition. HMM has also been adopted for the recognition of motion sequences to model temporal structure of action such as learning object or camera movement and behaviour models [88,101], gesture recognition [102], and more recently activities segmentation from continuous surveillance videos [103].

Yamato et al. [88] employed a HMM probabilistic model for the classification of different human motions. They use a sequence of symbols, one per frame, derived from a mesh feature at the image level. Example sequences are used to train HMMs to match an unknown sequence with a trained model by analysing the probability distribution.

Ariki and Sugiyama [101] developed a TV news retrieval system which could automatically classify articles using a keyword spotting technique. The keyword spotting technique can extract a keyword sequence with their probabilities and the extracted keywords are attached to the article for retrieval. The TV news article can be classified into topics such as politics, economy, and science and so on by integrating acoustic keyword probability and topic contribution probability of the keyword, which is the probability to show how a keyword contributes to classify the article. TV news is retrieved by speech including the keywords attached to the articles. They employ normalised Viterbi method which computes the word probability by forward heuristics only.

Psarrou et al. [102] describe a statistical dynamic framework to model and recognise temporal structures of human activities based on prior learning and continuous propagation of density distribution of behaviour patterns. In their approach, prior knowledge is learned from training sequences using HMMs and density models are temporarily augmented by current visual observations. In their study, walking motions and gestures are recognised.

Yacoob and Black [104] propose a framework for modelling and recognition of temporal activities. The modelling of sets of exemplar activities is carried out by parameterising their representations in the form of principal components. Recognition of spatio-temporal variants of modelled activities is carried out by parameterising the search in the space of admissible transformations that the activities can undergo.

Xu et al. [105] used a probability approach to characterise the motion patterns based on these features obtained from energy redistribution of the motion vector field to classify basketball videos into 16 events such as team offence at left court and fast break to left, etc. In their basketball analysis system, they consider these semantic events and trained a HMM model for each event. Shots in basketball video can be considered as sentences composed of those events. They automatically build the semantic net through training data to add knowledge rules in recognition. Finally, Viterbi algorithm is used to segment and recognise events in shots. A classification rate of 75% is obtained. Similarly, Petkovic and Jonker [106] also use HMMs to recognise events in tennis videos.

A few other studies have used probability distribution function (pdf)-based distance to compare similarity. Fablet and Bouthemy [107] rely on a statistical approach for motion-based object indexing and retrieval. The local motion of polygon regions marked by users is extracted. The motion of the marked object is modelled using a kernel density estimator. Similarity comparison is based on computing Kullback–Leibler divergence.

3.1.1.3. Distance- and classification-based comparison. Distance-based comparison is the most common approach for motion trajectory matching. When spatio-temporal information is represented as feature vectors, usually a similarity distance is calculated between a model and an unknown input feature vector, such as Petajan et al. [108], Finn and Montgomery [109], JACOB [110], VideoQ [111] and Ioka and Kurokawa [112]. The model with the smallest distance is taken to be the class of motion to which the input belongs.

Dagtas et al. [113] use a trail-based model for video retrieval. The motion of salient objects over a sequence of frames is traced. The trail image is generated to use for the trajectory comparison. The authors use three methods to compare trajectory images: the first one is absolute search by finding the summary value of normalised dot product between a query image and database trajectory images; the second one is spatial-invariant-search by finding maximum value of normalised dot product between a Fourier transformed query image and database trajectory images; the last one is scale-invariant-search using a Mellion transform by finding minimum value of normalised Euclidean distance between Mellion coefficients.

Stefanidis et al. [17] and Eickhorst et al. [114] model an object trajectory and its outline as a helix representation with 3D spatio-temporal point trajectory (x,y,t). Similarity comparison is performed by computing Euclidean distance at each node of the helixes to determine whether both are exhibiting similar behaviours.

Once spatio-temporal features are extracted from motion trajectory slice as feature vectors, they can then be clustered for mapping to spatio-temporal similarity. For example, Nelson and Polana [115] assume that similar motion trajectories would generate similar feature vectors, which could be classified using a nearest centroid classifier, or any other classifier. Dimitrova and Golshani [116] use macroblock tracing and clustering to derive trajectories, and dynamic programming to determine similarity between trajectories.

Goddard [117] uses neural networks to map temporal similarity. This representation consists of an ordered sequence of events which are coordinated by temporal and motion events. A hierarchy representation is used: at the low-level, the presence of a low-level feature triggers an event which is sent to the higher layer. Combination of events at the this level trigger other events at higher levels, and so on, until the coordinated sequence of events of a body in motion could trigger one motion model up to the output level, representing the global motion of walking, running or skipping.

3.1.1.4. Template matching-based comparison. This approach requires building a set of templates in advance. Comparison is performed by finding the maximum correlation between the existing templates and the current test pattern. For gesture recognition to identify actions, Martin and Shah [118] use dense optical flow fields over a region, and compute correlation between different sequences for matching. Niyogi and Adelson [119,120] use temporal matching trajectory slices (or called as the *xt* plane) for gait recognition based on reciprocating path traced by a walking person's shoe to reveal the special braded pattern generated by walking in space–time. Polana and Nelson [121] recognised repetitive motion activity by matching spatio-temporal templates of motion feature, such as total motion magnitude, to classify the activity into one of several known classes. Efros et al. [122] built a smoothed and half wave rectified motion descriptor for recognising low resolution object actions. Normalised correlation is computed to compare the similarity of actions.

More recently, Briassouli and Ahuja, [123] applied the short-term fourier transform (STFT) to capture and recognise repetitive actions. They created a frequency-modulated (FM) signal in x and y projection for STFT of object action and compute the corresponding time-varying power spectrum of each FM signal. Thereafter, they determine the correlation between their templates and an inputted object action.

3.1.1.5. Multi-resolution matching. The match of motion trajectory should be executed in multi-resolution to get good fitness of the curve. Scale-space of motion trajectories has been analysed in different ways. Rangarajan et al. [124] compute the diffused scale-space of speed and direction of different points extracted from their trajectory. They argue that for similar motions, the scale-space would be similar, such that the point by point difference between the scale space of the speed and direction curves from different points undergoing similar motion would be much smaller than of points with very different motions.

Sahouria and Zakhor [125] analysed surveillance videos based on wavelet analysis of object trajectories in x and y directions $X=[x_1,x_2,\ldots,x_n]$ and $Y=[y_1,y_2,\ldots,y_n]$. The first eight coefficients of the Haar wavelet transform are stored for matching.

Multi-scale analysis is also performed to represent trajectories for both spatial and spatio-temporal dimensions by Chen and Chang [126]. The motion is presented with trajectory signals $(x(t),y(t))$. Wavelet analysis is applied to partition them into sub-trajectories. The raw object trajectory is decomposed into a hierarchy of coefficients at different scales. The coarsest scale components are adopted to approximate a smoothed trajectory and the finer scale components are used to partition the global motion into sub-trajectories. Each sub-trajectory is then modelled as a feature vector. Mahalanobis metric is used to calculate distance between feature vectors.

Multi-scale pyramid tracking based on the Mean-Shift algorithm is used by Wexler et al. [127] for video completion or inpainting missing data with small object or local portion of object. The process involves three steps of moving object alignment, inpainting of missing object portions, and background texture synthesis. Feature points are characterised by pixel (R,G,B) value, horizontal motion, and vertical motion. Logarithm function of similarity between feature points is computed by L2-norm distance. To simplify object modelling, they adopted a mounted camera and without camera motion is considered.

3.1.1.6. Chain code-based comparison. Chain code approximates object trajectory by using a set of orientation primitives and represents each segment of trajectory with a directional primitive. The directional primitives quantise the space into eight basic directions with symbols from 0–7. The distance between two symbols (a) and (b) of chain code is $\min(|a-b|,8-|a-b|)$. On the other hand, differential chain code approximates the trajectory with piecewise linear segments and codes each segment with a directional primitive that is relative to the last segment with respect to direction (left or right) and segment length.

Chain codes describe the trajectory through a sequence of unit-size line segments with a given orientation [128]. Several video retrieval studies represent spatio-temporal trajectory in the form of chain codes, such as Lee and Kao [129], Yoshitaka et al. [130], Li et al. [131]. Examples of video retrieval systems using chain code-based trajectory comparison include VideoRoadMap [132] and VIOLONE [130]. Similarly, Lee and Kao [129] present a qualitative description which enables subsequence query matching by using chain code representation. Object motion is represented using a combination of the 12 primitive motion types for translation, translation in depth, rotation, and rotation in depth.

