Multilayer Video Semantic Feature Extraction Method

Xian Zhong[1], Tianbao Yu [1], Yansheng Lu [2]

1.School of Computer Science and Technology, Wuhan University of Technology, China;
2. School of Computer Science and Technology, Huazhong University of Science and Technology, China

Keywords: multilayer; video semantic; MVSS; feature extraction.

Abstract

To solve video semantic retrieval, the key problem is the difficulty of analyzing, searching and processing the massive video data. In this paper, we propose an innovative multilayer video semantic feature extraction method. The whole method can achieve automatic extraction, label of the concept ontology of video semantic and provide support for the semantic retrieval. The main idea of the method includes the MPE (Model Parameter Estimation) algorithm, MR(Model Replace) algorithm and MVSS(Multilayer Video Semantic Symbolic representation) algorithm. Express the semantic information of the whole video key-frame from three levels. Furthermore, we examine the validity of those algorithms by labelling the video object ontology. The experimental results show that the method proposed in this paper can improve the accuracy rate of the video semantic feature extraction.

1 Introduction

Video semantic feature extraction is the technical foundation of video semantic retrieval. A good video semantic feature extraction method can achieve high-quality and high-efficiency video semantic retrieval. Thus, the keys to the study are video semantic feature extraction methods and related technologies.

As we all know, the semantic information of the video data is extremely rich. So selecting the appropriate video semantic features and constructing the correct video semantic extraction method are the keys to accurate retrieval of video data for video content analysis, etc. [14]. Because the body has a strong ability to identify the semantics, and the difference between different ontology is large and invariance, in this paper we use ontology to characterize the semantic feature. Our main study is in connection with the semantic feature extraction of the video, that is, how to effectively extract the appropriate video semantic retrieval features.

2 Related research

Now about the widely used video semantic feature extraction methods can be summarized as the following categories.

Probability and statistics-based approach is to see the video semantic feature extraction as a classification problem of the video semantic feature to be extracted. Currently, the research and application of these methods are widely, the more mature among which are the Bayesian networks[1], Naive Bayes [1], Decision Tree [11], etc.

Statistical learning method makes use of the relevant statistical knowledge and machine learning to automatically construct video semantic feature extraction methods. Currently, the typical methods are Support Vector Machine [3], canonical correlation analysis[15], as well as Gaussian Mixture Model, Hidden Markov Model [12].etc.

Rule-based reasoning is mainly based on the characteristics of the video content, and combine with professional theory to develop the relevant rules, according to which to derive the semantic feature extraction video [9,6]. Currently, these methods focus on applying the fuzzy theory into the developing of the rules [5].

Through all above research [4] for video semantic feature extraction, this paper proposes a video semantic feature extraction framework, aiming the actual needs and the target of video semantic retrieval. The main idea of the framework is to explore the use of the machine automatically label video semantic features ontology approach, based on the method of probability statistics and relative acknowledge.

3 Multilayer video semantic feature extraction framework

Video semantic concepts include scenes, objects and events categories [7]. This paper focuses on those concepts in the video, where the video object refers to a physical entity representation of the object.

We set three systems: scenes, objects and events. Scenes include day and night, city and countryside. Objects include car, human and bridge, etc. Events are about the objects' relationship, such as the cars collision and traffic accident, etc.

Conditional Random Fields is a conditional probability distribution model which gives a set of input conditions and outputs another set of random variables. The feature of the method is to assume that the output random variables constituting a Markov random field [8,10,16] (Markov Random Field, MRF) . Conditional Random Fields is very suitable for time series of random variables labeling problem [2], and the semantics extraction of video data just satisfies these data features and applications. Thus, our object research work is mainly based on conditional random fields to build video semantic feature extraction method.

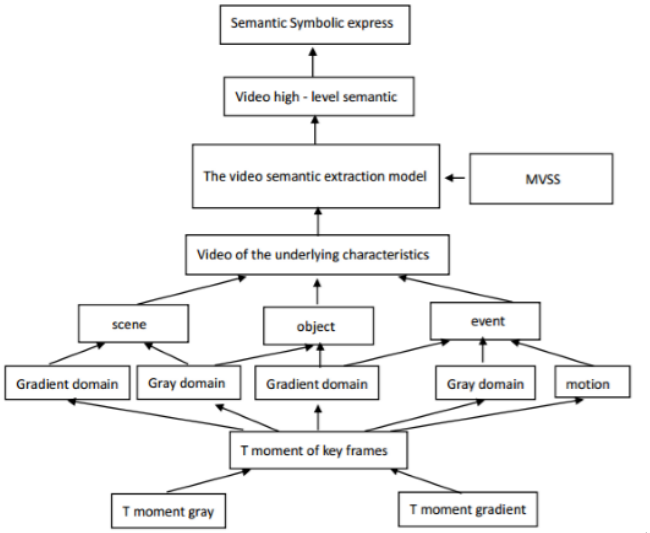

Figure 1: Video semantic feature extraction framework

Figure 1 shows the framework of video semantic features extraction. These methods are based on statistical probability knowledge, in order to facilitate formal description of each sub-method, where first define several related concepts.

