Event-Oriented Video Semantic Representation Methods: Postprint

Authors: Li Xuhui, Wu Qingfeng

Date: 2023-04-01T16:15:54+00:00

Abstract

[Purpose/Significance] Video content is profoundly influencing the information lives of a large population in our country, and effective representation of video semantics constitutes a critical foundation for advancing current research on video content and video application services. Existing video semantic representation methods suffer from limitations including a unitary perspective on event semantic representation and granularity partitioning, as well as a lack of flexible mechanisms for object semantic evolution. Therefore, investigating more effective video semantic representation methods holds significant importance.

[Method/Process] This paper proposes an event-oriented video semantic representation method. This method considers the human bidirectional cognitive process, can interpret and generate event semantics from different perspectives based on diverse user backgrounds and needs, and defines corresponding mechanisms for changes in semantic objects and roles.

[Results/Conclusion] The event-oriented video semantic representation method possesses a complete semantic representation framework, supports multi-perspective event semantic representation, enables flexible semantic expansion at the attribute level, object level, and event level, and can represent richer video semantics.

Full Text

Preamble

Video-based content such as short videos, online courses, mobile live streaming, and video blogs is profoundly influencing the information lives of a large population in China [1]. Compared to the television and PC eras, the demand for video content consumption and related research in the mobile internet era is becoming increasingly refined. Video content research initially focused only on external annotation information, then shifted to low-level features such as

color and motion trajectories, while the current emphasis is on research and applications based on video semantic features. Video data mining research requires a robust video semantic model as its foundational architecture, and the efficient organization and value extraction of library video resources depend on an appropriate video semantic representation framework [2-3]. Novel retrieval and recommendation systems must consider video semantics to fundamentally improve video content distribution efficiency. A sound video semantic representation method is fundamental to the aforementioned research and applications. As video analysis technology advances and user demands become more refined, video semantic representation research must not only effectively incorporate semantic objects and represent event semantics but also address the design of video event semantic structures, the extensibility of event semantics, and the corresponding mechanisms for semantic object changes.

Existing video semantic representation methods in current research, while possessing certain semantic representation capabilities, suffer from problems such as a single perspective and granularity division approach for event semantic representation and a lack of flexible mechanisms for semantic object changes. Based on the current state of research, this paper proposes an event-oriented video semantic representation method that follows a bottom-up semantic description process and fully considers the bidirectional nature of human cognitive processes [4], aiming to provide a video semantic representation method that supports multi-perspective representation of event semantics and corresponding multiple granularity division approaches. Centered on this research objective, this paper will first summarize the current state and shortcomings of video semantic representation methods, then analyze several key issues that urgently need resolution in current video semantic representation work, proceed to specifically define the event-oriented video semantic representation method and discuss how our method addresses these key issues, illustrate the method's effectiveness with a basketball game video clip example and compare it with existing related methods to demonstrate the innovation of our approach, and finally summarize the paper and propose future research directions.

2 Related Research

Early video semantic representation employed annotation-based methods, whose main idea was to overlay annotation information composed of natural text or structured data onto corresponding video sequences in the video stream. This method could represent a limited number of semantic objects, and annotations could not be interconnected, making it difficult to characterize complex semantic relationships in videos. It was primarily used to meet simple keyword- and attribute-based video query requirements. As video resources became richer and related research developed, researchers' and users' demands for video semantic representation became more complex. In the field of video data mining, research on video concept detection [5-6], video classification [7], content structure analysis [8], topic mining [9], and event mining [10-11] all

require more effective semantic representation methods as their preliminary modeling foundation and to provide interpretability for research results. Under growing new demands, on the one hand, current video semantic representation research is mostly based on video semantic data models, which are hierarchical models with the bottom layer corresponding to raw video data streams and the top layer to video semantic information. The top-layer semantic information is obtained through semantic abstraction and mapping of raw video data. Depending on the mapping mechanism, the model may have different types and numbers of intermediate layers, such as semantic object layers and event scene layers. On the other hand, since the concept of domain events [12] was proposed, event-based semantic representation has become a consensus among researchers in video semantic representation. An event is a relatively complete and comprehensive semantic information unit formed by the features, relationships, and background information of one or more semantic objects. Although subsequent video semantic representation research has different emphases, it basically reflects the idea of hierarchical models with events as the core semantics, and our method is also based on this foundation.