3.1.1.7. Autocorrelation-based matching. Autocorrelation-based matching can be used to detect a variety of periodic and repetitive motion trajectories. It seems reasonable and sensible to use Fourier transforms to emphasise self-similarity and detect repetitive motion. Furthermore, it is more robust to uncorrelated noise, and it possesses several desirable invariances, for instance to spatial and temporal translation and scale.

Tsai et al. [133] introduced the spatio-temporal curvature for cycle detection. The 2D trajectory of a point on an object that performs cyclic motion is used to compute curvature as a function of time. Autocorrelation is calculated to emphasise self-similarity within the curvature function. Fourier transform is finally applied to detect the presence of cycles and their period. A high impulse indicates the presence of cycles and their fundamental frequency. Albu et al. [134] found the temporal segmentation of cyclic human motion by computing periodicity score through comparing the corresponding maxima of autocorrelation of a non-ideal periodic signal of average period with the one of an ideal periodic signal of exact period. Then iteratively merge coherent neighbour individual segments based on the periodicity score into a global segmentation.

3.1.1.8. Other methods of motion trajectory matching. Motion trajectory modelling based on simple chain-code or B-splines does not completely capture the motion trajectory characteristics [126]. The motion trajectory-based approach is sensitive to noise and time-scale. No matter which technical approach is taken, motion trajectory-based approaches are highly dependent on time, whereas most spatial relations remain consistent during the movement of video objects. Hence, combination of recognition of motion trajectory with analysing evolution of spatial relations for video retrieval can make video retrieval approaches more robust against time-scale dependency.

Li et al. [131] presented a novel scheme for matching the trajectory of a single moving object by comparing the moving object's directional and topological relation difference for each moving step. They store the object movement, directional relation and topological relation in a linked list. The topological rationale is based on the *closeness* of these relations when they evolve from one into the other within a time interval. Extending Egenhofer and AI-Taha's [50] qualitative definition of eight typological relations, they further elaborated on the qualitative representation of eight differential/distance of direction relations and of four positional relations, when these relations evolve from one to the other.

Wai and Chen [98] proposed the second approach for curve querying. The motion trajectory is modelled as a sequence of peaks by extracting high curvature points. The orientation of the peak is coded by chain code or modified chain code. The angle of the peak is binned by dividing 0°–180° into eight partitions. The temporal information "*time*" records the total number of frames in which the

symbol object moves through the peak. With each peak characterised by its orientation, angle and temporal information, the motion track is transformed into a string of code triples. Consequently, the curve matching problem is converted into a string matching problem. A finite automata-based matching method is used for efficient query processing.

3.1.2. Space–time trajectory volume representation and matching

In contrast to the analysis of trajectory slices that only exploit a small portion of the available video data, space–time trajectory volume is more attractive due to its rich information content. Recent research trend has been to unify the analysis of spatial and temporal information by building a volume of spatio-temporal data in which consecutive images are stacked to form a temporal dimension. A sequence of such 2D moving object contours with respect to time generates a spatiotemporal object trajectory volume in (x,y,t) space which can be treated as a concatenation of object silhouettes. The frame intervals are assumed to be sufficiently short to allow treating 3D trajectory as a continuous-time 3D shape as well as to allow derivatives to be calculated. The full benefits of analysing this volume are realised when the images are frequently sampled sufficiently to preserve spatial and temporal domains continuity. In such a scenario, the complexity of feature correspondence is significantly reduced, and occlusion events are made much easier to untie and detect.

Existing approaches of matching 3D shapes can be divided into two categories: volumetric feature- and model-based. Feature-based approaches are inherently more general-examining raw pixel data (x,y,t) at the expense of higher sensitivity to noise. A video sequence can be represented as a feature vector, with straightforward indexing, and retrieval can be implemented efficiently using nearest neighbour search. Shapes can be compared quickly by computing their distance in this space. Alternatively, a classifier can be trained with priori knowledge to identify an unseen sequence. Further data analysis can assist with pre-processing, such as normalisation, rotation, scale, translation to increase the discriminative abilities.

Local descriptor-based approaches build an interest point detector, then identify local structures in space–time where the pixel values (x,y,t) have significant local variations in both space and time, and finally construct the correspondence among interest points and associated events or gestures. The technique assumes that one can reliably detect a sufficient number of stable interest points in the video sequence. In the spatial domain, interest points are the points with a high local intensity gradient, such as corners and edge points. Interest points detection have been extensively investigated in the past with successful applications for image indexing, object tracking and recognition. These space–time interest points will be spatial interest points with a distinct temporal location corresponding to an instance characterised by variations of object motion in magnitude or direction in local spatio-temporal vicinity. There points should be distinctive enough to reliably identify the local spatio-temporal change. In the video sequence volume, if there are the points where an object rapidly changes its direction of motion, an instance of critical events can be reported. For example, consider scenes with a car crash event or with a person walking. Of course, local descriptors should also be robust to geometric perturbations and noise. Local descriptors also have the advantage of trajectory alignment with properties of free-viewpoints and robustness to temporal extents.

Laptev [135] extend the notion of interest points of image into the spatio-temporal domain. A scale-invariant Harris–Laplace interest point detector is derived by using a second moment matrix integrated over a Gaussian window over space and time (x,y,t). They show that analysis of "space–time" corner feature pointers can generate an understanding of events and actions in video data.

Yilmaz and Shah [136] generate 3D action trajectory volume by contour tracking cross time interval. Their contour tracking is performed based on the correspondence between contour points by computing the maximum matching of the weighted bipartite graph. Subsequently, high curvature points extracted from the surface of the action volume that are used for matching two actions. At the end, matching is performed by computing the correlation between two matrices composed by high curvature point sets, respectively. The observation matrix is constructed using coordinates of high curvature points from the two matrices. The smallest singular value (the ninth eigenvalue) of the observation matrix corresponds to the best match of trajectories. A similar method is also studied by Rao et al. [137], in which 3D epipolar geometry constraint for view-invariant alignment is used.

Unfortunately, since only the points from the contour are used for tracking, these techniques fail to detect actions when insufficient useful space–time interest points are available, in situations where the smooth motions contain no sharp extrema, or such points are missing or have errors due to occlusions, changing illumination, reflectance, shadows as well as noise. It is quite possible that there are so few such points in a typical motion, it is difficult to trace these interest points to support event recognition. In order to solve the problem of actions with very few feature points, an approach of space–time patches (ST-patches) template comparison for registering two action video clips is presented in Shechtman and Irani [138]. They attempt to match two space–time patches (ST-patches) on gradient of motion fields by examining the rank-increase measure of their joint matrix. Finally, multiple templates are correlated against the same video sequence to detect multiple different activities.

A number of other volumetric feature-based approaches have been used in literature. Examples include Lihi and Irani (2001), Ke et al. (2005), Weinland et al. [139], and Gorelick et al. [140]. Zelnik-Manor and Irani [141] regard an event as a stochastic temporal process associating an empirical distribution of space–time gradients, which are generated from an entire video volume. Weinland et al. [139] compute the Fourier magnitudes from the motion history volume as the descriptors for human action recognition. They adopt cylindrical coordinates, centered on bodies, to express motion view invariance to locations and rotations around the z-axis. Their work assumes stationary multi-camera and static background scenes, and that similar actions only differ by rigid transformations.

In Gorelick et al. [140], human action is considered to be a motion history volume derived from a continuous silhouette sequence of a moving torso and protruding limbs. The moving silhouette and time axis constitute space–time volume and represent it as a shape trajectory. Human activity recognition problem is expressed as the task of comparing two 3D shapes by solving the Poisson equation for each point in the silhouette stack. A set of features are extracted to identify space–time saliency of moving parts and locally judge the orientation and rough aspect ratios of the space–time shape. The nearest neighbour and a variant of the median Hausdorff distance are used for action classification and recognition.

A drawback of volumetric feature-based approaches is that partial shape matching is not supported, because features are extracted from the whole of 3D shape. In contrast, model-based methods are well suited for partial shape matching. Geometry model-based methods such as skeletons take into account approximate topology matching. Model-based approaches generally resort to either fitting a predefined structure to a 3D trajectory volume, or matching against predefined motion models [142].