Definition 1 Video Key-frame Sequence: extracts the key frames of one video, and obtains the sequence of the key frames from combined extraction in the chronological order, called a video key frame sequence, which is expressed as$\{K_t|t \in N\}$. Here, the variable t is defined as the time scale video key frame sequence. We will compare each key-frame to get the whole semantic information [16].

Definition 2 Corpus: is one space to store the three levels information of the video: scenes corpus we call it **B**, which includes many members called $\mathbf{B}_i\{i \in N\}$, objects corpus we call it **O**, the member call it $\mathbf{O}_j\{j \in N\}$,and the events call **E**, the members call $\mathbf{E}_k\{k \in N\}$.Events will show what happened in the video. And for objects level, we use the ontology to express the semantic information. For example, the video object of the sequence video key frame at time t can be expressed domain ontology Ot . So that if at the time t, x is a pixel video key frame, then the corresponding body region can be labelled as **O**t(x)

We will locate each video frame into a multilayer object map. Each video frame contains the background. In this context, there are many objects and a definite relation between each object, evolved into said events and occurrence of an event which also happen in this context. The relationships are shown in Figure 2. For each level tagging, according to the structure, you can get what we call semantic information, and these labels correspond to the elements of our corpus.

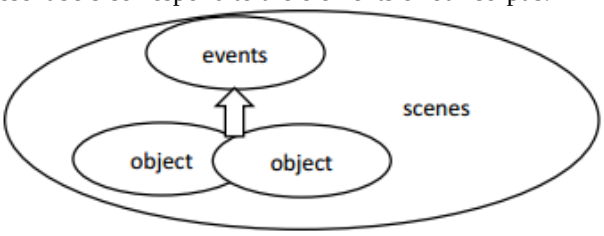

Figure 2: Multilayer semantic information relationship of video

Definition 3 Video Corpus Store: We approach each matrix element storage corpus for training early samples, stored as shown in the video after the characteristics of the underlying key frame, key frame extracted compare corresponding to each corpus, and find the corresponding elements formed label, finish building the string, it is in the form of a word to show the contents of a video screen as described, that is, video semantic retrieval executives express.

Definition 4 Video Low level Feature Field:a random domain which constituted by the image characteristics of the lower pixel is called low-level features video domain. As video key frame region, which can be represented as **F**. For example, the low-level features of pixel x at time t can be represented as a random vector field $\mathbf{F}_t(x)$.

Definition 5 Video Image Pixel Field: given a video sequence, the set constituted by all the pixels of the key frame on one moment is called the video image pixel domain, and also called the time video key frame space Domain, then expressed as **X**. For example, at t time, the pixel domain of video key frame can be represented as a set $\mathbf{X}_t$.

Definition 6 Video Semantic Rule Define: Scene: We use the new and the best methods to do the scenes' express, and compare for the video. Objects: we use the MRF to support the whole system, use MPE and MR algorithm to catch the objects' semantic information. Events: we store some common scenes, such as the cars collision, fire, pedestrian retrograde and vehicle breakdowns. Etc. The main judge whether a person walking in the tunnel for the video surveillance, is part of abnormal behavior. We can create a combination vehicle event features such as video surveillance main fire vehicles, vehicles retrograde, vehicle collision, the interaction of people and vehicles and so on.

4 Multilayer video semantic feature extraction method

4.1 Video low-level features extraction method

For low-level features of the video, each pixel will be mapped to the corresponding high-level semantics. We use gray feature, gradient characteristics, and motion feature to represent low-level features of the video. Final presentation form will be a string of characters.

In this context, low-level video feature is a random fields **F**t Statute, constituted by the gray, gradient, motion feature. Either of a pixel x on the Statute of the video sequence {**K**t} is ruled as a random vector and expressed as:

$$\mathbf{F}t(x)=(\mathbf{I}t(x),\mathbf{G}t(x),\mathbf{E}t(x)) \quad (1)$$

Definition 7 Gray Field: at time t, the video key frame pixel x can be expressed as: **I**t(x).