Although there is this common conceptual foundation, constrained by video analysis technology levels or the limitations of researchers' fields, many related studies have relatively limited representation of high-level semantics. In the graph model-based soccer video semantic representation method proposed by Wang Haoran et al. [13], event units are composed of shot and audio features. Zhang Jing et al. defined event templates based on pedestrian motion features, mapping motion trajectories to event semantics [14]. Liu Xiaolu proposed a knowledge element model for security video [15], mapping security video content to three aspects: basic video information, carrier objects, and security events. Xie Xiao et al. defined a multi-level structure for geographic video semantics [16], abstracting geographic video semantics into three interrelated layers: feature domain, behavior process domain, and event domain. The above research mostly focuses on security, geographic, and transportation domains, fully considering low-level features and spatiotemporal information of video content and combining professional domain knowledge, but lacks support for high-level semantics, has relatively primitive event semantic representation, and lacks universality. Therefore, our method will focus on representing more complex event semantics in videos and strive to achieve universality.

Researchers have attempted to propose various video semantic representation methods, aiming to cover relatively complete low-level feature information of videos while accurately expressing high-level semantic information. The AVIS model proposed by S. Adali et al. [17] was an earlier video semantic model that introduced high-level semantic information. AVIS clearly defined semantic concepts such as video objects, events, and roles, and formed tree structures with sequence nodes based on temporal inclusion relationships. However, this model ignored most low-level features of semantic objects, and users could not extend and complete object semantics in a top-down manner. Although the VIDEX method [18] integrated low-level video features, its high-level semantic

structure design was relatively simple and could not represent complex associations between semantic objects and events. Bao Hong et al. proposed a hierarchical semantic association model [19], which used concept hierarchy trees to represent inheritance relationships between abstract concepts and could effectively represent relatively complex abstract concepts, but did not consider the hierarchical nature of event semantic structures. The THVDM proposed by Y. Wang [20] distinguished objects and events at the concept level, predefined some event semantic structures, and could represent associations between events of different granularities. However, existing representation methods all suffer from problems of a single perspective and granularity division approach for event semantic representation and a lack of flexible mechanisms for semantic object changes. The specific manifestations are: Single perspective and granularity division approach for event semantic representation. Existing representation methods have a unique interpretation perspective for events, while different user groups do not have unified understanding of video semantics. Taking basketball game videos as an example, a coach might interpret from a global tactical perspective, while ordinary fans might interpret from the perspective of a single player's performance. The granularity division of event semantics is also related to the interpretation perspective. A coach's event granularity division for an entire game video might be based on the overall tactical confrontation process, while the event granularity from an ordinary fan's perspective might be based on specific behaviors such as scoring or fouling by a single player. Lack of flexible mechanisms for semantic object changes. In existing methods, semantic objects and roles participating in events are usually static and cannot well support semantic changes of the same semantic object under different interpretation perspectives. For semantic objects that may participate in events of different granularities, the number and meaning of their instantiated roles also require specific intermediate mechanisms, which existing methods fail to provide good support for. Therefore, this paper proposes an event-oriented video semantic representation method that focuses on multi-perspective representation of event semantics and corresponding multiple granularity division approaches, and provides flexible mechanisms for semantic object changes.

3 Key Issues in Video Semantic Representation

3.1 Bottom-Up Description Process

Human understanding of video semantics is a bottom-up description process from low-level physical features to high-level semantic information. A video semantic representation method that conforms to the actual cognitive process must effectively and flexibly support this bottom-up description process. The representation of semantic objects is the foundation of the bottom-up description process. When representing low-level semantic objects, it is necessary to focus on characterizing features independent of specific events to ensure that the same semantic object can be repeatedly invoked in different events.

When representing high-level event semantics in the bottom-up description pro-

cess, the same semantic object will have different semantic information when participating in different events. For example, in a basketball game video, the semantic object "a certain athlete" may participate in the "ball possession offense" event and the "under-basket defense" event with the semantics of "offender" and "defender" respectively. Therefore, intermediate roles that specifically represent event-related features of semantic objects can be introduced. In the bottom-up description process, the main cognitive work is to judge and determine the existence and types of semantic objects and events. Once the semantic description framework is determined, it is usually necessary to fill in more semantic information for semantic objects or events in the framework according to different needs. This information can come from low-level features, background information, etc. For example, in a basketball game video segment, after bottom-up determination of the "scoring" event based on spatiotemporal associations of athletes and assigning the role of "scorer" to the semantic object "a certain athlete" in this event, in addition to the event-independent attributes of "a certain athlete," it may also be necessary to complete more semantic information for the "scorer" role, such as whether the "scoring method" is "jump shot," "layup," or "free throw."

3.2 Multi-Perspective Interpretation of Video Events

Event semantic representation is the core of video semantic representation. Existing video semantic representation methods lack attention to multi-perspective interpretation of events and ignore the bidirectional nature of human cognitive behavior. Specifically, in the bottom-up description process, people may choose to focus on different semantic objects or different aspects of the same semantic object, and therefore may interpret different event semantics and corresponding event structures. For example, when watching the same basketball game video segment, some users focus on score information, some focus on the behavior of a particular star player, some focus on the referee's behavior, and some focus on team tactical coordination. These different focuses may all serve as bases for interpreting video semantics and dividing event structures (see Figure 1 [Figure 1: see original paper]). Therefore, supporting multi-perspective interpretation of events in video semantic representation is very important. The semantic representation method must be able to represent semantic information from various perspectives and support different event structure division approaches under different perspectives.