Since the structures of the spatio-temporal volume are inherently non-rigid, an alternative popular approach is to model the space–time volume, i.e., to construct a deformable model for iterative evolvement to fit a 3D spatio-temporal surface. A good example of this is Hamameh et al. [74], who extend the principle of 2D active shape model (ASM) to build a deformable spatio-temporal shape model by incorporating *a priori* knowledge of object shape. The statistical shape model is defined by the constraints from a set

of landmark points picked from the associated 3D trajectory volume. A 2D object shape varying with time evolves dynamically to approximately fit the real geometry of the 3D object trajectory. An energy function is minimised through the object shape deformed iteratively by using dynamic programming until the energy function converges.

Similar studies have been performed by Niyogi and Adelson [119,120] and Baumberg and Hogg [143]. Baumberg and Hogg derive a deforming model for fitting 3D trajectory shape surface to recognise human gait. They use unconstrained second order system identification technique to automatically learn physics-based 'vibration modes' for localising and tracking a specific deformable object. The $n \times n$ second order system allows to be decoupled into n independent second order systems. Each pedestrian contour shape is represented by a B-spline with 40 uniformly spaced control points. A 3D eigenshape-based generic walk deforming model is built from a set of training data for recognising non-rigid motions. The signal to noise ratio is calculated based on the nodal displacements relative to the mean shape over the whole training set.

Hsieh et al. [144] introduce triangulation-based skeleton modelling scheme to analyse human behaviour by tracking silhouette sequences. Triangulation-based skeleton and the centroid context (CC), as two key features, are extracted to perform coarse-to-fine search for posture recognition. The skeleton is built from triangular meshes and is used for spanning tree-based pruning and search. CC descriptor utilises a polar labeling scheme to label every triangular mesh with a unique number. Finally, a silhouette sequence is coded into a semantic symbol string. Weighted edit distance is used to measure the similarity between posture strings for posture identification.

A drawback of the model-based approach is that an elaborate model must be constructed for each object motion to be tracked. Despite that 3D spatio-temporal volume is more informative than 2D spatio-temporal slice for classifying activities or postures, these approaches are is of limited use for real-time applications due to the inherent correspondence problem and high computational cost [144]. Moreover, the presence outliers or noises from shadow, clutter, or occlusion, et al. restrict the accuracy of space–time shapes. A more detailed survey on 3D shape matching can be found in Tangelder and Veltkamp [145].

3.1.3. SfM representation and recognition

Motion features are generally extracted within a shot by matching consecutive frames through pixel blocks-based search. Motion information can be classified as global or dominant motion, and region motion. Global motion is statistically identified as camera motion, such as pan, tilt and zoom, whereas region motion is characterised as object motion. Using global information, object motion can be isolated by compensating for the global motion. Motion information can be used for quantifying objects within the video sequence and tracking trajectories of objects. Temporal information, usually addressed in the context of motion detection, can provide extra cues about the content, shape, structure, and other high or low level information present in a shot for video retrieval.

Motion vectors and optical flow are two commonly used methods for extracting 2D motion information. Motion vectors are approximated based on the movement of macroblocks in MPEG, while optical flow consists of the computation of the displacement of each pixel between frames in uncompressed videos. Motion vectors provide a rough and sparse approximation to real optical flows. Frame-based motion is used to identify camera motion, whereas region-based motion indicates object movement.

The SfM approach tries to capture 3D geometry from relative motion information between a camera and an object. Motion-based object segmentation is based on the understanding that pixels associated with an object tend to move in a coherent fashion. The motions reveal the contour of a shape. Three-dimensional shape can be perceived and inferred from motions. SfM recovers a shape from motion-induced spatial and temporal changes occurring in an image sequence. The techniques exploit the relative motion between camera and scene. SfM methodology addresses two subproblems: feature correspondence and shape and structure reconstruction.

Several solutions have been proposed to tackle the SfM problem. One of the most influential of these was proposed by Tomasi and Kanade [146]. They recovered a shape matrix from image sequences with the use of the singular value decomposition for rigid objects. Vidal and Hartley [147] also introduced a geometric approach in three perspective views to recover 3D multiple rigid-body motions from point correspondences by ranking trilinear constraints and computing its associated multi-body trifocal tensor.

These methods are limited to rigid objects or/and static scenes, whereas there are many applications with the scenarios of non-rigid or dynamic objects in the real world. Examples include a walking person, moving vehicles, human faces expressions, lip movements, etc. Since the shape of non-rigid object deformation varies from frame to frame, minimising the registration error is a more difficult task than that of the rigid ones. Several extensions have been proposed to relax the rigidity constraint. For example, Bregler et al. [148] detail how to rebuild the non-rigid shape by using factorisation. They express a 3D shape of non-rigid object as a weighted linear combination of a set of shape bases that define the principal modes of object deformation under weak perspective viewing conditions. In Torresani et al. [149], an algorithm is presented from learning the time-varying shape instance from the motions of 2D tracking points by estimating a Gaussian distribution in each frame. The non-rigid object's motions are supposed to consist of a rigid component plus a non-rigid probabilistic shape deformation. Prior information on the motion or shape is introduced to avoid ambiguities.

Unlike previous work that analyses 2D deformations, recently Wang and Wu [150] assume that the non-rigid object is composed of a rigid part and a deformation part. They address the problem of 3D reconstruction of non-rigid object parts by using a deformation weight constraint for non-rigid factorisation and using constraint power factorisation (CPF) on an uncalibrated affine camera model. Their algorithm recovers the structure in affine space and separates the rigid features from the deformed ones, before estimating the transformation from affine to metric space.

3.1.4. Multi-camera tracking with view invariance

Recently significant research effort has focused on understanding object activities under multi-camera systems for surveillance, security and industry assembly purposes (e.g., Ref. [151] used fixed-location multiple cameras to detect unusual events by thresholding optic flow histogram distance for either its direction or magnitude). The understanding of object activities can form the basis of understanding video content which is key to video retrieval and similarity matching. A detailed survey on multi-camera tracking is given by Stoykova et al. [152]. With multiple cameras, an action may be observed from different viewpoints simultaneously. For object activity analysis using multiple cameras, a widely used approach to achieve view invariance is to use the fundamental matrix between two different views of an action to solve the correspondence problem. By imposing geometric constraints, the geometrical relationships across cameras between the motion trajectories of each object are reconstructed. At the same time, the geometric error function is minimised by imposing temporal consistency in terms of a motion model.

Estimating the fundamental matrix from a set of point correspondences to express the constraint relationships among pixel coordinates, camera orientations and positions has been suggested earlier by Hartley and Zisserman [153], Rao et al. [137], and Yilmaz and Shah [154]. These approaches can be classified into linear and nonlinear methods. The most common linear methods are

the normalised eight-point algorithm [155], and the rank theorem [146,156,157].

A number of approaches have addressed how to track human and other object actions in videos, especially by considering kinematic constraints using inherent landmarks of human body (e.g., 13 articulated joints). Example studies include Parameswaran and Chellappa [158], Yilmaz and Shah [154], and Gupta et al. [159]. Yilmaz and Shah [154] track trajectories of 13 articulated joints of human body with two moving camera views. Action recognition is performed by modelling camera motions and recovering scene geometry for each video frame. View invariance is achieved by estimating a temporal fundamental matrix of the epipolar geometry between two different views of an action.

Similarly, Gupta et al. [159] unify of constraints, including the kinematic constraints, the occlusion between body-parts and appearance consistency across body-parts, to track human skeleton for multi-view occluded human pose estimation. For body search initialisation, epipolar constraints are applied to match across views to obtain a rough localisation of faces in 3D. The posterior of each body-part is optimised by using non-parametric belief propagation.

One of the key difficulties in analysing multi-camera systems is the dependence of the analysis on viewpoint. Viewpoint invariance can be achieved by extracting view invariant features from the video. Parameswaran and Chellappa [158] use five body joints as the landmark points aligned into a plane in 3D to extract a set of projective invariants for actions matching. Another strategy is to achieve invariance is through transformation, rotation and translation an object of individual views to a new coordinate system associated cameras, and then performing feature extraction and normalisation. Weinland et al. [139] align actions in variety of viewpoints around the central vertical axis of the human body using Fourier transform in cylindrical coordinates.