Definition 8 Gradient Field: At time t, the gradient model video key frame on a pixel x can be expressed as:

$$\mathbf{G}t(x)=[\mathbf{I}t(x+\triangle x)-\mathbf{I}t(x),\mathbf{I}t(x+\triangle y)-\mathbf{I}t(x)]^T \quad (2)$$

And the gradient approximation model is subject to a Gaussian distribution.

Definition 9 Model Parameter Estimation: assume that the parameters δ represents the concept of the object ontology

category, set $\theta = (\lambda, \mu, \Sigma)$. Use iterative algorithm to estimate the parameters in this model.

Algorithm 1 MPE

Input: x and δ

While !convergence

$$\sum_{i=1}^{m} \log p(x_i;\theta) = \sum_{i=1}^{m} \log \sum_{\delta} p(x_i, \delta;\theta) =$$

$$\sum_{i=1}^{m} \log \sum_{\delta} q_i(\delta_i) \frac{p(x_i, \delta_i;\theta)}{q_i(\delta_i)} \leq$$

$$\sum_{i=1}^{m} \sum_{\delta} q_i(\delta_i) \log \frac{p(x_i, \delta_i;\theta)}{q_i(\delta_i)}$$

$$q_i(\delta_i) = \frac{p(x_i, \delta_i;\theta)}{\sum_{\delta} p(x_i, \delta_i;\theta)} = \frac{p(x_i, \delta_i;\theta)}{p(x_i;\theta)} = p(\delta_i \mid x_i;\theta)$$

For each x

$$q_i(\delta_i) := p(\delta_i \mid x_i;\theta)$$

End For;

$$\sum_{i=1}^{m} \sum_{\delta} q_i(\delta_i) \log \frac{p(x_i, \delta_i;\lambda,\mu,\Sigma)}{q_i(\delta_i)} =$$

$$\sum_{i=1}^{m} \sum_{\delta} q_i(\delta_i = O_i) \log \frac{p(x_i \mid \delta_i = O_i;\mu,\Sigma)\, p(\delta_i = O_i;\lambda)}{q_i(\delta_i = O_i)} =$$

$$\sum_{i=1}^{m} \sum_{\delta} \lambda_{O_i} \log \frac{\frac{1}{(2\pi)^{n/2} \mid \Sigma_{O_i} \mid^{1/2}} \exp[-\frac{1}{2}(x_i - \mu_{O_i})^T \Sigma_{O_i}^{-1} (x_i - \mu_{O_i})]\lambda_{O_i}}{\lambda_{O_i}}$$

Set the partial with respect to λ, μ, Σ equal to zero

$$\theta^* := arc\max_{\theta} \sum_{i=1}^{m} \sum_{\delta} q_i(\delta_i) \log \frac{p(x_i, \delta_i;\theta)}{q_i(\delta_i)}$$

End While;

Output: $\theta = (\lambda, \mu, \Sigma)$

Definition 10 Motion Feature: the inter-frame difference is used to get the motion point, and then the whole motion feature can be get from every frame, expressed as **E**t(x)**.**

4.2 Video high-level semantics extraction method

The following content comes from the analysis of spatial and temporal information between statistical pixel point for video semantic feature extraction framework and the further object semantics extraction methods, which help to enhance the accuracy of video semantic extraction.

4.2.1 Scenes information analysis and modeling

In the video sequences, the pixels of the same key frame influence with each other. Rationally using such relationship helps to improve the quality of video semantic extraction. Here, we focus on the same key-frame pixel x in the spatial domain immediately adjacent four pixels vertically and horizontally, expressed as: **S**={x1,x2,x3,x4}. x1, x2, x3 ,x4 are the surround pixels of the pixel x**.** Background mode takes the edges and corners background to define, that is to say ,uses the whole underlying characteristics.

