論文の内容の要旨

論文題目

Study on Semantic Content Generation Support System for Multimedia Information Analysis and Indexing

（マルティメディア情報分析・記述におけるセマンティックコンテンツ生成支援システムに関する研究）

氏名　　武 小萌

1. Introduction

In recent years, the number of videos produced has risen dramatically as a result of advances in digital broadcasting services. Connection technologies are making it easier to view multi-channel images and videos online. Although many multimedia applications have been developed, multimedia applications still lack a content management capability that would allow users to handle video content more effectively. Viewing video without the ability to interact with the video content cannot yet satisfy the user. Conventional applications cannot provide services to such an extent due to the lack of the annotation to the video objects. For browsing, searching, and manipulating video documents in a more effective and efficient way, improved multimedia systems must include new technology and tools that can manage and manipulate video content, and even generate new content.

2. System Architecture

In this thesis, a semantic content generation support system is proposed for multimedia information analysis and indexing. This system aims at mapping low-level descriptions to semantic concepts so as to identify identical object and facilitate indexing, browsing, searching, and managing the multimedia database.

This system includes three layers: a content provider layer, an application service provider layer, and an end user layer. In the content provider layer, raw video data is processed to scenes, shots, keyframes, and regions. By applying semantic object modeling and annotation with these fundamental video units as the target, a semantic content database, in which high-level indexes and low-level descriptions of each object in the video are annotated, is generated as the result of this layer. The application service provider layer provides application-oriented operations that are used to establish application-dependent services based on the output from the content provider layer. Finally, the end user layer includes an interface planner that connects the user side and the system side.

3. Semantic Content Generation Support System

Content Modeling: Low-Level Features to Semantic Concepts

Movie and variety video is the main target video source in this work. There is a limit if we only use the fixed set of low-level visual features to retrieve semantic object from the complex video source. Therefore, semantic object modeling is a desired demand. In this work, a correspondence relation between keywords and the corresponding models is created. Video is first partitioned into shots, and the middle frames in each shot are extracted to simply represent these shots. A color image segmentation algorithm is used to segment these keyframes into regions. Color, texture, area ratio, and coordinates of the Minimal Bounding Rectangle (MBR) of regions are obtained from each segmented region. An information-theoretic measure is used to automatically measure illumination instability of video.

Based on these low-level features, a database of ontological semantic object models is constructed, which allows content provider to get information about specific semantic objects. We proposed a semantic object model, which has a hierarchy structure from semantic concept, salient regions to low-level features. The system searches the database for keyframes in the video to detect similarities in detailed low-level features. Since image recognition techniques are limited in their ability to fully-automatically identify and recognize images, our proposed approaches recover a greater number of similar keyframes and thus provide higher recall rate and more relevant results. From these results, content provider can select relevant keyframes interactively, and the matched objects in them are then automatically annotated according to descriptions that are added to the model in advance.

Supporting Object Query Based on Background Classification

Semantic concepts of objects do not occur in isolation and there is always a context to the co-occurrence of objects and backgrounds in a video scene. It is believed that it can be beneficial to model this context. In this work, a block-based background classification engine is used to recognize the backgrounds where the objects appear in the video. Unlike in specific object, there are no steady features such as size, shape, and etc. in specific background to represent the location where the shots are taken. Here, to extract similar backgrounds from the video, a model image, which is a template representing the feature of the background and serves as the background model, is first selected from keyframes. Each model image is inspected to find the closest match based on the similarity measure between the model image and the target keyframe. To compute the background similarity, blocks comprised in both the query image and the target keyframe are clustered based on low-level visual features. The degree that the blocks of two images falling within each cluster overlap each other, which is referred to as the overlapping degree, is used as the background similarity.

On the other hand, a probabilistic Bayesian network is used to model the context between objects and backgrounds, and to demonstrate how this leads to an improvement in the performance of object query and annotation. The results of object annotation and background classification are used for training and the trained network is used to support the quality of further object matching.

Fuzzy Mode Similarity Measure Based on Illumination Instability

In the case of model matching, a fuzzy mode similarity measure is proposed to adaptively calibrate the feature matching criterion based on the illumination instability. The purpose of this method is to solve the video and image analysis problem without being impacted by the lighting changes. In this work, we first use an information-theoretic measure as a quantitative measure of the information distribution within an image. This measure is further extended to the video case and used to quantitatively represent the lighting condition of each single scene. The illumination instability of the video is thus measured by calculating the instability of the features extracted from the extended measure.

