POSTER: Video High-level Semantic Recognizing

Cheng Zeng, Xiaozhu Liu, Jun Li, Feng Dan
State Key Lab of Software Engineering, Wuhan University, 430072, China
zengc@whu.edu.cn

ABSTRACT

This paper presents an event semantic recognizing method based on Markov chain by stages, which takes object semantic as the bridge and is valid for recognizing complex event semantic. Semantic concept mapping mechanism based on semantic template is used to realize the automatic recognition of video semantic. In the experiment contrasting with IMAT system, our method shows more extensive recognition range and higher accuracy. Experimental results are encouraging, and indicate that the performance of the proposed approach is effective.

Keywords

Markov chain, STTS, Semantic template, Object semantic

1. INTRODUCTION

At present, the retrieval technology based on keywords has been adopted widely by many commercial search engines. Many retrieval systems of images or video based on content have also come into being in some research institutions. However, these technologies depend too much on user's participating, or retrieval results betray the user requirement. It is urgent to strengthen the research on media semantic information mining.

Video semantic mining is one of the crucial problems of intending cross-media search engine. The current researches about it focus on mining object semantic information, static scene. But they are restricted within some certain application domains, and don't adapt to deal with complex semantic mining such as dynamic structure, long term, multi-objects, multi-relations etc.. This paper presents a novel method which integrates object semantic with relationship information among objects to realize the transition from low-level feature to high-level semantic. Semantic template, which stores the mapping relationship between physical features and semantic, is constructed to automatically recognize video semantic.

2. RELATIVE RESEARCH

Virage[And0a] is a video automatic annotating system which provides an open framework to expand other video or audio analysis tools. But it doesn't refer to the gap problem between low-level features and high-level semantic. V. Mezaris[Jou0a] corresponds low-level features description in MPEG-7 to relevant middle-level description, and constructs a simple word database called object ontology.

Benitez[Con0a] presents an intelligent information system framework MediaNet, which integrates low-level features and media knowledge concepts into a single system. ASSAVID project[Jou0b] supported by Europe IST fund develops a video retrieval system. It realizes the automatic classification of sport scenes by researching special characters of different sports. Each class corresponds to a sport name that is convenient for user to retrieve based on keywords.

Navid[Jou0c] presented a method which recognizes some special regions, such as sky, grass and so on by low dimensions color feature based on support vector machine and wavelet texture analysis. Multi-label[Jou0d] is utilized to recognize more objects and classifies visual media containing multi-objects. IBM develops a MPEG-7 video annotation system [Jou0e], which realizes video retrieval based on object or event concepts by manual annotating beforehand. In addition, the system provides a training and learning mechanism based on HMM for automatic annotation.

Jurgen et al.[Jou0f] realize the automatic marking of some objects or wonderful scenes information such as goal, football, athlete and so on in football game. Visual and audio features are utilized at the same time to recognize video event semantic in [Jouog].

3. OBJECT SEMANTIC TEMPLATE

Semantic Template Training System

Semantic template is looked as the bridge between feature layer and semantic layer. Each template maps to several semantic concepts. This paper constructs different semantic template respectively corresponding to object and event semantic. Object

This work is supported by National Basic Research Program of China(973 Program) No.2007CB310800 and the Doctor Subject Fund of Education Ministry No.20070486064 and the Hubei Province Natural Science Foundation of China No. 2007ABA038 and the High School 111 Project of China No. B07037

semantic template (OST) contains not only static feature relationship obtained by statistic learning, but also best suitable segment granularity for each object region, object sub-region features, space relationship description and so on. In addition, best suitable feature will also be stored to realize retrieving with multi-granularities. Event semantic template(EST) is based on OST. It stores time-space topology relation among regions, state transferring through time, objects appearing list and so on.

The key problem of constructing OST is how to correspond object region in video with object semantic concept (OSC) in real-world. We train visual media set by Semantic Template Training system(STTS).

Pre-processing and Concept Matching

Video pre-processing includes features extracting, video segmenting, video frame segmenting based on object and others. We take the begin-frame of video segment as original information. STTS will be utilized to obtain the mapping relationship, between object region information and OSC.

After ascertaining the object regions, we will dynamically adjust the segmenting granularity of this region and verify its tracking effectiveness in follow-up frames for each semantic object.

Constructing OST

This paper improves parameters accuracy in semantic template by iterative processing based on statistical probability until all parameters level off. The aim is that each object region segmented could be automatically mapped to the best suitable semantic concept. Moreover, we take into account the relationship among different OSC which will improve the effect of mapping and enhance the validity of video semantic expression.