3.3 Granularity Division of Video Events

Granularity refers to the size of semantic fragments when video events are represented. The selection of event interpretation perspective mentioned in the previous key issue affects the division of event structures, which is mainly manifested as granularity division of events. Granularity division of events involves both the combination of events and semantic changes of related event roles. The combination of events means that multiple continuous events with smaller

granularity, lower level, and single semantics can be combined as sub-events into composite events with larger granularity, higher level, and richer semantics. For example, continuous "chasing," "subduing," and "escorting" events recorded in a surveillance video can be combined into a larger-granularity "arrest" event. In the process of event combination, it should be noted that the semantic objects involved in sub-events will undergo changes in quantity and semantics in composite events, and the semantic representation method must be able to support such changes. For instance, the "chased person" role in the sub-event "chasing" could evolve into the "criminal" role in the composite event "arrest."

3.4 Top-Down Semantic Completion

Semantic completion refers to the process of filling more semantic information into an already formed semantic representation framework. As mentioned earlier, different users may have different understandings of the same video semantics. Therefore, a good video semantic representation method must have flexible extensibility and be able to generate semantics extensibly based on users' different focuses. The semantics of objects and roles involved in events must also be able to expand with changes in event semantics.

Support for retrieval requirements is an important capability for video semantic representation methods to function in application scenarios. Under the current background of rich video content and complex user demands, existing retrieval methods based on keywords or low-level features cannot fully meet user needs, while video semantics-based retrieval methods are key to current research and development in the video retrieval field. Therefore, it is necessary to consider support for diverse retrieval in video semantic representation methods.

4 Video Semantic Representation Method

To address the key issues in video semantic representation mentioned above, this paper proposes an event-oriented video semantic representation method. This section will first define the semantic representation method framework, discuss how the method solves the aforementioned key issues, then use a basketball game video clip for application example description, and finally compare our method with existing related methods to illustrate the innovation of our approach.

4.1 Event-Oriented Video Semantic Representation Method

This paper proposes an event-oriented video semantic representation method. The logical framework of this method's semantic representation is shown in Figure 2 [Figure 2: see original paper]. The semantic representation framework characterizes three main entities and their association methods: semantic objects, roles, and events. Semantic objects are the foundation of semantic representation in this method. Roles are obtained by instantiating semantic objects in specific events and participate in the semantic construction of specific events. The specific definitions of the three entities are as follows:

(1) **Semantic Object (Object)**: A semantic object is an object with primary semantics obtained through automatic analysis of low-level video features. All semantic objects exist independently of specific events. A semantic object is represented using a tuple as Object = {oid, OT, Attrs}, where:

- oid is the unique identifier of the semantic object.
- OT = {ti, tj} records the start and end times of the time interval when the semantic object appears in the video, where ti represents the start time and tj represents the end time.
- Attrs = {k1:v1, ..., kn:vn} is an extensible attribute key-value pair. It includes the object's low-level features such as color and motion trajectory, as well as other event-independent information such as object naming. Other event-independent information about the object can be extended or completed as needed.

(2) **Role (Role)**: A role is obtained by instantiating a semantic object in a specific event. The semantic information of a role is related to a specific event. A role is represented using a tuple as Role = {rid, semrole, RT, Attrs}, where:

- rid is the unique identifier of the role.
- semrole is the semantic label of the role, representing the type of semantic role the role plays in the event, such as "police" and "criminal" in an "arrest" event.
- RT = {ti, tj} records the start and end times of the time interval of this role in the video.
- Attrs = {k1:v1, ..., kn:vn} is an extensible attribute key-value pair. The attributes are role information obtained by combining relevant semantic object features with the event semantics they participate in, and can be extended or completed with other event-related information as needed.

(3) **Event (Event)**: An event is a high-level semantic block formed by integrating one or more meaningful roles and the semantic relationships between them. The generation of video event semantics is directly related to users' focus angles on video content. An event is represented using a tuple as Event = {eid, name, focus, ET, Attrs}, where:

- eid is the unique identifier of the event.
- name is the event name.
- focus is the focus perspective based on which the event semantics is generated.
- ET = {ti, tj} records the start and end times of the time interval of the event in the video. The time interval of a non-composite event is obtained by taking the union of the time intervals of the event's roles, while the time interval of a composite event is obtained by taking the union of the time intervals of its sub-events.
- Attrs = {k1:v1, ..., kn:vn} is an extensible attribute key-value pair. It can include the event's semantic time information, semantic location informa-

tion, etc., and can be extended or completed as needed.

