Extraction of Semantic Video Content using Ontology Model

N.H.Angela Lincy,K.Selva

Abstract — Video-based applications has recently revealed the need for extracting the content in videos. Raw data and low-level features alone are not sufficient to fulfill the user 's needs; that is, a deeper understanding of the content at the semantic level is required. Currently, manual techniques, which are inefficient, subjective and costly in time and limit the querying capabilities, are being used to bridge the gap between low-level representative features and high-level semantic content. In this proposal, a semantic content extraction system that allows the user to query and retrieve objects, events, and concepts are extracted. An ontology-based video semantic content model that uses spatial/temporal relations in event and concept definition is introduced here. This metaontology definition provides a wide-domain applicable rule construction standard that allows the user to construct an ontology for a given domain. In addition to domain ontologies, additional rule definitions (without using ontology) are used to lower spatial relation computation cost and to be able to define some complex situations more effectively. Genetic Algorithm-based object extraction method is integrated to capture and classify the semantic content. Thus by using this proposal, the objects, events and concepts of videos are extracted more accurately.

Index Terms —Semantic content extraction, video content modeling, ontology**.**

1 INTRODUCTION

The rapid increase in the available amount of video data has caused an urgent need to develop intelligent methods to model and extract the video content. Typical applications in which modeling and extracting video content are crucial include surveillance, video-on-demand systems, intrusion detection, border monitoring, sport events, criminal investigation systems, and many others. The ultimate goal is to enable users to retrieve some desired content from massive amounts of video data in an efficient and semantically meaningful manner. There are basically three levels of video content which are raw video data, low-level features and semantic content. First, raw video data consist of elementary physical video units together with some general video attributes such as format, length, and frame rate. Second, low-level features are characterized by audio, text, and visual features such as texture, color distribution, shape, motion, etc. Third, semantic content contains high-level concepts such as objects and events. The first two levels on which content modeling and extraction approaches are based use automatically extracted data, which represent the low-level content of a video, but they hardly provide semantics which is much more appropriate for users.

Users are mostly interested in querying and retrieving the video in terms of what the video contains. Therefore, raw video data and low-level features alone are not sufficient to fulfill the user's need; that is, a deeper understanding of the information at the semantic level is required in many video-based applications. However, it is very difficult to extract semantic content directly from raw video data. This is because video is a temporal sequence of frames without a direct relation to its semantic content [1].

Therefore, many different presentations using different sets of data such as audio, visual features, objects, events, time, motion, and spatial relations are partially or fully used to model and extract the semantic content. No matter which type of data set is used, the process of extracting semantic content is complex and requires domain knowledge or user interaction. There are many research works in this area. Most of them use manual semantic content extraction methods. Manual extraction approaches are tedious, subjective, and time consuming [2], which limit querying capabilities. Besides, the studies that perform automatic or semiautomatic extraction do not provide a satisfying solution. Although there are several studies employing different methodologies such as object detection and tracking, multimodality and spatiotemporal derivatives, the most of these studies propose techniques for specific event type extraction or work for specific cases and assumptions. In [3], simple periodic events are recognized where the success of event extraction is highly dependent on robustness of tracking. The event recognition methods described in [4] are based on a heuristic method that could not handle multiple-actor events. Event definitions are made through predefined object motions and their temporal behavior. The shortcoming of this study is its dependence on motion detection. In [5], scenario events are modeled from shape and trajectory features using a hierarchical activity representation extended from [4]. Hakeem and Shah [6] propose a method to detect events in terms of a temporally related chain of directly measurable and highly correlated low level actions (subevents) by using only temporal relations.