We define $f_i \ = [\, f_i^1, f_i^2, \cdots, f_i^X \,]$ (i∈[1,N]) to denote video feature vector, where N is the number of video segments in training set and X is feature dimension. c^j (j ∈[1,Q])denotes an OSC, Q is the amount of OSC. OST is represented as o ∈O= {o$_1$,o$_2$,…o$_K$}, where K is the amount of OST waiting for being constructed. Each feature f is able to belong to several OST, but each OST corresponds to a single concept. The essence of overcoming semantic gap is to compute P(c|f), where OST plays the bridge role.

($f_i \ , c^j$) constructed by feature vector and semantic concept is supposed to be independent each other that could be represented as follows:

$$P(f_i \ , c^j \mid o_k) = P_\lambda(f_i \mid o_k) P_\nu(c^j \mid o_k) \quad (1)$$

Feature vector and concept are looked as random distributing in video frame. The probability, that a OST is chosen, is represented as P(o$_k$). $P_\lambda(f_i \mid o_k)$ and $P_\nu(c^j \mid o_k)$ respectively denote conditional probability of feature f$_i$ and concept c^j. If the effect of OST o$_k$ is able to be ignored, we will compute the probability of ($f_i \ , c^j$):

$$P(f_i \ , c^j) = P(c^j) \sum\nolimits_{k=1}^{K} P_\lambda(f_i \mid o_k) P(o_k \mid c^j) \quad (2)$$

$P(o_k \mid c^j)$ could be transformed by Bayesian formula. Suppose that feature vectors in whole media document accord with K Gaussian mixture distribution. So each OST o$_k$ will correspond to a Gaussian distribution.

$$p_\lambda(f_i \mid o_k) = \frac{1}{(2\pi)^{X/2} \left| \sum\nolimits_k \right|^{1/2}} e^{-\frac{1}{2}(f_i - \mu_k)^T \sum\nolimits_k^{-1} (f_i - \mu_k)} \quad (3)$$

where μ$_k$, $\sum\nolimits_k$ respectively denote the mean and covariance matrix of f$_i$ in the object region corresponding to o$_k$. In term of maximum likelihood principle, function L(θ) and ln (L(θ)) will tend to maximum with the same independent variable. Therefore, this paper takes $p_\nu(c^j \mid o_k)$ as independent variable to form the probability formula for annotating the whole training set.

$$\ln \prod\nolimits_{j=1}^{Q} \prod\nolimits_{i=1}^{N} [\, p_\nu(c^j \mid O)^{\delta_i} \prod\nolimits_{t,s=1,t\neq s}^{\delta_i} P(c^t \mid c^s)]$$

$$= \sum\nolimits_{j=1}^{Q} \sum\nolimits_{i=1}^{N} [\delta_i \ln(\sum\nolimits_{k=1}^{K} P(o_k) P_\nu(c^j \mid o_k))$$

$$+ \ln \sum\nolimits_{t,s=1,t\neq s}^{\delta_i} P(c^t \mid c^s)] \quad (4)$$

$P(c^t \mid c^s)$ denotes relationship between OSC. By similar principle, $P_\nu(f_i \mid o_k)$ could be calculated.

The above formulas gain the probability that feature f$_i$ and concept c^j are chosen, while o$_k$ is taken as prior probability. However, it is necessary to gain the posterior probability of o$_k$ for transiting low-level feature to object semantic. We deduce the following formulas based on Bayesian rule:

$$P(o_k \mid f_i, c^j) = \frac{P(o_k) P_\lambda(f_i \mid o_k) P_\nu(c^j \mid o_k)}{\sum\nolimits_{a=1}^{K} P(o_a) P_\lambda(f_i \mid o_a) P_\nu(c^j \mid o_a)} \quad (5)$$

$P(o_k \mid f_i)$ is calculated by similar way. The expectation of full likelihood estimate of $\ln P(\lambda, \upsilon, O)$ is calculated based on $P(O \mid \lambda, \upsilon)$:

$$\sum\nolimits_{(i,j)=1}^{K} \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{Q} s(c_i^j) \ln[P(o_{i,j}) P_\lambda(f_i \mid o_{i,j}) P_\nu(w^j \mid o_{i,j})] P(O \mid \lambda, \nu)$$

$$P(O \mid \lambda, \upsilon) = \prod\nolimits_{i=1}^{N} \prod\nolimits_{j=1}^{Q} P(o_{i,j} \mid f_i, c^j) \quad (6)$$

where $s(c_i^j)$ denotes the probability that the concept corresponding to feature vector f$_i$ emerges in media, o$_{i,j}$ denotes the OST corresponding to(f_i, c^j).

With the similar method, we gain the expectation of likelihood estimate $\ln P(\lambda, O)$ based on $P(O \mid \lambda)$.