The three types of entities defined above are interconnected and together constitute a complete semantic representation framework to represent rich video semantics. The specific definitions of the association methods between entities are as follows:

(1) **Semantic Object-to-Object Association (Object-Object Relation)**: Our method focuses on characterizing event-independent spatiotemporal associations (Spatio-Temporal Relation) between semantic objects. Temporal association mainly refers to the relative positions of two semantic objects within time intervals, such as appearing in the same time interval (equal), appearing in a previous time interval (before), appearing in a later time interval (after), etc. Spatial association includes directional association and topological association. Directional association includes east, south, above, below, etc. Topological association includes cover, touch, etc. Detailed definitions of spatiotemporal associations can be found in [21-23]. Spatiotemporal associations between semantic objects are directional and are represented as a series of directed edges in the graphical framework. A spatiotemporal association between semantic objects is represented using a tuple as STRel = {type, oid1, oid2}, where:

- type is the type of spatiotemporal association, such as "cover."
- oid1 is the identifier of the starting semantic object of the directed relation.
- oid2 is the identifier of the ending semantic object of the directed relation.

(2) **Semantic Object-Role Association (Object-Role Relation)**: Event-independent low-level semantic objects are instantiated into roles in specific events. The association between semantic objects and roles is called an instantiation relation (Instantiation Relation). In this association, one-to-one, one-to-many, and many-to-one quantity relationships are allowed between semantic objects and semantic roles. An instantiation relation is represented using a tuple as InsRel = {Os, Rs, eid}, where:

- Os = {oid1, …, oidn} is the set of identifiers of semantic objects participating in the instantiation.
- Rs = {rid1, …, ridn} is the set of identifiers of roles participating in the instantiation.
- eid is the identifier of the event targeted by the instantiation process.

(3) **Role-to-Role Association (Role-Role Relation)**: The semantic information of roles is related to specific events, while the underlying foundation of roles is semantic objects. Therefore, associations between roles are semantic associations (Semantic Relation) formed based on spatiotemporal associations between objects combined with specific event semantics. Semantic associations between roles are also directional. A semantic association between roles is represented using a tuple as SemRel = {type, rid1, rid2}, where:

- type is the type of semantic association, which is related to specific event semantics.
- rid1 is the identifier of the starting role of the directed relation.
- rid2 is the identifier of the ending role of the directed relation.

(4) **Event-to-Event Association (Event-Event Relation)**: Our method focuses on composition associations (Composition Relation) between events. An event-to-event composition association is represented using a tuple as ComRel = {Eid, Subs}, where:

- Eid is the identifier of the composite event.
- Subs = {eid1, …, eidn} is the set of identifiers of sub-events.

(5) **Event-Role Association (Event-Role Relation)**: Events own roles. The association between events and roles (Owning Relation) is represented using a tuple as OwReal = {eid, Roles}, where:

- eid is the identifier of the event.
- Roles = {rid1, …, ridn} is the set of identifiers of roles owned by the event.

The above definitions constitute the framework of the video semantic representation method proposed in this paper. Our method focuses on supporting multi-perspective representation of event semantics and corresponding multiple granularity division approaches. In the above definitions, the focus attribute of events provides the foundation for multi-perspective expansion of event semantics. Multiple event granularity division approaches are achieved through time interval division and composition associations between events under each different perspective. To clarify this implementation process, we define an extensible object T = {T1, T2, …, Tn}, where T1 = {focus1, ETs} represents the event granularity division approach when the focus perspective is focus1, and ETs = {ET1, ET2, …, ETn} represents the set of time intervals of composite and non-composite events under the current division approach. Existing methods only support a single division approach under a single perspective, while in our method, the extensibility of object T represents the diversity of event semantic perspectives and event granularity division approaches.

As can be seen from Figure 3 [Figure 3: see original paper], taking a certain video segment as an example, its multi-perspective event semantic representation and multiple event granularity division process can be formalized as object T = {T1, T2}, where T1 = {focus1, {ET1, ET2}} and T2 = {focus2, {ET1, ET2, ET3}}, representing two granularity division approaches under two perspectives. In the figure, when the perspective is focus1, ET1 = {et1, et2, et3}, representing the composition association between event Eid1 and its sub-events, which demonstrates the correspondence between event composition association and time interval division and shows the feasibility of implementing this process.