Another key issue in semantic content extraction is the representation of the semantic content. Many researchers have studied this from different aspects. A simple representation could relate the events with their low-level features (shape, color, etc.) using shots from videos, without any spatial or temporal relations. However, an effective use of spatiotemporal relations is crucial to achieve reliable recognition of events. Employing domain ontologies facilitate use of applicable relations on a domain. There are no studies using both spatial relations between objects, and temporal relations between events together in an ontology-based model to support automatic semantic content extraction. Studies such as BilVideo [7], [8], extended-AVIS [9], multiView [10] and classView [11] propose methods using spatial/temporal relations but do not have ontology-based models for semantic content representation. Bai et al. [12] present a semantic content analysis framework based on a domain ontology that is used to define semantic events with a temporal description logic where event extraction is done manually and event descriptions only use temporal information. Nevatia and Natarajan [13] propose an ontology model using spatiotemporal relations to extract complex events where the extraction process is manual. In [14], each linguistic concept in the domain ontology is associated with a corresponding visual concept with only temporal relations for soccer videos. Nevatia et al. [15] define an event ontology that allows natural representation of complex spatiotemporal events in terms of simpler subevents.

A Video Event Recognition Language (VERL) that allows users to define the events without interacting with the low-level processing is defined. VERL is intended to be a language for representing events for the purpose of designing an ontology of the domain, and, Video Event Markup Language (VEML) is used to manually annotate VERL events in videos. The lack of low-level processing and using manual annotation are the drawbacks of this study. Akdemir et al. [16] present a systematic approach to address the problem of designing ontologies for visual activity recognition. The general ontology design principles are adapted to the specific domain of human activity ontologies using spatial/temporal relations between contextual entities.

However, most of the contextual entities which are utilized as critical entities in spatial and temporal relations must be manually provided for activity recognition. Yildirim [17] provide a detailed survey of the existing approaches for semantic content representation and extraction. Considering the above-mentioned needs for content based retrieval and the related studies in the literature, methodologies are required for automatic semantic content extraction applicable in wide-domain videos. In this study, a new Automatic Semantic Content Extraction Framework (ASCEF) for videos is proposed for bridging the gap between low-level representative features and high-level semantic content in terms of object, event, concept, spatial and temporal relation extraction.

In order to address the modeling need for objects, events and concepts during the extraction process, a wide-domain applicable ontology-based fuzzy VIdeo Semantic COntent Model (VISCOM) that uses objects and spatial/temporal relations in event and concept definitions is developed. VISCOM is a metaontology for domain ontologies and provides a domain-independent rule construction standard. It is also possible to give additional rule definitions (without using ontology) for defining some special situations and for speeding up the extraction process. ASCEF performs the extraction process by using these metaontology- based and additional rule definitions, making ASCEF wide-domain applicable. In the automatic event and concept extraction process, objects, events, domain ontologies, and rule definitions are used. The extraction process starts with object extraction. Specifically, a semiautomatic Genetic Algorithm-based object extraction approach [18] is used for the object extraction and classification needs of this study. For each representative frame, objects and spatial relations between objects are extracted. Then, objects extracted from consecutive representative frames are processed to extract temporal relations, which is an important step in the semantic content extraction process. In these steps, spatial and temporal relations among objects and events are extracted automatically allowing and using the uncertainty in relation definitions. Event extraction process uses objects, spatial relations between objects and temporal relations between events. Similarly, objects and events are used in concept extraction process. This study proposes an automatic semantic content extraction framework. This is accomplished through the development of an ontology-based semantic content model and semantic content extraction algorithms. Our work differs from other semantic content extraction and representation studies in many ways and contributes to semantic video modeling and semantic content extraction research areas.

First of all, we propose a metaontology, a rule construction standard which is domain independent, to construct domain ontologies. Domain ontologies are enriched by including additional rule definitions. The success of the automatic semantic content extraction framework is improved by handling fuzziness in class and relation definitions in the model and in rule definitions. A domain-independent application for the proposed system has been fully implemented and tested. As a proof of wide-domain applicability, experiments have been conducted for event and concept extraction for basketball, football, and office surveillance videos. Satisfactory precision and recall rates in terms of object, event, and concept extraction are obtained by the proposed framework. Our results show that the system can be used in practical applications. Our earlier work can be found in [19], [20]. The organization of the paper is as follows. In Section 2, the proposed video semantic content model is described in detail. The automatic semantic content extraction system is explained in Section 3. In Section 4, the performed experiments and the performance evaluation of the system are given. Finally, in Section 5, our conclusions and future research directions are discussed.

2. VIDEO SEMANTIC CONTENT MODEL

In this section, the proposed semantic video content model and the use of special rules (without using ontology) are described in detail.