3.2. Video sequence-to-sequence-based matching

Unlike trajectory-based approaches that track moving objects only, sequence-to-sequence-based matching takes video frames as input and the analysis is based on all pixels in video frames. Trajectory-based approach can align temporal video sequences with different background and viewpoints by using explicit geometric constraint. The main limitation of tracking trajectory-based approaches is that its accuracy is affected heavily by false alerts due to tracking failure or creation of false targets (fragmentation of targets, shadows, self-occlusion, clutters, etc). Since sequence-to-sequence-based approach uses low-level information without the need for tracking, it is simple, direct and suitable to general purpose video retrieval. Within the overall umbrella of sequence-to-sequence matching methods, *motion-based comparison*, *string matching* and *knowledge-based modelling* are the three key approaches.

Video similarity can be computed based on *motion feature comparison*. For instance, Polana and Nelson [160] compute several features from the normal flow (component parallel to the gradient) of the whole image. One of them is the average flow magnitude divided by its standard deviation. In Dimitrova and Golshani [161], motion trajectories are obtained by tracing the position of a macroblock. The trajectory of a macroblock is computed from forward and backward motion vectors that belong to the macroblock. The position of a macroblock in a P-frame is computed using block coordinates and forward motion vectors. The position of a macroblock in a B-frame is computed by averaging the positions obtained from (1) the next predicted block coordinates and the backward motion vector and (2) the previous block coordinates and forward motion vector. Each trajectory can be thought of as an n-tuple of motion vector. The macroblock trajectories are feature vectors used for indexing. Virage Video Engine [162] also includes motion as an indexing feature. They have four descriptors for motion description: motion content, motion uniformity, motion panning, and motion tilting. Motion content is defined as the total amount of motion within a given video. Motion uniformity is used to capture the smoothness of the motion as a function of time. Motion panning and motion tilting are used to compute the horizontal and vertical motion components of the motion within a video sequence. Motion query is performed by ranking a collection of videos based on the motion properties of choice. JACOB [110] is a colour- and motion-based video search engine. The user is asked to choose global motion orientation and magnitude within each quadrant of the frames. Motion-based descriptors are based on the optical flow field of the r-frame. VideoQ [111] emphasises motion as a key attribute in searching video databases and uses optical flow. Object colour regions are segmented and tracked over the duration of a video shot. Motion estimation uses a hierarchical block matching method. During the tracking process, global motion compensation algorithms are applied so that the final object trajectory is independent of camera motion. Similarly, Ioka and Kurokawa [112] propose a method for retrieving image sequences by using motion information. Motion information is derived from motion vectors from the image sequences using block matching. After aggregation analysis, several representative motion vectors are generated and their representative trajectories are stored in the database to use for video retrieval.

The principle of the *string matching* is to find the longest common sub-sequence across two videos [49,163,164]. For this purpose, minimum editing distance is determined for video sequence matching. The method is quite flexible and can be used to compare different length video sequences.

Adjeroh et al. [163] propose a method called v-string matching. The video sequence is initially described by a sequence of feature values and transformed into a sequence of symbols which is a string representation—called *vstring*. Edit distance is used on this to calculate the distance between two *vstrings*. The similar study is also investigated by Yazdani and Ozsoyoglu [164].

Knowledge-based modelling approaches have been successfully applied in spatio-temporal sequence-to-sequence matching. They rely on rules derived from training data and can provide better predictive performance when the training set is large and comprehensive. These methods include the use of recursive techniques, either as an estimation fusing mechanism or as a state estimator.

Rui and Anandan [165] address the problem of detecting action boundaries in a video sequence containing unfamiliar and arbitrary visual actions. Their approach is based on detecting temporal discontinuities of the spatial pattern of object region motion which correspond to the action temporal boundary to capture the action. They represented frame-to-frame optical flow in terms of the coefficients calculated from all of the flow fields in a sequence, after principal components analysis to determine the most significant flow fields. The temporal trajectories of those coefficients of the flow field are analysed to determine locations of the action segment boundaries of video objects.

Naphade et al. [166] introduced a method for video annotation using Gaussian mixture model for frame object recognition based on visual appearance features. Bayes rule was applied temporally in image sequence forward and backward directions.

Smith et al. [167] use knowledge-based approach to learn action or pose configurations from training images. They track human activities for office sign language recognition by using a boosting classifier to temporally group individual frame classifiers, which acquire actions by considering spatial tracks of hands and head regions in all views as features.

Nguyen et al. [168] introduced a multi-target tracking approach where the distribution of targets in current frame are statistically measured with an incrementally learning PPCA (probabilistic principal component analysis) from the targets' past PCA appearance

models. Temporally updates are performed upon the arrival of new classification results against local spatial context.

Non-parametric model-based approaches have also been investigated in video indexing. Fablet and Bouthemy [169] propose an approach for non-parametric motion analysis in video sequence. This method relies on the statistical modelling of distributions of local random walks. The local motion-related measurement is a weighted local average of the normal flow. Drawback of this typed approach lies in that recursively update estimation from frame to frame may fail catastrophically if visual ambiguities caused by noises or occlusions persist over several consecutive frames [170].

4. Spatio-temporal video indexing systems

An integrated system for spatio-temporal video retrieval is LucentVision. LucentVision [12] was developed at the Visual Communications Research Department within Bell Labs. It was effectively used for tennis video indexing through spatio-temporal activity maps. This system analyses video from multiple cameras in real-time and captures the activity of the players and the ball in the form of motion trajectories. The system stores these trajectories in a database along with video, 3D models of the environment, scores, and other domain-specific information. LucentVision enhances live television and Internet broadcasts with game analyses and virtual replays. LucentVision uses eight cameras placed around a tennis stadium to track the players and the ball. Real-time analysis of video from these cameras determined the motion trajectory for each player and the ball. For a query on tennis match, selection includes *score-based queries* (e.g., *all points won by a player against opponent's serve*), *statistics-based queries* (e.g., distances run by players on average) or *space-based queries* (e.g., *all points at the net*) and *historical queries*. Each query could be refined further using a time constraint, for example limiting it to one set, one game, or even any particular match period (e.g., the first 30 s of the second set). LucentVision stores 250 videos of international tennis matches in its database. Each data selection generates an SQL query to the database. Retrieved motion trajectories can be viewed using a number of visualisation and animation choices, such as coverage maps of motion trajectories, end-position maps, and speed charts (the speeds of the players). End-position maps are created by mapping the last node of a player's trajectory onto the virtual court.

5. Conclusions

Video retrieval is essentially the task of finding the most similar video based on a query video. Traditionally, text-based labels attached to videos were used for matching. Since the 1980s, significant research into image analysis opened up the possibility of extracting image content information from these videos which could form the basis of matching, ranking and retrieving them. Over the recent years, it has been recognised that raw pixel information and basic statistical features of colour distribution are not enough in discriminating video content and consequently video retrieval quality is severely affected. Advancements in the areas of image content understanding and pattern recognition approaches have now made it possible to recognise image objects, model their behaviour and use spatio-temporal change information in the retrieval process. Our review has focused on how best the spatio-temporal information can be extracted from videos, represented and matched for improving video retrieval measures of precision and recall. We hope that this has provided the reader with an overview of the wide-ranging research efforts in this area and will help focus their own research.

References

[1] M. Roach, J. Mason, N. Evans, L.Q. Xu, F. Stentiford, Recent trends in video analysis: a taxonomy of video classification problems, in: M.A. Hamza (Ed.), Proceedings of Internet and Multimedia Systems Conference, Kauai, USA, Acta Press, 2003.

[2] J. Li, N. Allinson, D. Tao, X. Li, Multi-training support vector machine for image retrieval, IEEE Trans. Image Process. 15 (11) (2006) 3597–3601.

[3] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machine-based relevance feedback in image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1088–1099.

[4] D. Tao, X. Tang, X. Li, Y. Rui, Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm, IEEE Trans. Multimedia 8 (4) (2006) 716–727.

[5] D. Tao, X. Li, S.J. Maybank, Negative samples analysis in relevance feedback, IEEE Trans. Knowl. Data Eng. 19 (4) (2007) 568–580.

[6] D. Tao, X. Tang, X. Li, Which components are important for interactive image searching?, IEEE Trans. Circuits Systems Video Technol. 18 (1) (2008) 3–11.