In our video semantic representation method, video semantic representation is based on low-level semantic objects and their spatiotemporal associations. Semantic objects are further instantiated into roles in specific events, enabling

them to be reused in the semantic representation of multiple specific events. Event semantics in videos can be generated based on different focus perspectives, forming multiple granularity divisions across multiple perspectives. The semantics of events can be flexibly extended and changed, and the roles involved in events and semantic associations between roles also change accordingly, while the low-level semantic objects in the framework remain unchanged, only producing different instantiation processes based on different event semantics. Instantiated roles serve as the bridge between changing high-level event semantics and unchanged low-level semantic objects. This is the semantic representation process of the event-oriented video semantic representation method.

The following sections will introduce some relevant details of our method and discuss how it addresses the key issues in video semantic representation mentioned in the previous section.

4.2 Resolution of Key Issues

To support the bottom-up description process, the semantic objects defined in our method are objects with primary semantics that can be obtained using current recognition technologies, and their attributes are all event-independent. In terms of associations between objects, we focus on characterizing event-independent spatiotemporal associations. Low-level semantic objects are instantiated upward into event-related roles, and roles form high-level semantic associations in events based on the spatiotemporal associations of underlying objects. Events then form higher-level event semantics through aggregation of semantic granularity. Here we illustrate the details of how spatiotemporal associations between semantic objects form semantic associations between roles. As shown in Figure 4 [Figure 4: see original paper], the spatiotemporal association of "spatiotemporal position approaching" between two low-level semantic objects evolves into the semantic association of "chasing" between the roles of "police" and "criminal" in the specific "arrest" event.

In supporting multi-perspective interpretation of video events, unlike previous research that only supports a priori, single-perspective event semantic representation, our method allows extensible, multi-perspective event semantic representation. Events can be distinguished by the "focus" attribute defined above. Event semantics can be interpreted from different perspectives, and each perspective can produce different event granularity division approaches. Previous research methods could generally only characterize tree-structured event semantics aggregated under a single perspective, while our method enables the final event semantics to form a network-structured event semantics aggregated from multiple perspectives.

In supporting event granularity division, our method designs mechanisms for changes of semantic objects and roles in events at different levels. The design of semantic objects and roles changing with events lies in two aspects. First, semantic objects are sliced in terms of quantity. Composite events only focus on

semantic objects in sub-events that better reflect high-level semantics. Therefore, the set of objects included in a composite event is a subset of the sets of objects included in all its sub-events. For example, in the composite event "Score" composed of sub-events "Pass" and "Shoot," all sub-events involve 4 semantic objects, while the composite event only involves 2 of them (see Table 1). Second, the same semantic object uses different instantiated roles in events at different levels. For example, the semantic object "Player1" is instantiated as the "Passer" role in the sub-event "Pass" and as the "Assistant" role in the composite event "Score" (see Figure 5 [Figure 5: see original paper]).

Table 1 Semantic Object Changes in Event Composition Process

Event	Semantic Objects Involved
Pass, Shoot	Player1, Player2, Player3, Ball
Score	Player1, Player2

In supporting top-down semantic completion, we define extensible attribute key-value pairs for semantic objects, roles, and events, which can add personalized semantic information as needed to meet the requirements of top-down semantic completion. Moreover, due to the existence of roles as an intermediate layer between semantic objects and events, when completing event-related semantics for roles in specific events, the basic semantics of the objects themselves will not be changed.

In supporting extensibility and retrieval requirements, first, support for extensibility is reflected in all aspects of the method design. The extension of attribute key-value pairs, the instantiation process of objects, and the multi-perspective generation approach of events support attribute-level, object-level, and event-level semantic extensions respectively. In terms of retrieval capability, our representation method can not only support graph data retrieval methods rich in semantics but also conveniently perform retrieval based on semantic objects or based on event focus perspectives. It can retrieve all roles generated by object instantiation through instantiation relations.

4.3 Method Application Example

To demonstrate the application effect of the event-oriented video semantic representation method, this paper selected a clip from a basketball game video for semantic representation.

The video clip used in this paper is from a game between Beijing Royal Fighters (hereinafter "Team A") and Sichuan Blue Whales (hereinafter "Team B") in the CBA. The continuous key frames of the video clip are shown in Figures 6 [Figure 6: see original paper] and 7 [Figure 7: see original paper]. The basic scene information is: when the game is about to end, player "Athlete A3"

from Team A passes the ball to “Athlete A1,” and “Athlete A1” makes a shot, reversing the score between the two teams before the game ends.

Using our method to represent the semantics of this video segment, as shown in Figure 8 [Figure 8: see original paper], the figure shows the event semantics from three perspectives of this video segment, with focuses on “ball holder,” “Athlete A2,” and “score” respectively. For convenient display, the figure simplifies the representation of entity attributes by directly showing the names of semantic objects, semantic labels of roles, event names, and focus attributes in corresponding rectangular boxes, while other attributes are not fully displayed. The figure does not show all object-to-object and role-to-role associations, mainly demonstrating the evolution from spatiotemporal associations between objects to semantic associations between roles through the associations of objects “Athlete A1” and “Athlete B1” and their corresponding roles “Shooter” and “Defender.” The spatiotemporal association of “spatiotemporal position approaching” between the two semantic objects evolves into the semantic association of “interception” between roles.