The calculated $P(o_k \mid f_i)$ and formula (3) will provide the initial distribution of features in OST. And $P(o_k \mid f_i, c^j)$ provides the probability that OST is chosen.

$$P(o_k) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{Q} s(c_i^j) P(o_k \mid f_i, c^j)}{\sum_{i=1}^{N} \sum_{j=1}^{Q} s(c_i^j)} \quad (7)$$

Feedback Study of OST

After an OST has been constructed, it must be validated by retrieving in larger media set to update itself. The formula is as follows:

$$P(f_i \mid c^j) = \int P_v(f_i \mid o) P(o \mid c^j) do$$

$$= \mathrm{E}_o[P_v(c^j \mid o) \frac{P_\lambda(f_i \mid o)}{P(c^j)}]$$

$$\approx \frac{\sum_{k=1}^{K} P_v(c^j \mid o_k) P_\lambda(f_i \mid o_k)}{\sum_{t=1}^{K} P_v(c^j \mid o_t)} \quad (8)$$

In terms of formula (8), system will return M documents of maximal $P(f_i \mid c^j)$. Then user marks those satisfying results. We design a decision function ε(ξ), $\varepsilon(\xi) = \varepsilon_i$ which denotes that ξ belonging to the template o_k is accepted.

The key that judges the suitability of the constructed OST for object semantic recognition is to make the risk minimal. Otherwise, it is necessary to update OST by increasing training examples.

4. EVENT SEMANTIC TEMPLATE

This paper takes the event semantic as the result of semantic objects evolving with some certain rules, structure, or motion relation. Event semantic is inclined to describe certain interested object regions, but ignore other regions, namely observing different regions in different degrees.

For expediently constructing EST, we divide event semantic into unitary and multiunit event semantic. The first is possible to refer to one or two interested objects, but exist only one mainly observed semantic object. Another is combined with lots of unitary events with certain relationships. The interested object of each unitary event contained in multiunit event is likely to be different.

1. Recognizing unitary event semantic

Unitary event semantic is looked as the minimal unit describing event semantic. If a video segment contains one interested object, the object will directly be taken as observed object. On the contrary, if it contains several interested objects with different motion state, it is necessary to classify them.

Unitary event is actually used to describe semantic objects (classes), theirs own motion states and possibly existing action relation between two objects (classes). We are inclined to recognize abstract event semantic without domain restriction. As a result, we define 5 and 6 motion state types for single object and two objects, respectively.

Thereby, the recognition of unitary event semantic is transformed into the problems of recognizing object semantic and classifying motion states of interested semantic objects.

2. Recognizing of multiunit event semantic

Multiunit event semantic is looked as the combination of some unitary event semantic with certain temporal sequence relation. If each unitary event semantic is taken as a kind of state, multiunit event semantic will evolve into the result that a series of states transfer in turn.

This paper utilizes the theory of Markov chain to construct EST. Suppose the probability that unitary event E^t happens in time t is $\pi_t(E^t)$ and similarly E^{t+1} corresponding to $\pi_{t+1}(E^{t+1})$. If $\pi_t(E^t)$ is known, we could calculate the sum of these products which is equal to $\pi_{t+1}(E^{t+1})$:

$$\pi_{t+1}(E^{t+1}) = \sum_{E^t} \pi_t(E^t) p(E^t \rightarrow E^{t+1}) \quad (9)$$

If π_t and π_{t+1} are equal, it means the stable distribution of Markov chain. In other word, multiunit event semantic has been combined with stablest unitary event set with certain temporal sequence relation. We utilize the Sheskin algorithm to calculate the stable distribution probability by decreasing dimension and analogizing. At last, we could gain all stable distribution vectors, and then ascertain each parameter in EST.

5. EXPERIMENTS

We gain 15862 video segments in 866 video documents and train semantic templates based on different initial training sets and feedback times. The initial training numbers are 57, 130, 286 and 457. In the experiment, we define 59 OSC, 21 event semantic concepts. Each template will be used to automatically recognize and annotate the whole video database. We update these templates based on multiple times of feedback study where increase additional 20 examples at a time.

The experimental result shown in Fig.1 proves the average recognizable rate of object semantic more rely on the initial example number than feedback times. If the initial example number is much little, it will require more times of feedbacks, and it stays at a lower recognizable rate even after many times of feedbacks.

Fig.2 reflects the change of high-level semantic recognizable rate when that of object semantic gradually increases. We found event semantic has fairly high requirement for object semantic recognition accuracy. Fig.3 displays the average semantic recognizable ratio for the whole video database. It shows that STT is better than IBM' IMAT in the average recognizing ratio for most of semantic concepts. Especially for multi-objects and multi-relations event semantic, STT improves recognition accuracy and expands the recognition range.