4.2.2 Object information analysis and modeling

For video sequences, rational using MRF and space and time domain information of video sequence also helps to improve the quality of video semantic extraction. Here, we focus on two adjacent key frames, such as, in the video sequence of time-domain information associated with the same position between pixels and pixels. For object annotation, the entire process using Jensen inequality with the model updating algorithm

Algorithm 2 MR

Input: x

While !end

$$p(O_{t+1} \mid F_{1\ldots t+1}) = \frac{p(O_{t+1} \mid F_{1\ldots t})\, p(F_{t+1} \mid O_{t+1})}{p(F_{t+1} \mid F_{1\ldots t})}$$

$$\propto p(O_{t+1} \mid F_{1\ldots t})\, p(F_{t+1} \mid O_{t+1})$$

$$= (\sum_{O_t} p(O_{t+1}, O_t \mid F_{1\ldots t}))\, p(F_{t+1} \mid O_{t+1})$$

$$= (\sum_{O_t} p(O_{t+1} \mid O_t)\, p(O_t \mid F_{1\ldots t}))\, p(F_{t+1} \mid O_{t+1})$$

$$\propto \exp(-\sum_{x \in C} (\Psi_{I_t(x)}(O_t(x) \mid F_{1\ldots t}(x)) + \sum_{y \in N_x} \Psi_{G_t(x)}(O_t(x), O_t(y))))$$

$$\prod \exp(\sum_{x \in X} \alpha_t \omega_t(x_t, x_{t+1}) + \sum_{y \in N_{x(t+1)}, t} \beta_t \gamma_t(x_{t+1}, y))$$

$$\prod_{x \in X} N(I_{t+1}(x); \mu_{O_t, t+1}, \sigma_{O_t, t+1}) N(G_{t+1}(x); \mu_{G, t+1}, \Sigma_{G_t, t+1})$$

End While

Output: Object semantic

4.2.3 Events information analysis and modeling

We can get the information of the objects' activities, then get the Events. We use the motion feature, combine Gradient and Gary domain, and use the former to make rule.For the former mentioned events rule, for the video, we cut the video to get the key frame, then compare each key frame. The events can be captured by the corpus.

For video retrieval, the exact extent of the rules largely determines the accuracy of the events that happen in the video identified. Thus, creating and updating the rules is particularly important. To better describe an event, usually at least three to five job picture composition are needed. So creating a rule is described by several groups of standard templates conduct training and extraction process.

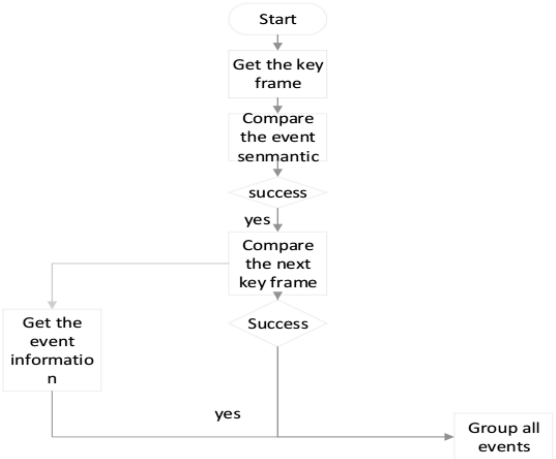

Figure 3: The progress of the events semantic

There are two ways to create a general rule. One is to enter a path to set the standard template of the same event, the groups of pictures under the path for training, and to give the more

precise rules. Another is a combination of rules, according to simple rules combined to create complex rules. Create a new rule is a simple process, so users can create their own input through the image path composition rules. The sequence of events for the object, the strategy adopted rules eigenvalues of each group objects determined to give final rule characteristic value averaging. Eigenvalues generate new rules require verified before use the basic rules. General, getting events semantic process is shown as figure 3.

4.3 Video semantic annotation method

According to the above video semantic retrieval framework, we will eventually know that the video object annotation is a string of characters that **SOE** (scenes, object, event) structure. In the process of learning how to use a video sample based on the realization of the object of the video semantic annotation, the entire process, we use the MVSS algorithm.

Algorithm 3 MVSS
Input: sample x
Step 1: Extract video key frame sequence, and make redundant processing;
Step 2: For each frame matching;
Step 3: Match corpus, extract the contents of each frame;
Step 4: The semantic content of the sign;
Output: symbolic semantic

According to our library corresponding semantics Semantic, respectively corresponding to the selected elements, as follows compositions **B**i⊕**O**j⊕**E**k to obtain the desired structure.

Here, we use parameters **λ** to achieve the reference of model parameter estimation method. **λ** is taking from 0 to 1 to analyze its impact on the performance of the methods.

Select accuracy rate (Precision) and recall (Recall) to measure the performance of this method, which are defined as follows:

$$P = \frac{TP}{TP + FP} \text{ (Accuracy rate)} \quad (3)$$

$$R = \frac{TP}{TP + FN} \text{ (Recall)} \quad (4)$$

Where, TP refers to the correct number of objects semantics; FN indicates the wrong number of object semantics; FP is the number of correct object semantics are extracted as errors.