In the semantic representation with “ball holder” as the focus, objects “Athlete A1,” “Athlete A3,” and “Ball” are instantiated as “Receiver,” “Passer,” and “Basketball” respectively, becoming roles participating in the “Pass” event. Since the shot is made, the continuous “Pass event” and “Shoot event” are combined into a larger-granularity “Score” event. In the “Score” event, the semantic objects involved in its sub-events are only “Athlete A2” and “Athlete A1,” which have strong relevance to the semantics at this level. Therefore, the semantic representation of the “Score” event only involves these two objects, which are instantiated as new roles “Assistant” and “Scorer” respectively. The latter also shows in the figure the extensible attribute key “scoring method” with its attribute value “jump shot.”

The semantic representation processes with “Athlete A2” and “score” as focuses are similar to the above process. The figure shows the “Positioning” event with “Athlete A2” as the focus and the “Buzz Beater” event with “score” as the focus. A “Buzz Beater” event in basketball refers to a scoring event that reverses the score and determines the game’s outcome when the game is about to end. The roles “Winning Team” and “Losing Team” participating in the Buzz Beater event shown in the figure are both instantiated from multiple semantic objects in a many-to-one manner, which is an instantiation mechanism supported by our method.

The above three-perspective semantic representation is only used as an example illustration in this paper. In practical applications, the event-oriented video semantic representation method supports extending more event semantic representations from different perspectives.

4.4 Comparison with Related Methods and Innovation Explanation

To better understand the differences among various video semantic representation methods and more intuitively reflect the advantages of our method, this section compares our method with other research and explains the innovation of our method.

This section selected six video semantic representation methods mentioned in the related work to compare with our method. All these methods have basic video event semantic representation capabilities. Our method mainly addresses the problems of single perspective and granularity division approach for event semantic representation and lack of flexible semantic object change mechanisms in existing methods. Therefore, this comparison mainly examines the following four requirements related to these issues: distinguishing objects and roles, allowing event composition, supporting multi-perspective event semantic representation, and having object semantic change mechanisms. The comparison results are shown in Table 2 , where "√" indicates that the method can directly meet the requirement or can indirectly meet it in a similar way, and "×" indicates that the method cannot meet the requirement or does not define related content.

Table 2 Comparison of Typical Video Semantic Representation Methods

Method	Distinguish Objects and Roles	Allow Event Composition	Support Multi-Perspective	Object Semantic Change Mechanism
AVIS [17]	√	√	×	×
VIDEX [18]	×	×	×	×
THVDM [20]	√	√	×	×
Graph Model-Based Soccer Video Semantic Modeling [13]	×	√	×	×

Method	Distinguish Objects and Roles	Allow Event Composition	Support Multi-Perspective	Object Semantic Change Mechanism
Multi-Level Geographic Video Semantic Model [16]	×	√	×	×
Video Hierarchical Semantic Association Model [19]	√	√	×	×
Event-Oriented Video Semantic Representation Method	√	√	√	√

Among the above indicators, the event-oriented video semantic representation method proposed in this paper performs the best. On the one hand, this is because some of the above methods did not have much reference work when they were proposed, and the complexity of event semantics had not yet been

addressed. They mainly clarified basic concepts related to events when exploring video semantic representation methods, providing a foundation for later research. On the other hand, based on previous research, this paper focuses on aspects related to event complexity such as multi-perspective representation of event semantics, making specialized considerations and designs for the distinction between objects and roles involved in complex event semantic representation, event composition and granularity division, and object semantic changes. Therefore, our method can better meet the needs when representing semantics oriented toward events.

Specifically, our method has the following innovations: It has a complete semantic representation framework. Our method follows a bottom-up description process for video semantic representation, covering different levels of semantic information in the representation framework and reasonably associating them. It can represent event semantics from multiple perspectives. Event semantics can be interpreted and generated from different perspectives according to different user backgrounds and needs, producing multiple event granularity division approaches and forming a network-structured event semantic structure aggregated from multiple perspectives. It can flexibly perform semantic extensions. In our method, semantic objects and events have a low coupling relationship. The number of semantic objects participating in events and the semantics of their instantiated roles have corresponding change mechanisms. The semantics of objects and roles can be flexibly extended with changes in event semantics from different perspectives and granularities.

This paper proposes an event-oriented video semantic representation method around the research theme of video semantic representation, illustrates the process of using it for semantic representation through examples, compares it with existing related research, and explains the innovation of our method. Our method solves the problems of single perspective and granularity division approach for event semantic representation and lack of flexible semantic object change mechanisms in existing event semantic representation methods, and provides better support for video semantic completion and extension.