2.1 Overview of the model

Ontology provides many advantages and capabilities for content modeling. Yet, a great majority of the ontology based video content modeling studies propose domain specific ontology models limiting its use to a specific domain. Besides, generic ontology models provide solutions for multimedia structure representations. In this study, we propose a wide-domain applicable video content model in order to model the semantic content in videos. VISCOM is a well-defined metaontology for constructing domain ontologies. It is an alternative to the rule-based and domain-dependent extraction methods. Constructing rules for extraction is a tedious task and is not scalable. Without any standard on rule construction, different domains can have different rules with different syntax. In addition to the complexity of handling such difference, each rule structure can have weaknesses. Besides, VISCOM provides a standardized rule construction ability with the help of its metaontology. It eases the rule construction process and makes its use on larger video data possible. The rules that can be constructed via VISCOM ontology can cover most of the event definitions for a wide variety of domains. However, there can be some exceptional situations that the ontology definitions cannot cover. To handle such cases, VISCOM provides an additional rule based modeling capability without using ontology. Hence, VISCOM provides a solution that is applicable on a wide variety of domain videos. Objects, events, concepts, spatial and temporal relations are components of this generic ontology-based model. Similar generic models such as [13], [21], [22] which use objects and spatial and temporal relations for semantic content modeling neither use ontology in content representation nor support automatic content extraction. To the best of our knowledge, there is no domain-independent video semantic content model which uses both spatial and temporal relations between objects and which also supports automatic semantic content extraction as our model does. The starting point is identifying what video contains and which components can be used to model the video content. Keyframes are the elementary video units which are still images, extracted from original video data that best represent the content of shots in an abstract manner. Name, domain, frame rate, length, format are examples of general video attributes which form the metadata of video.

2.2 Ontology-based modeling

The linguistic part of VISCOM contains classes and relations between these classes. Some of the classes represent semantic content types such as Object and Event while others are used in the automatic semantic content extraction process. Relations defined in VISCOM give ability to model events and concepts related with other objects and events. VISCOM is developed on an ontology-based structure where semantic content types and relations between these types are collected under VISCOM Classes, VISCOM Data Properties which associate classes with constants and VISCOM Object Properties which are used to define relations between classes. In addition, there are some domain independent class individuals.

2.3 Rule-Based Modeling

Additional rules are utilized to extend the modeling capabilities. Each rule has two parts as body and head where body part contains any number of domain class or property individuals and head part contains only one individual with a value, μ, representing the certainty of the definition given in the body part to represent the definition in the head part where $0 \leq \mu \leq 1$.

2.4 Domain Ontology Construction With VISCOM

VISCOM is utilized as a metamodel to construct domain ontologies. Basically, domain specific semantic contents are defined as individuals of VISCOM classes and properties.

3. AUTOMATIC SEMANTIC CONTENT EXTRACTION FRAMEWORK

The Automatic Semantic Content Extraction Framework is illustrated in Fig. 1. The ultimate goal of ASCEF is to extract all of the semantic content existing in video instances. In order to achieve this goal, the automatic semantic content extraction framework takes video instance, domain ontology, and the set of rules for domain. The output of the extraction process is a set of semantic contents, named video semantic content represented as object instance, event instance, concept instance.