[7] A.D. Bimbo, E. Vicario, D. Zingoni, Symbolic description and visual querying of image sequences using spatio-temporal logic, IEEE Trans. Knowl. Data Eng. 7 (4) (1995) 609–622.

[8] M. Erwig, M. Schneider, Visual specifications of spatio-temporal developments, in: Proceeding of 15th IEEE Symposium on Visual Languages, 1999, pp. 187–188.

[9] J.K. Wu, A.D. Narasimhalu, B.M. Mehtre, C.P. Lam, Y.J. Gao, CORE: a content-based retrieval engine for multimedia information systems, ACM Multimedia Systems 3 (1995) 25–41.

[10] J.P. Cheylan, S. Lardon, Toward a conceptual model for the analysis of spatio-temporal processes, in: A.U. Frank, I. Campari (Eds.), Spatial Information Theory, Springer, Berlin, 1993, pp. 158–176.

[11] C. Claramunt, M. Theriault, Toward semantics for modelling spatio-temporal processes within GIS, in: M.J. Kraak, M. Molenaar (Eds.), Advances in GIS Research 1, Taylor & Francis, London, 1996, pp. 27–43.

[12] G.S. Pingali, A. Opalach, Y.D. Jean, I.B. Carlbom, Instantly indexed multimedia databases of real world events, IEEE Trans. Multimedia 4 (2) (2002) 269–282.

[13] B.F. Buxton, H. Buxton, Monocular depth perception from optical flow by space–time signal processing, Proc. R. Soc. London Ser. B Biol. Sci. 218 (1210) (1983) 27–47.

[14] E. Aldelson, J.R. Bergen, Spatiotemporal energy models for the perception of motion, J. Opt. Soc. Am. 2 (1985) 284–299.

[15] R.C. Bolles, H.H. Baker, D.H. Marimont, Epipolar plane image analysis: an approach to determining structure from motion, Int. J. Comput. Vision 1 (1987) 7–56.

[16] H.H. Baker, R.C. Bolles, Generalizing epipolar plane image analysis on the spatiotemporal surface, Int. J. Comput. Vision 3 (1989) 37–39.

[17] A. Stefanidis, P. Agouris, P. Partsinevelos, Spatio-temporal helixes for event modelling, in: Proceedings of the Third (USA) National Conference on Digital Government Research (dg.o), 2002, pp. 219–224.

[18] Y. Ricquebourg, P. Bouthemy, Real-time tracking of moving persons by exploiting spatiotemporal image slices, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 797–808.

[19] R. Wildes, J. Bergen, Qualitative spatio-temporal analysis using an oriented energy representation, Proc. Eur. Conf. Comput. Vision 2 (2000) 768–784.

[20] W.-S. Li, K.S. Candan, SEMCOG: a hybrid object-based image and video database system and its modeling, language, and query processing, Theory Prac. Object System (TAPOS) 5 (3) (1999) 163–180.

[21] J.F. Allen, Maintaining knowledge about temporal intervals, Commun. ACM 26 (11) (1983) 832–843.

[22] W.Y. Ma, B.S. Manjunath, A comparison of wavelet transform features for texture image annotation, in: Proceeding of IEEE International Conference on Image Processing, 1995.

[23] M. Stricker, A. Dimai, Color Indexing with weak spatial constraints, in: Proceeding of Storage and Retrieval for Image and Video Databases IV, SPIE, vol. 2670, 1996, pp. 29–40.

[24] R.M. Haralick, K. Shanmugam, I. Dinstein, Texture features for image classification, IEEE Trans. Systems Man Cybernet. 3 (6) (1973) 610–621.

[25] J. Huang, R. Zabih, Combining color and spatial information for content-based image retrieval, in: Proceeding of European Conference on Digital Libraries, 1998.

[26] J.R. Smith, C.F. Li, Image classification and querying using composite region templates, J. Comput. Vision Image Understand. 75 (1–2) (1999) 165–174.

[27] C.C. Chang, S.Y. Lee, Retrieval of similar pictures on pictorial databases, Pattern Recognition 24 (7) (1991) 675–680.

[28] S.K. Chang, Q.Y. Shi, C.W. Yan, Iconic indexing by 2D strings, IEEE Trans. Pattern Anal. Mach. Intell. 9 (3) (1987) 413–428.

[29] M.J. Egenhofer, R. Franzosa, Point-set topological spatial relations, Int. J. Geogr. Inf. System 5 (2) (1991) 161–174.

[30] V.N. Gudivada, V.V. Raghavan, Design and evaluation of algorithms for image retrieval by spatial similarity, ACM Trans. Inf. Systems 13 (2) (1995) 115–144.

[31] A. Cohn, Qualitative Spatial Representation and Reasoning Techniques, Lecture Notes in Artificial Intelligence, vol. 1303, Springer, Berlin, 1997 pp. 1–30.

[32] S. Lee, F. Hsu, Spatial reasoning and similarity retrieval of images using 2D C-string knowledge representation, Pattern Recognition 25 (1992) 305–318.

[33] S.-Y. Lee, F.-J. Hsu, 2D C-String: a new spatial knowledge representation for image database systems, Pattern Recognition 23 (10) (1990) 1077–1087.