Video is becoming an increasingly important medium in both public domains for brand transmission and ideological shaping [24] and personal domains for knowledge learning [25] and entertainment consumption [26]. In practice, the event-oriented video semantic representation method can provide an information representation framework for the organization and management of video data resources, support the design of systems that meet users' refined video acquisition needs, and provide good intermediate data structures for video data mining research. The event-oriented video semantic representation method can be specifically applied in the following scenarios: Organization and management of electronic library video resources. Video is an important medium for readers to obtain information currently. Reorganizing video resources based on a sound semantic representation method can better realize the value and usability of library video resources [27]. Design of video semantics-based retrieval or

recommendation systems. Applying video semantic representation methods to system design can bring new breakthroughs to refined video content acquisition solutions in current practical applications. Support for video data mining research. Structurally represented video semantic information can support data mining research related to advanced semantics such as video topics and provide interpretability for mining results.

In future research, we will continue work in the following aspects: Improvement of semantic representation work. This paper emphasizes that under multi-perspective event semantic representation, there should be multiple granularity division approaches. Therefore, in terms of inter-event associations, we focused on composition associations most relevant to event granularity division. Subsequent improvements can be made in temporal associations, causal associations, etc., based on multi-perspective event semantic representation. Construction of semantic data models. We will establish a universal data model for the video semantic representation method proposed in this paper and plan to implement it in a graph database-based data schema. System design and implementation. Using purely manual annotation methods cannot maximize the value of the model. Based on the universal data model, we plan to design a video semantic analysis system to achieve automatic or semi-automatic video semantic analysis and structured representation and storage of semantic information.

References

[1] CNNIC Internet Research. The 43rd CNNIC China Internet Report Released [J]. China Broadcasts, 2019(4): 48. [2] Deng Luohua, Deng Dongning, Chen Sheng. On the Construction of Video Libraries [J]. Journal of Academic Libraries, 2010, 28(2): 70-73. [3] Zhao Kun. Research on the Development and Construction of Library Audio-Video Resources in Big Data Environment [J]. Library Development, 2015, 248(2): 64-68. [4] Zhu Zhixian. A Review of Modern Cognitive Psychology [J]. Journal of Beijing Normal University, 1985(1): 1-6. [5] Ca0 Liubin. Research on Deep Learning-Based Image and Video Description Methods [D]. Taiyuan: Shanxi University, 2018. [6] Zhou Jiaosheng. Video Semantic Concept Detection Method Based on Latent Semantic Analysis [J]. Information & Communications, 2018(2): 141-143. [7] Chen Chen. Research on Video Classification Based on Action Semantic Association Rule Mining [D]. Zhenjiang: Jiangsu University, 2018. [8] Vijayakumar V, Nedunchezhian R. Mining video association rules based on weighted temporal concepts [J]. ProQuest, 2012, 9(4): 297-303. [9] Li GR, Zhang WG, Pang JB, et al. Online web video topic detection and tracking with semi-supervised learning [J]. Multimedia Systems, 2016, 22(1): 115-125. [10] Luan Xidao, Xie Yuxiang, Han Zhiguang, et al. Research on News Video Mining Technology [J]. Computer Science, 2007, 34(2): 1-6. [11] Wang Shuo. Research on Basketball Video Exciting Event Detection Methods [D]. Xi'an: Xidian University, 2015. [12] Gupta A, Weymouth T, Jain R. Semantic queries with pictures: the VIMSYS model [C]//Proceedings of the seventeenth international conference on Very