Based on this information-theoretic illumination instability measure, characterizing the similarity between the model regions and the retrieved keyframes becomes two issues, one being model distribution prediction and the other being similarity measure according to the degree of lighting changes and the shape of model distribution. In the fuzzy mode, the video is first segmented into scenes using the annotation obtained based on the background classification engine. With the video frame as the target, the illumination instability of each scene is evaluated. When performing model matching, the scene where the model image exists and the scene where the retrieved keyframe exists are integrated as a fuzzy set. A membership function is computed to estimate the impact of lighting changes to object distribution. Finally, the similarity between the model and the retrieved regions is measured using this computed membership function, and further to adaptively calibrate the feature matching criterion based on the illumination instability.

4. Experimental Results

We store 15 videos with total of 15,708 keyframes to generate experiments for examining the performance of the proposed approaches. In case of one video, the total processing time is 655.2 minutes, less than 11 hours. The experimental data includes 36 models from the 15 movies. 20 background models are used to examine the performance of background classification engine. For context relationship modeling, we use 40% of the keyframes for training and others for testing. Regions segmented from keyframes in training set images are used to extract objects. We compare the detection performance of object query using context relationship network as against the original one. To examine the wellness that the proposed measure characterizes the instability of the lighting condition, the relationship between the instability measure and the performance of the implemented

system is statistically simulated. Comparison between fuzzy mode similarity measure and previous mode is also carried out.

From the experimental results illustrated in the manuscript of this thesis, we can see that the proposed instability measure successfully reflects the relationship that the performance of color-based video system decreases as the lighting condition of the object video becomes instable. Two simpler and more straight-forward measures for this purpose are also implemented to compare to the proposed approach. Compared to the results of these two measures, our proposed approach shows a better reflection of illumination instability. The experimental result of semantic object model matching, fuzzy mode similarity measure and context relationship network (CRN) demonstrates significant improvement in performance of semantic object extraction by using CRN than without using it. Furthermore, we can see that the retrieval performance is improved by the fuzzy mode similarity measure, which adaptively calibrates the feature matching criterion based on the illumination instability, and the performance is uniformly better for any threshold.

The same character wearing the same costume may appear in various locations during various seasons or times of the day. The position and size of the same object in different scenes may also differ depending upon the photographic technique or filming angle used. It will be very tedious and time consuming to manually annotate all of the objects of the video content. In this work, by controlling the recall in our model to be not lower than 90% and the precision not lower than 50% respectively, a user would be able to choose one results in two for annotation. Compared to conventional approaches, the proposed system greatly holds down the indexing cost of content provider.

5. Object-Oriented Application Services

As the result, we build up a semantic content database, which stores video content metadata and the operations that manipulate such kind of metadata. Based on this database and the corresponding operations, the proposed system generates a general video management and utilization framework to provide a great potential environment for development of interactive multimedia applications. In this thesis, two novel practical applications are constructed and implemented to show the practicability of the proposed semantic content generation support system.

The purpose of the application referred to as Video Characters' Popularity Voting System (VCPVS) is to provide description annotation, retrieval and statistics to the video with an object such as a character as the basic unit over the Internet or other network. With the use of the constructed application, the user can not only view the video, but also interact with the video content, retrieve scenes in which the user is interested, and discuss video characters with other users who are also viewing the same video over the Internet.

Another application, referred to as Semantic Video Summarization System, is constructed to generate video summary on semantic level, adjusted to the taste of the viewer, and adapted to the contents which the viewer desires. Compared to conventional video summarization applications, the constructed Semantic Video Summarization System plays a better role in capturing the essence of a long video sequence and in producing a meaningful display that optimally presents the semantic events in the summarized video.

6. Conclusions

In this thesis, we have proposed a semantic content generation support system based on low-level descriptions for supporting rich video indexing and retrieval. This system aims at mapping low-level descriptions to semantic concepts so as to identify identical object and facilitate indexing, browsing, searching, and managing the multimedia database. Compared to related researches, the main contributions of this thesis include:

- Semantic objects in the video can be extracted out automatically and annotated semi-automatically merely based on low-level visual features with predefined semantic object models.
- Based on a semantic object modeling approach and a region-based model matching approach, our proposed system shows high recall rate to enhance the practicability of semantic video indexing and allows interactive video content annotation.
- A probabilistic approach is used to explicitly model the interaction between objects and the backgrounds where these objects appear based on a background classification engine.
- A fuzzy mode similarity measure is used to adaptively calibrates the feature matching criterion based on the illumination instability, and further to solve the video and image analysis problem without being impacted by the lighting changes
- Our proposed system generates a general video management and utilization framework to provide a great potential environment for development of application-oriented operations and interactive multimedia applications.

Following BS digital broadcasting from 2001, the terrestrial digital broadcasting started on Dec. 1 of 2003 in Japan, and the analogue broadcasting will be ended in 2011. Against this background the necessities of semantic video indexing system and interactive multimedia application increase rapidly. We believe that our research will contribute greatly to the related fields.