[34] S.K. Chang, E. Jungert, A spatial knowledge structure for image systems using symbolic projections, in: Proceedings of the Fall Joint computer Conference, 1986, pp. 79–86.
[35] S.K. Chang, C.W. Yan, D.C. Dimitroff, T. Amdt, An intelligent image database system, IEEE Trans. Software Eng. 14 (5) (1988) 681–688.
[36] E. Jungert, Extended symbolic projections as a knowledge structure for spatial reasoning, in: Proceeding of Fourth BPRA Conference on Pattern Recognition, Springer, Berlin, 1988, pp. 343–351.
[37] S. Chang, E. Jungert, T. Li, Representation and retrieval of symbolic pictures using generalized 2D strings, Proc. Visual Commun. Image Process. IV SPIE 1199 (1989) 1360–1372.
[38] M.-C. Yang, 2D B-string representation and access methods of image database, Master Thesis, Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan, 1990.
[39] P.W. Huang, Y.R. Jean, Using 2D C+-string as spatial knowledge representation for image database systems, Pattern Recognition 27 (1994) 1249–1257.
[40] E.G.M. Petrakis, S.C. Orphanoudakis, A generalized approach to image indexing and retrieval based on 2-D strings, in: Intelligent Image Database Systems, World Scientific Publishing Co., Singapore, 1996, pp. 197–218.
[41] E.G.M. Petrakis, Image representation, indexing and retrieval based on spatial relationships and properties of objects, Ph.D. Thesis, University of Crete, Department of Computer Science, March 1993.
[42] P.W. Huang, Y.R. Jean, Spatial reasoning and similarity retrieval for image database systems based on RS-strings, Pattern Recognition 29 (1996) 2103–2114.
[43] S. Lee, M. Yang, J. Chen, Signature file as a spatial filter for iconic image database, J. Visual Lang. Comput. 3 (4) (1992) 373–397.
[44] J.R. Smith, S.-F. Chang, Integrated spatial and feature image query, Multimedia System 7 (2) (1999) 129–140.
[45] G. Costagliola, G. Tortora, T. Arndt, A unifying approach to iconic indexing for 2-D and 3-D scenes, IEEE Trans. Knowl. Data Eng. 4 (3) (1992) 205–222.
[46] A.D. Bimbo, M. Campanai, P. Nesi, A three-dimensional iconic environment for image database querying, IEEE Trans. Software Eng. 19 (10) (1993) 997–1011.
[47] C.C. Liu, A.L.P. Chen, 3D-list: a data structure for efficient video query processing, IEEE Trans. Knowl. Data Eng. 14 (1) (2002) 106–122.
[48] A.J.T. Lee, H.-P. Chiu, P. Yu, 3D C-string: a new spatio-temporal knowledge representation for video database systems, Pattern Recognition 35 (2002) 2521–2537.
[49] F.-J. Hsu, S.-Y. Lee, B.-S. Lin, Video data indexing by 2D C-Trees, J. Visual Lang. Comput. 9 (4) (1998) 375–397.
[50] M. Egenhofer, K.K. Al-Taha, Reasoning about gradual changes of topological relationships, in: Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, Springer, Berlin, 1992, pp. 196–219.
[51] M. Vazirgiannis, Y. Theodoridis, T. Sellis, Spatio-temporal composition in multimedia applications, in: Proceeding of IEEE International Workshop Multimedia and Software Development, Software Engineering, 1996.
[52] J.Z. Li, M.T. Ozsu, D. Szafron, Modeling of video spatial relationships in an object database management system, in: Proceeding of International Workshop on Multi-media Database Management Systems, 1996, pp. 124–132.
[53] R.Z. Liang, S. Venkatesh, D. Kieronska, Video indexing by spatial representation, in: Proceedings of the Third Australian and New Zealand Conference on Intelligent Information Systems (ANZIIS-95), 1995, pp. 99–104.
[54] J.Z. Li, M.T. Ozsu, Stars: a spatial attributes retrieval system for images and videos, in: Proceedings of the Fourth International Conference on Multimedia Modeling, 1997, pp. 69–84.
[55] C.-C. Hsu, W.W. Chu, R.K. Taira, A knowledge-based approach for retrieving images by content, IEEE Trans. Knowl. Data Eng. 8 (4) (1996) 522–532.
[56] M. Nabil, A.H.H. Ngu, J. Shepherd, Picture similarity retrieval using the 2D projection interval representation, IEEE Trans. Knowl. Data Eng. 8 (4) (1996) 533–539.
[57] P.R. Lipson, Context and configuration based scene classification, Ph.D. Thesis, MIT, EECS Department, 1996.
[58] S. Gautama, et al., Relevance criteria for spatial information retrieval using error-tolerant graph matching, IEEE Trans. Geosci. Remote Sensing 45 (4) (2007).
[59] B.T. Messmer, Efficient graph matching algorithms, Ph.D. Thesis, University of Bern, Switzerland, 1995.
[60] E.G.M. Petrakis, C. Faloutsos, K.-Ip. Lin, ImageMap: an image indexing method based on spatial similarity, IEEE Trans. Knowl. Data Eng. 14 (5) (2002) 979–987.
[61] E.G.M. Petrakis, C. Faloutsos, Similarity searching in medical image databases, Technical Report at University of Maryland, UMD: {CS-TR-3388}, UMIACS-TR-94-134 (extended version), 1994.
[62] V.N. Gudivada, G.S. Hung, An algorithm for content-based retrieval in multimedia databases, in: Proceedings of the 1996 International Conference on Multimedia Computing and Systems (ICMCS '96), 1996, pp. 56–61.
[63] I. Ahmad, W.I. Grosky, Spatial similarity-based retrievals and image indexing by hierarchical decomposition, in: Proceedings of the International Database Engineering and Application Symposium, Montreal, Canada, 1997, pp. 269–278.
[64] H. Yu, W. Wolf, A visual search system for video and image databases, in: Proceedings of IEEE multimedia computing and systems'97, 1997, pp. 517–524.
[65] Y.A. Aslandogan, C. Thier, C.T. Yu, C. Liu, K.R. Nair, Design, implementation and evaluation of SCORE (a System for COntent based REtrieval of Pictures), in: Proceedings of 11th International Conference on Data Engineering, IEEE-ICDE-11, 1995, pp. 280–287.
[66] A.P. Sistla, C. Yu, C. Liu, K. Liu, Similarity based retrieval of pictures using indices on spatial relationships, in: Proceedings of the 21st International Conference on Very Large Databases, Zurich, Switzerland, 1995, pp. 619–629.
[67] M.S. Lew, K. Lempinen, N. Huijsmans, Webcrawling using sketches, in: Proceeding of the Second International Conference on Visual Information Systems (VISUAL97), 1997, pp. 77–84 ⟨http://www.wi.leidenuniv.nl/~mlew/⟩.
[68] W. Hsu, T.S. Chua, H.K. Pung, An integrated color-spatial approach to content-based image retrieval, in: Proceeding of ACM Multimedia Conference, 1995, pp. 305–313.
[69] W.Y. Ma, B.S. Manjunath, Netra: a toolbox for navigating large image databases, Multimedia System 7 (1999) 184–198.
[70] W. Niblack, X. Zhu, J.L. Hafner, T. Bruel, D.B. Ponceleon, D. Petkovic, M. Flickner, E. Upfal, S.I. Nin, S. Sull, B.E. Dom, Updates to the QBIC system, in: Proceeding of IS&T SPIE, Storage and Retrieval for Image and Video Databases VI, vol. 3312, San Jose, 1998, pp. 150–161.
[71] J.R. Smith, S.F. Chang, Tools and techniques for color image retrieval, in: Proceeding of on Storage and Retrieval for Image and Video Database IV, SPIE, vol. 2670, 1996.
[72] M.A. Rodrfguez, M.C. Jarur, A genetic algorithm for searching spatial configurations, IEEE Trans. Evol. Comput. 9 (3) (2005).
[73] C. Freksa, Temporal reasoning based on semi-intervals, Artif. Intell. 54 (1992) 199–227.
[74] G. Hamarneh, T. Gustavsson, Deformable spatio-temporal shape models: extending ASM to 2D+ time, J. Image Vision Comput. 22 (2004) 461–470.
[75] M. Vazirgiannis, M. Hatzopoulos, Integrated multimedia object and application modeling based on events and scenarios, in: Proceedings of IEEE International Workshop for MMDBMSs, 1995, pp. 48–55.
[76] T.D.C. Little, A. Ghafoor, Interval-based conceptual models for time-dependent multimedia data, IEEE Trans. Knowl. Data Eng. 5 (4) (1993) 551–563.
[77] R. Hjelsvold, R. Midtstraum, O. Sandst, A temporal foundation of video databases, in: Proceeding of International Workshop Temporal Database, 1995, pp. 295–314.
[78] R. Weiss, A. Duda, D.K. Gifford, Composition and search with a video algebra, IEEE Multimedia 2 (1) (1995) 12–25.
[79] E. Oomoto, K. Tanaka, OVID: design and implementation of a video-object database system, IEEE Trans. Knowl. Data Eng. 5 (4) (1993) 629–643.
[80] R. Hjelsvold, R. Midtstraum, O. Sandst, Searching and Browsing a Shared Video Database, Multimedia Database Systems, Design and Implementation Strategies, Kluwer Academic Publishing, Boston, 1996 (Chapter 4).
[81] E.J. Hwang, V.S. Subrahmanian, Querying video libraries, J. Visual Commun. Image Representation 7 (1) (1996) 44–60.
[82] J.Z. Li, M.T. Ozsu, D. Szafron, Modeling video temporal relationships in an object database management system, in: Proceeding of SPIE Multimedia Computing and Networking (MMCN97), 1997, pp. 80–91.
[83] Y.F. Day, S. Dagstas, A. Ghafoor, Spatio-temporal modeling of video data for on-line object-oriented query processing, in: Proceeding of IEEE International Conference Multimedia (ICMCS '95), 1995, pp. 98–105.
[84] D. Papadias, Y. Theodoridis, Spatial relations, minimum bounding rectangles and spatial data structures, Technical Report KDBSLAB-TR-94-06, National Technical University of Athens, Greece, 1994, Also Int. J. Geogr. Inf. Systems (1996), to appear .
[85] N. Pissinou, I. Radev, K. Makki, W.J. Campbell, Spatio-temporal composition of video objects: representation and querying in video database systems, IEEE Trans. Knowl. Data Eng. 13 (6) (2001) 1033–1040.
[86] P. Salembier, R. Qian, N. O'Connor, P. Correia, I. Sezan, P. van Beek, Description schemes for video programs, users and devices, Signal Process. Image Commun. 16 (1–2) (2000) 211–234.
[87] W. Ren, S. Singh, Video sequence matching with spatio-temporal constraint, in: Proceeding of 17th International Conference on Pattern Recognition (ICPR'04), vol. 3, August 2004, pp. 834–837.
[88] J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden Markov model, in: Proceeding of Conference on Computer Vision and Pattern Recognition, 1992, pp. 379–385.
[89] B. Jahne, Spatio-temporal Image Processing: Theory and Scientific Applications, Springer, Berlin, 1993.
[90] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Comput. Vision Image Understand. 73 (3) (1999) 428–440.
[91] T. Pavlidis, Polygonal approximation by Newton's method, IEEE Trans. Comput. 26 (1977) 800–807.
[92] F.S. Cohen, Z. Huang, Z. Yang, Invariant matching and identification of curves using B-splines curve representation, IEEE Trans. Image Process. 4 (1995) 1–10.
[93] F. Bashir, A. Khokhar, Curvature scale space based affine-invariant trajectory retrieval, in: Proceeding of IEEE International Multitopic Conference (INMIC 2004), 2004.
[94] C.C. Chang, S.M. Hwang, D.J. Buehrer, A shape recognition scheme based on relative distances of feature points from the centroid, Pattern Recognition 24 (11) (1991) 1053–1063.
[95] H.H. Chen, J.S. Su, A syntactic approach to shape recognition, in: Proceeding of International Computer Symposium, Taiwan, 1986, pp. 103–122.
[96] J.J. Little, Z. Gu, Video retrieval by spatial and temporal structure of trajectories, in: Proceeding of SPIE Storage and Retrieval for Media Databases 2001, San Jose, 2001.
[97] F. Mokhtarian, A. Mackworth, Scale-based description and recognition of planar curves and two-dimensional shapes, IEEE Trans. Pattern Anal. Mach. Intell. 8 (1) (1986).