Large Data Bases. San Francisco: Morgan Kaufmann, 1991: 69-79. [13] Wang Haoran, Bai Liang, Lao Songyang. Graph Model-Based Soccer Video Semantic Modeling Method [J]. Computer Science, 2011, 38(6): 266-269, 297. [14] Zhang Jing, Gao Wei, Liu Anan, et al. Video Semantic Event Modeling Method Based on Motion Trajectory [J]. Electronic Measurement Technology, 2013, 36(9): 31-36, 40. [15] Liu Xiaolu. Scene-Based Representation and Retrieval of Security Video Content Based on Knowledge Elements [D]. Dalian: Dalian University of Technology, 2017. [16] Xie Xiao, Zhu Qing, Zhang Yeting, et al. Multi-Level Geographic Video Semantic Model [J]. Acta Geodaetica et Cartographica Sinica, 2015(5): 555-562. [17] Adali S, Candan K, Chen S S, et al. The advanced video information system: data structures and query processing [J]. Multimedia Systems, 1996, 4(4): 172-186. [18] Tusch R, Kosch H, Boszormenyi L. VIDEX: an integrated generic video indexing approach [C]//ACM international conference on multimedia. Los Angeles: ACM, 2000: 448-451. [19] Liu Hongzhe, Bao Hong, Xu De. Content-Based Video Hierarchical Semantic Association Model [J]. Computer Applications, 2005, 25(8): 1797-1800. [20] Wang Y, Xing C X, Zhou L Z. THVDM: a data model for video management in digital library [C]//Proceedings of the sixth international conference of Asian digital libraries. Berlin: Springer International Publishing, 2003: 178-192. [21] Allen J F. Maintaining knowledge about temporal intervals [J]. Readings in qualitative reasoning about physical systems, 1990, 26(11): 361-372. [22] Li J Z, Ozsu M T, Szafron D. Modeling video temporal relationships in an object database management system [C]//Proceedings of the multimedia computing and networking. San Jose: SPIE, 1997: 80-91. [23] Egenhofer M J, Franzosa R D. Point-set topological spatial relations [J]. International Journal of Geographical Information Science, 1991, 5(2): 161-174. [24] Zhu Xu. Exploring the Advantages of Short Video Information Transmission to Strengthen Ideological Education for College Students [J]. Ability and Wisdom, 2019(24): 77. [25] Kilpatrick C, Storr J, Lim K, et al. Exploring the use of entertainment-education YouTube videos focused on infection prevention and control [J]. American Journal of Infection Control, 2018, 46(11): 1218-1223. [26] Gao Shijie, Wu Lili, Guo Chen. Research on Mobile Short Video Advertising Creation and Consumer Psychology [J]. China Market, 2019(2): 139-140. [27] Chen Chun, Li Na, Ma Jianxia. Analysis of the Current Situation of Foreign Libraries' Non-Text Resource Construction and Services and Its Enlightenment to China [J]. Library and Information Service, 2015, 59(10): 53-59.

Author Contributions: Li Xuhui: Proposed the research topic, determined the research 思路 and paper framework, revised the paper; Wu Qingfeng: Collected literature materials, wrote and revised the paper.

Research on Video Semantic Representation for Events

Li Xuhui Wu Qingfeng

School of Information Management, Wuhan University, Wuhan 430072

Abstract: [Purpose/Significance] Video content is affecting the information lives of a large number of people in China. Proper representation of video semantics is the key foundation for advancing current video content research and application services. Existing video semantic representation methods suffer from problems such as single perspective and granularity division approach for event semantic representation and lack of flexible mechanisms for semantic object changes. Therefore, exploring more effective video semantic representation methods is of great significance. [Method/Process] This paper proposes an event-oriented video semantic representation method. This method considers the bidirectional nature of human cognitive processes, can interpret and generate event semantics from different perspectives according to different user backgrounds and needs, and defines corresponding change mechanisms for semantic objects and roles. [Result/Conclusion] The event-oriented video semantic representation method has a complete semantic representation framework, supports multi-perspective event semantic representation, can flexibly perform attribute-level, object-level, and event-level semantic extensions, and can represent richer video semantics.

Keywords: video semantic representation; multi-perspective; semantic extension

---

Publication Ethics Statement

To strengthen and enhance academic norms, research integrity, and academic ethics in the entire process of academic paper writing, reviewing, and editing, establish good academic atmosphere, promote scientific spirit, resolutely resist academic misconduct, and establish and maintain a fair, just, and open academic exchange ecological environment, the *Library and Information Service* Magazine (including the editorial departments of *Library and Information Service* and *Knowledge Management Forum*) has formulated a publication ethics statement and officially released it in February 2020.

The publication ethics statement commits that the magazines will strictly abide by and implement national policies and regulations related to academic ethics and editing and publishing, standardize the behavior of authors, peer review experts, and journal editors in the entire editing and publishing process, and accept supervision from the academic community and the whole society. It includes three parts with a total of fifteen clauses: 1. Authors' Publishing Ethics (Academic papers are an important part of scientific research; Academic misconduct is a cancer of academic papers; Authors are the main contributors to academic papers; Authorship reflects authors' intellectual property and academic contributions; Academic papers must attach great importance to intellectual property and information security; Standardized citation of references is an important representation of academic norms; Great importance must be attached to the standardization of research data and management; Establish a mechanism for correction and academic self-purification). 2. Peer Review

Experts' Publishing Ethics (Peer review is an important quality control mechanism for papers; Review experts should comply with relevant ethical guidelines and codes of conduct). 3. Editors' Publishing Ethics (Editors should become the guardians of academic paper quality; Editors should play a monitoring role in academic ethics construction; Editors should become the last barrier against academic misconduct; Zero tolerance for academic misconduct).

Full text available at: http://www.lis.ac.cn/CN/column/column291.shtml

(Journal News)

Note: Figure translations are in progress. See original paper for figures.