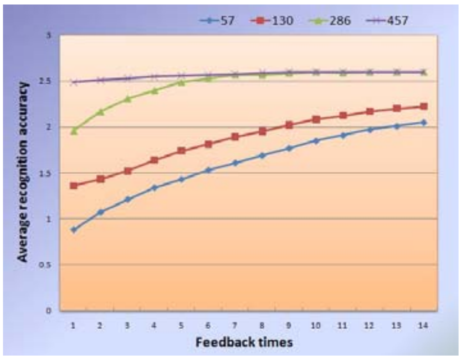

Figure.1 Object semantic recognizable rate

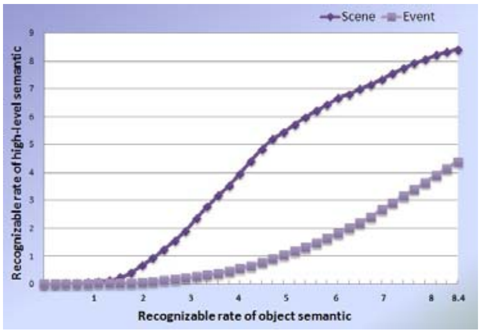

Figure.2 The relation of recognizable rate

6. CONCLUSION

This paper presents an object semantic mining method based on maximum likelihood estimate. It constructs OST by interactive study which stores the mapping relation between low-level features and semantic concept. Object semantic is taken as a bridge for realizing the transiting from low- feature to event semantic by stages. During contrasting our system with IMAT, our method is better in the range and accuracy of recognizing semantic concepts. It promotes the progress of video retrieval based on semantic.

7. REFERENCES

[And0a] Virage Inc. http://www.virage.com.

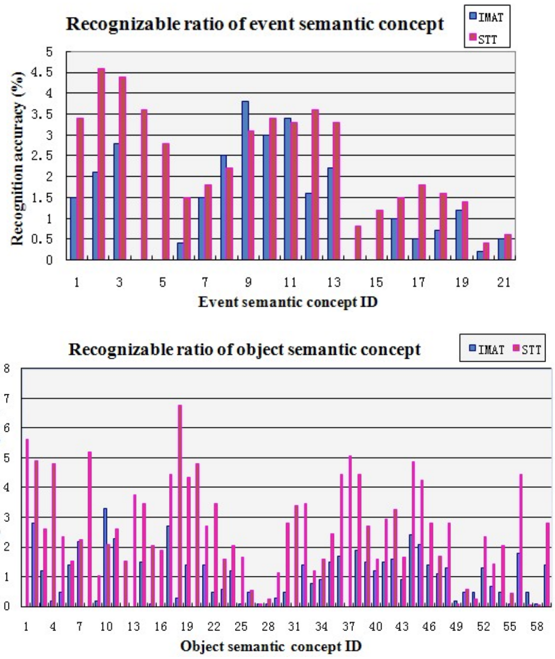

Figure.3 Average recognizable ratio of semantic concepts

[Jou0a] V. Mezaris et al.. Real-time Compressed-domain Spatiotemporal Segmentation and Ontologies for Video Indexing and Retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 14(5):606– 621, May 2004.

[Con0a] Benitez A.B., Smith J.R., Chang S.F., "MediaNet: A Multimedia Information Network for Knowledge Representation", Proceedings of the SPIE 2000 Conference on Internet Multimedia Management Systems, Vol. 4210, 2000.

[Jou0b] K. Messer, et al. A Unified Approach to the Generation of Semantic Cues for Sports Video Annotation. Signal Processing 85 (2005) 357–383

[Jou0c] Navid Serranoa, Andreas E. Savakis, Jiebo Luo. Improved Scene Classification Using Effcient Low-level Features and Semantic cues. Pattern Recognition, 37 (2004) , p1773–1784

[Jou0d] Matthew R. et al. Learning Multi-label Scene Classiffcation. Pattern Recognition 37 (2004):1757 -1771

[Jou0e]Arnon Amira, Sankar Basub, Giridharan Iyengar. A Multi-modal System for the Retrieval of Semantic Video Events. Computer Vision and Image Understanding, V96, 2004, p216-236

[Jou0f] Jurgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo, Walter Nunziati. Semantic Annotation of Soccer Videos: Automatic Highlights Identification. Computer Vision and Image Understanding, v92, v2003, p285–305

[Jou0g] Goh, et al. Audio-Visual Event Detection based on Mining of Semantic Audio-Visual Labels. http://www. merl.com/papers/docs/TR2004-008.pdf, 2004