[98] T.Y. Wai, A.L.P. Chen, Retrieving video data via motion tracks of content symbols, in: Proceeding of ACM International Conference on Information and Knowledge Management, 1997.
[99] A.B. Poritz, Hidden Markov models: a guided tour, in: Proceeding of IEEE International Conference on Acoustics, Speech and Signal, 1988, pp. 7–13.
[100] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[101] Y. Ariki, Y. Sugiyama, A TV news retrieval system with interactive query function, in: Proceeding of Second IFCIS International Conference on Cooperative Information Systems, 1997, pp. 184–192.
[102] A. Psarrou, S. Gong, M. Walter, Recognition of human gestures and behaviour based on motion trajectories, Image Vision Comput. 20 (2002) 349–358.
[103] T. Xiang, S. Gong, Activity based surveillance video content modelling, Pattern Recognition 41 (7) (2008) 2309–2326.
[104] Y. Yacoob, M.J. Black, Parameterized modelling and recognition of activities, in: Proceeding IEEE International Conference on Computer Vision, 1998, pp. 120–127.
[105] G. Xu, Y.-F. Ma, H.-J. Zhang, S. Yang, Motion based event recognition using HMM, in: Proceeding of 16th International Conference on Pattern Recognition (ICR'02), vol. 2(1), 2002, pp. 831–834.
[106] M. Petkovic, W. Jonker, Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events, in: IEEE International Workshop on Detection and Recognition of Events in Video, 2001.
[107] R. Fablet, P. Bouthemy, Statistical motion-based object indexing using optic flow field, Proc. Int. Conf. Pattern Recognition 4 (2000) 40287–40290.
[108] E.D. Petajan, B. Bischoff, D. Bodoff, N.M. Brooke, An improved automatic lipreading system to enhance speech recognition, in: Proceeding of Human Factors in Computing Systems (SIGCHI'88), 1988, pp. 19–25.
[109] K.E. Finn, A.A. Montgomery, Automatic optically-based recognition of speech, Pattern Recognition Lett. 8 (1988) 159–164.
[110] E. Ardizzone, M.L. Cascia, Automatic video database indexing and retrieve, Multimedia Tools Appl. (MTAP) 4 (1) (1997) 29–56.
[111] S.-F. Chang, et al., A fully automated content-based video search engine supporting spatio-temopral queries, IEEE Trans. Circuits Systems Video Technol. 8 (5) (1998) 602–615.
[112] M. Ioka, M. Kurokawa, A method for retrieving sequences of images on the basis of motion analysis, in: Proceedings of the SPIE Image Storage and Retrieval Systems, vol. 1662, 1992, pp. 35–46.
[113] S. Dagtas, W. Al-Khatib, A. Ghafoor, A. Khokhar, Trail-based approach for video data indexing and retrieval, in: Proceeding of IEEE International Conference on Multimedia Computing and System, vol. 2, 1999, pp. 235–239.
[114] K. Eickhorst, P. Agouris, A. Stefanidis, Modelling and comparing spatio-temporal events, in: Proceedings of the Fifth (USA) National Conference on Digital Government Research (dg.o), 2004, pp. 163–172.
[115] R.C. Nelson, R. Polana, Qualitative recognition of motion using temporal texture, CVGIP: Image Understand. 56 (1) (1992) 78–89.
[116] N. Dimitrova, F. Golshani, Motion recovery for video content analysis, ACM Trans. Inf. Systems 13 (4) (1995) 408–439.
[117] N.H. Goddard, The perception of articulated motion: recognizing moving light displays, Ph.D. Thesis, University of Rochester, 1992.
[118] G.A. Martin, M. Shah, Lipreading using optical flow, in: Proceeding of National Conference Undergraduate Research, 1992.
[119] S.A. Niyogi, E.H. Adelson, Analyzing and recognizing walking figures in xyt, in: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, June 1994.
[120] S. Niyogi, E. Adelson, Analyzing gait with spatiotemporal surfaces, in: IEEE Workshop on Motion of Non-rigid and Articulated Objects, IEEE Press, Silver Spring, MD, 1994, pp. 64–69.
[121] R. Polana, R.C. Nelson, Detection and recognition of periodic, nonrigid motion, Int. J. Comput. Vision 23 (3) (1997) 261–282.
[122] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, Proc. Int. Conf. Comput. Vision 2 (2003) 726–733.
[123] A. Briassouli, N. Ahuja, Extraction and analysis of multiple periodic motions in video sequences, IEEE Trans. Pattern Anal. Mach. Intell. 29 (7) (2007) 1244 –1251.
[124] K. Rangarajan, W. Allen, M.A. Shah, Recognition using motion and shape, in: Proceeding of 11th International Conference on Pattern Recognition, Hague, Netherlands, 1992.
[125] E. Sahouria, A. Zakhor, Motion indexing of video, Proc. IEEE Int. Conf. Image Process. 2 (1997) 526–529.
[126] W. Chen, S.-F. Chang, Motion trajectory matching of video objects, in: Proceedings of SPIE/IS&T Storage and Retrieval for Media Data-based 2000, SPIE, vol. 3972, 2000, pp. 544–553.
[127] Y. Wexler, E. Shechtman, M. Irani, Space–time video completion, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 463–476.
[128] M. Sonka, V. Hlavac, R. Boyle, Image Processing Analysis and Machine Vision, second ed., PWS Publishing, 1999.
[129] S.Y. Lee, H.M. Kao, Video indexing-an approach based on moving object and track, SPIE Storage Retrieval Image Video Databases 1908 (1993) 25–36.
[130] A. Yoshitaka, Y. Hosoda, M. Yoshimitsu, M. Hirakawa, T. Ichikawa, VIOLONE: video retrieval by motion example, J. Visual Lang. Comput. 7 (4) (1996) 423 –443.
[131] J.Z. Li, M.T. Ozsu, D. Szafron, Modeling of moving objects in a video database, in: Proceedings of IEEE International Conference on Multimedia Computing and Systems, 1997, pp. 336–343.
[132] S. Panchanathan, Y.C. Park, F. Golshani, VideoRoadMap: a system for interactive classification and indexing of still and motion pictures, in: Proceeding of IEEE Instrumentation and Measurement Technology Conference, vol. 1, 1998, pp. 18–21.
[133] P.-S. Tsai, M. Shah, K. Keiter, T. Kasparis, Cyclic motion detection for motion based recognition, Pattern Recognition 27 (12) (1994) 1591–1603.
[134] A.B. Albu, R. Bergevin, S. Quirion, Generic temporal segmentation of cyclic human motion, Pattern Recognition 41 (1) (2008) 6–21.
[135] I. Laptev, On space–time interest points, Int. J. Comput. Vision 64 (2/3) (2005) 107–123.
[136] A. Yilmaz, M. Shah, Action sketch: a novel action representation, Proc. Conf. Comput. Vision Pattern Recognition 1 (2005) 984–989.
[137] C. Rao, A. Gritai, M. Shah, T. Syeda-Mahmood, View-invariant alignment and matching of video sequences, in: Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV2003), vol. 2, 2003, pp. 939–945.
[138] E. Shechtman, M. Irani, Space–time behavior-based correlation—or—how to tell if two underlying motion fields are similar without computing them?, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 2045–2056.
[139] D. Weinland, R. Ronfard, E. Boyer, Motion history volumes for free viewpoint action recognition, in: IEEE International Workshop on Modeling People and Human Interaction, 2005.
[140] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (29) (2007).
[141] L. Zelnik-Manor, M. Irani, Event-based analysis of video, in: Proceedings of IEEE Conference Computer Vision and Pattern Recognition (CVPR2001), 2001, pp. 123–130.
[142] T. Goodhart, P. Yan, M. Shah, Action recognition using spatio-temporal regularity based features, in: Processing IEEE International Conference on Acoustics, Speech and Signal (ICASSP 2008), 2008, pp. 745–748.
[143] A. Baumberg, D. Hogg, Generating spatiotemporal models from examples, Image Vision Comput. 14 (1996) 515–532.
[144] J.W. Hsieh, Y.T. Hsu, H.Y.M. Liao, C.C. Chen, Video-based human movement analysis and its application to surveillance systems, IEEE Trans. Multimedia 10 (3) (2008) 372–384.
[145] J. Tangelder, R. Veltkamp, A survey of content based 3D shape retrieval methods, in: Proceedings of International Conference on Shape Modeling Applications, 2004, pp. 145–156.
[146] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, Int. J. Comput. Vision 9 (2) (1992) 137 –154.
[147] R. Vidal, R. Hartley, Three-view multibody structure from motion, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 214–227.
[148] C. Bregler, A. Hertzmann, H. Biermann, Recovering non-rigid 3d shape from image streams, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 690–696.
[149] L. Torresani, A. Hertzmann, C. Bregler, Learning non-rigid 3D shape from 2D motion, in: Advances in Neural Information Processing Systems (NIPS), 2003, pp. 1555–1562.
[150] G. Wang, Q.M.J. Wu, Stratification approach for 3-D Euclidean reconstruction of nonrigid objects from uncalibrated image sequences, IEEE Trans. Systems Man Cybernet. Part B 38 (1) (2008) 90–101.
[151] A. Adam, et al., Robust real-time unusual event detection using multiple fixed-location monitors, IEEE Trans. Pattern Anal. Mach. Intell. 30 (3) (2008).
[152] E. Stoykova, et al., 3-D Time-varying scene capture technologies—a survey, IEEE Trans. Circuits systems video Technol. 17 (11) (2007) 1568–1586.
[153] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, Cambridge, UK, 2003.
[154] A. Yilmaz, M. Shah, Matching actions in presence of camera motion, Comput. Vision Image Understand. 104 (2) (2006) 221–231.
[155] Q.T. Luong, R. Deriche, O. Faugeras, T. Papadopoulo, On determining the fundamental matrix: analysis of different methods and experimental results, INRIA, Technical Report, Raport de Recherche 1894, 1993.
[156] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, Int. J. Comput. Vision 50 (2) (2002) 203–226.
[157] S.M. Seitz, C.R. Dyer, View-invariant analysis of cyclic motion, Int. J. Comput. Vision 25 (1997) 1–25.
[158] V. Parameswaran, R. Chellappa, Quasi-invariants for human action representation and recognition, in: Proceeding of ICPR2002, vol. 1, 2002, pp. 307–310.
[159] A. Gupta, A. Mittal, L.S. Davis, Constraint integration for efficient multiview pose estimation with self-occlusions, IEEE Trans. Pattern Anal. Mach. Intell. 30 (3) (2008) 406–493.
[160] R. Polana, R.C. Nelson, Recognition of motion from temporal texture, in: Proceeding of Conference on Computer Vision and Pattern Recognition, Champaign, IL, 1992, pp. 129–134.
[161] N. Dimitrova, F. Golshani, Rx for semantic video database retrieval, in: Proceeding of the Second ACM International Conference on Multimedia, San Francisco, CA, 1994, pp. 219–226.
[162] A. Hampapur, A. Gupta, B. Horowitz, C.-F. Shu, C. Fuller, J.R. Bach, M. Gorkani, R. Jain, Virage video engine, in: Proceedings of the SPIE Storage and Retrieval for Image and Video Databases V, vol. 3022, 1997, pp. 188–198.
[163] D.A. Adjeroh, M.C. Lee, I. King, A distance measure for video sequences, Comput. Vision Image Understand. 75 (1/2) (1999) 25–45.
[164] N. Yazdani, Z.M. Ozsoyoglu, Sequence matching of images, in: Proceedings of Eighth International Conference on Scientific and Statistical Database Management, 1996, pp. 53–62.