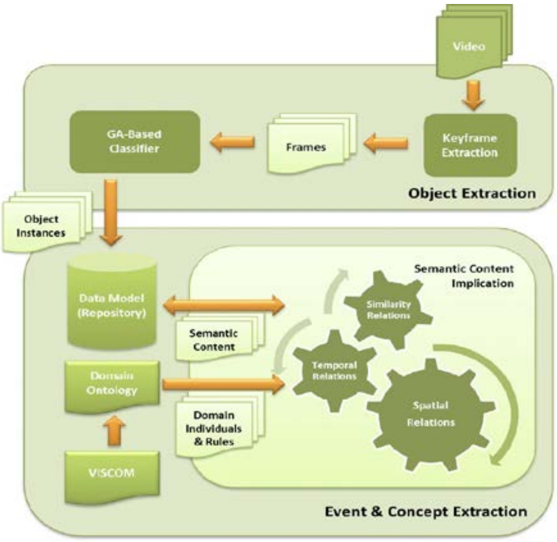

Fig. 1. Automatic semantic content extraction framework

There are two main steps followed in the automatic semantic content extraction process. The first step is to extract and classify object instances from representative frames of shots of the video instances. The second step is to extract events and concepts by using domain ontology and rule definitions. A set of procedures is executed to extract semantically meaningful components in the automatic event and concept extraction process. The first semantically meaningful components are spatial relation instances between object instances. Then, the temporal relations are extracted by using changes in spatial relations. Lastly, events and concepts are extracted by using the spatial and temporal relations by using changes in spatial relations. Lastly, events and concepts are extracted by using the spatial and temporal relations. Details about these procedures are described in below given sections.

3.1 Object Extraction

Object extraction is one of most crucial components in the framework, since the objects are used as the input for the extraction process. However, the details of object extraction process is not presented in detail, considering that the object extraction process is mostly in the scope of computer vision and image analysis techniques. It can be argued that having a computer vision-based object extraction component prevents the framework being domain independent. However, object extraction techniques use training data to learn object definitions, which are usually shape, color, and texture features. These definitions are mostly the same across different domains. Thus, using training data in such object extraction techniques does not necessarily make those techniques domain dependent. As long as the object extraction technique can identify a large number of different object types, such a technique is usable in ASCEF. In order to meet the object extraction and classification need, a semiautomatic Genetic Algorithm-based object extraction approach [18], [20] is utilized in this study. The approach is a supervised learning approach utilizing eight MPEG-7 descriptors to represent the objects. During the object extraction process, for each representative keyframe in the video, above-mentioned object extraction process is performed and a set of objects is extracted and classified. The extracted object instances are stored with their type, frame number, membership value, and Minimum Bounding Rectangle data.

IV. EMPIRICAL STUDY

The experimental part of the system contains evaluation tests on office surveillance, basketball, and football videos. Precision and recall rates and Boundary Detection Accuracy (BDA) [31] score, that are important metrics to see the performance of the retrieval systems, are used in this study to evaluate the success of the proposed framework. A semantic content is accepted as a correctly extracted semantic content when its interval intersects with the manually extracted semantic content interval. In addition, precision and recall rates are calculated according to the detected content
boundary /interval compared with the manually labeled boundary/interval with the formulas given below:

$$Precint = Tmb \cap Tdb / Tdb$$

$$Recint = Tmb \cap Tdb / Tdb$$

where _db and _mb are the automatically detected event/ concept interval and the manually labeled event/concept interval, respectively. Initially, the framework is tested with five office surveillance videos, each being 10 minutes in length. Totally, 1,026 keyframes are extracted and utilized in the extraction process. During this test, first the object extraction is done automatically and all of the semantic content extraction process is executed. Then, in order to see the effect of object extraction on the success of the system, the test is performed by providing the objects manually. The test results for the case with automatic object extraction. The videos in this test contains 50 semantic entities. After retrieving 50 entities, it is observed that 45 of them are correctly extracted, five wrongly extracted and five missed. The missed entities are the result of the automatic object extraction process that has misclassified or not extracted some of the objects. Wrong extractions are result of sensitivity of ontological rules on the object positions (i.e., small movements of person object that are not walking/ casting extracted as walking/casting) and an unsuitable class individual definition in the ontological rules (i.e., the similarity definitions of typing). Both precision and recall rates are calculated as 90.00 percent and BDA score is calculated as 78.59 percent, which shows the success of our proposed framework.

4.CONCLUSION

The primary aim of this research is to develop a framework for an automatic semantic content extraction system for videos which can be utilized in various areas, such as surveillance, sport events, and news video applications. The novel idea here is to utilize domain ontologies generated with a domain-independent ontology-based semantic content metaontology model and a set of special rule definitions. Automatic Semantic Content Extraction Framework contributes in several ways to semantic video modeling and semantic content extraction research areas. First of all, the semantic content extraction process is done automatically. In addition, a generic ontology-based semantic metaontology model for videos (VISCOM) is proposed. Moreover, the semantic content representation capability and extraction success are improved by adding fuzziness in class, relation, and rule definitions. An automatic Genetic Algorithm-based object extraction method is integrated to the proposed system to capture semantic content. In every component of the framework, ontology-based modeling and extraction capabilities are used. The test results clearly show the success of the developed system.