[165] Y. Rui, P. Anadan, Segmenting visual actions based on spatio-temporal motion patterns, Proc. IEEE Conf. Computer Vision Pattern Recognition 1 (2000) 111 –118.
[166] M. Ramesh Naphade, I.V. Kozintsev, T.S. Huang, Factor graph framework for semantic video indexing, IEEE Trans. Circuits Systems Video Technol. 12 (1) (2002) 40–52.
[167] P. Smith, N. Lobo, M. Shah, Temporalboost for event recognition, in: Proceeding IEEE International Conference Computer Vision (ICCV2005), vol. 1, 2005, pp. 733–740.
[168] H.T. Nguyen, et al., Spatio-Temporal context for robust multitarget tracking, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 52–64.
[169] R. Fablet, P. Bouthemy, Motion recognition using spatio-temporal random walks in sequence of 2D motion-related measurements, in: Proceeding of International Conference on Image Processing, 2001, pp. 652–655.
[170] F. Fleuret, J. Berclaz, P. Fua, Multicamera people tracking with a probabilistic occupancy map, IEEE Trans. Pattern Anal. Mach. Intell. 20 (2) (2008) 267–282.
[171] D. Papadias, Y. Theodoridis, T. Sellis, M. Egenhofer, Topological relations in the world of minimum bounding rectangles: a study with R-trees, in: Proceedings of the ACM Conference on the Management of Data (SIGMOD), ACM Press, San Jose, CA, 1995.

About the Author—SAMEER SINGH is Professor of Autonomous Systems in the Department of Computer Science, and is the Director of Research School of Informatics, Loughborough University, UK. He also heads Computer Vision and Autonomous Systems research group at Loughborough with more than 55 members. His main research focus is on the development of novel sensor data analysis and machine learning techniques that can support semi- and fully automated intelligent systems for transportation, security and surveillance, mobile phone networks, and biomedical applications. These diverse applications are complex in nature, depend heavily on advances in machine learning and sensor technology for solving problems, and can benefit enormously from automation. Over the last two decades, Prof. Singh has worked at the interface between computer science, engineering, health sciences and mathematics to develop novel algorithms in the areas of computer vision (quantitative evaluation of image enhancement, evolutionary approaches to object tracking, novelty detection in video sequences, optimisation of image analysis tools, classifying human dynamics, audio–video fusion, and handwriting recognition), and machine learning (multi-resolution pattern recognition, pareto-evolutionary neural networks, sensor fusion, predictive systems, and multi-objective optimisation). Most of this research has been published in various IEEE Transactions and other leading journals. Altogether, Prof. Singh has published over 170 papers in his career, and currently has more than £2 million research grant income to support his research team. His work is supported by several leading companies, for example HP Labs, Motorola, Corus Rail, QinetiQ, and government agencies working on transport and national security. He is also highly active in serving on various conference committees, and journals. Notably, he is currently serving as Editor-in-Chief of Pattern Analysis and Applications journal by Springer, and is Associate Editor of IEEE Transactions on SMC B, IEEE Transactions on Knowledge and Data Engineering, Real Time Image Analysis, and Neural Computing and Applications journal.

About the Author—MANEESHA SINGH was born in India. She received the B.S. degree in computer science from Kurukshetra University, Kurukshetra, India and the M.Phil. and Ph.D. degrees from the University of Exeter, Exeter, UK in 1999, 2001 and 2004, respectively. Her Ph.D. was in the area of machine learning for image analysis in aviation security. Her main research interests include image processing, natural scene analysis, video analysis, and neural networks. She has published more than 30 papers in the area of machine learning for image analysis in peer reviewed journals and conferences. Currently she is a Senior Research Fellow at Loughborough University leading the project on imaging for road transport applications.

About the Author—WEI REN graduated from University of Exeter with a Ph.D. in 2005 in the area of spatiotemporal analysis for video retrieval. Her key research interests are in the areas of image analysis, neural networks and machine learning. During the course of her Ph.D. she developed novel framework for video retrieval and a publicly available benchmark Minerva for video retrieval. She is currently working as a Post-Doctoral research in Peking University.