As a further study, one can improve the model and the extraction capabilities of the framework for spatial relation extraction by considering the viewing angle of camera and the motions in the depth dimension.

REFERENCES

[1] M. Petkovic and W. Jonker, “An Overview of Data Models and Query Languages for Content-Based Video Retrieval,” Proc. Int’l Conf. Advances in Infrastructure for E-Business, Science, and Education on the Internet, Aug. 2000.

[2] M. Petkovic and W. Jonker, “Content-Based Video Retrieval by Integrating Spatio-Temporal and Stochastic Recognition of Events,” Proc. IEEE Int’l Workshop Detection and Recognition of Events in Video, pp. 75-82, 2001. 60 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013 Fig. 7. Rule effect on spatial relation computation.

[3] L.S. Davis, S. Fejes, D. Harwood, Y. Yacoob, I. Haratoglu, and M.J. Black, “Visual Surveillance of Human Activity,” Proc. Third Asian Conf. Computer Vision (ACCV), vol. 2, pp. 267-274, 1998.

[4] G.G. Medioni, I. Cohen, F. Bre´mond, S. Hongeng, and R. Nevatia, “Event Detection and Analysis from Video Streams,” IEEE Trans. Pattern Analysis Machine Intelligence, vol. 23, no. 8, pp. 873-889, Aug. 2001.

[5] S. Hongeng, R. Nevatia, and F. Bre´mond, “Video-Based Event Recognition: Activity Representation and Probabilistic Recognition Methods,” Computer Vision and Image Understanding, vol. 96, no. 2, pp. 129-162, 2004.

[6] A. Hakeem and M. Shah, “Multiple Agent Event Detection and Representation in Videos,” Proc. 20th Nat’l Conf. Artificial Intelligence (AAAI), pp. 89-94, 2005.

[7] M.E. Do¨nderler, E. Saykol, U. Arslan, O ¨ . Ulusoy, and U. Gu¨du¨ kbay, “Bilvideo: Design and Implementation of a Video Database Management System,” Multimedia Tools Applications, vol. 27, no. 1, pp. 79-104, 2005.

[8] T. Sevilmis, M. Bastan, U. Gu¨du¨ kbay, and O ¨ . Ulusoy, “Automatic Detection of Salient Objects and Spatial Relations in Videos for a Video Database System,” Image Vision Computing, vol. 26, no. 10, pp. 1384-1396, 2008.

[9] M. Ko¨pru¨ lu¨, N.K. Cicekli, and A. Yazici, “Spatio-Temporal Querying in Video Databases,” Information Sciences, vol. 160, nos. 1-4, pp. 131-152, 2004.

[10] J. Fan, W. Aref, A. Elmagarmid, M. Hacid, M. Marzouk, and X. Zhu, “Multiview: Multilevel Video Content Representation and Retrieval,” J. Electronic Imaging, vol. 10, no. 4, pp. 895-908, 2001.

[11] J. Fan, A.K. Elmagarmid, X. Zhu, W.G. Aref, and L. Wu, “Classview: Hierarchical Video Shot Classification, Indexing, and Accessing,” IEEE Trans. Multimedia, vol. 6, no. 1, pp. 70-86, Feb. 2004.

[12] L. Bai, S.Y. Lao, G. Jones, and A.F. Smeaton, “Video Semantic Content Analysis Based on Ontology,” IMVIP ’07: Proc. 11th Int’l Machine Vision and Image Processing Conf., pp. 117-124, 2007.

[13] R. Nevatia and P. Natarajan, “EDF: A Framework for Semantic Annotation of Video,” Proc. 10th IEEE Int’l Conf. Computer Vision Workshops (ICCVW ’05), p. 1876, 2005.

[14] A.D. Bagdanov, M. Bertini, A. Del Bimbo, C. Torniai, and G. Serra, “Semantic Annotation and Retrieval of Video Events Using Multimedia Ontologies,” Proc. IEEE Int’l Conf. Semantic Computing (ICSC), Sept. 2007.