5 Experimental analysis

Data Source: This study based on Wuhan municipal projects: Wuhan tunnel intelligent monitoring system (Hankou Railway Station, Yue Ma chang, Wuchang Railway Station , etc.), we randomly selected part of the monitoring video data as our experiment data set.

5.1 Experimental preparation and environment

Experimental contents are car and pedestrian on the urban transportation video surveillance system main roads and tunnels, and the video scenes relatively fixed.

Experimental system used Microsoft's Visual Studio 2010 as a development tool. Middleware used OpenCV2.3.1 running under the Dell PowerEdge T620 Main Server Chassis.

5.2 Training samples and performance evaluation

We used some pictures for the experiment preparation in each part and randomly selected some key frame sequences as the test samples and for the test of λ. Use marked symbol to replace the semantic information.

5.3 Experimental results

For the three parts of the semantic information, different semantic categories of images require different combinations of features to get the better described. In the recognition of background module, using the HSV color feature and co-occurrence matrix can describe what we call scene, and then store and retrieve the background. For the object extraction sectors, give the introduction of the concept of the body, then get use of the method based on conditional random fields, last identify objects and extract the information. In the event extraction session, we can do the motion detection and track vehicles on the video. Figure 4 shows the operation of the interface in this experiment Platform. Finally we use the symbolic representation to express the semantic content of the video.

Figure 4: Interface of the experiment Platform

Fig5-Fig7 are the semantic ontology results of the video. We can see four lanes in the right of the Figure5, the background scene semantic is S=road, which has lanes during the day. B=vehicle, E=running. So the figure shows that many vehicles running on the road at the tunnel.

Figure5: Vehicles ontology marked results at the tunnel.

Figure 6: Person and car ontology marked results on the road

From the figure 6, we find that a car park in a wrong way on the right, and a person is front of the car, this may be a dangerous phenomenon. And we can get the objects from the picture information, B_1=car, B_2=person, B_3=road.

Figure7: The irregular events

From the figure 7 we can get car drive in a wrong direction, with fire and make trouble on the road. The experimental system in video motion detection and tracking vehicles and pedestrians on the basis of complete video objects automatic semantic ontology annotation, Figure 8 gives the impact of parameters on the precision and recall rates.

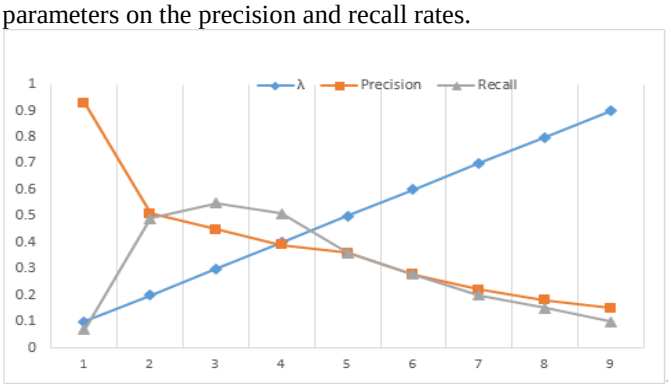

Figure 8: the effect of parameter λ on precise and recall rate

The accuracy rate and the recall rate of all the pixels in the video frame are extracted as the basic test unit of the object semantics and the semantic entities are extracted as the semantic objects respectively. The data are as follows: P1 R1 is this paper method and P2, R2 refers to Ren's method [13].

	pixels in the video frame extracted as the semantic objects		semantic ontology extracted as the semantic objects	
P	P1=78.95%	P2=67.42%	P1'=60.00%	P2'=50.53%
R	R1=71.42%	R2=62.50%	R1'=69.23%	R2'=46.80%

Table 1: the compare of pixels and semantic ontology.

From this table and these pictures, the proposed method of semantic entities can track at the same time also concept of semantic ontology annotation. Comparing with related methods currently, we have a higher precision and recall rate.

6 Conclusion and future work

This paper analyzes the main technical methods and related work on video semantic retrieval at home and abroad. From the video scenes, objects, events, combined with the each frame changes, a video semantic feature extraction framework is proposed to achieve a video semantic features extraction. And a series of experiments verify the reasonableness of the research methods and effectiveness, as well as related algorithms proposed availability. For video semantic retrieval demands a video semantic feature extraction model is established to achieve a low-level features to high-level semantic feature mapping, analyze the semantic extraction efficiency, effectively implement the video semantic extraction. The following, a more accurate and high efficient method is needed to obtain a better result.

Acknowledgements

This work was supported by Natural Science Foundation of Hubei Province (2015CFB525), National Natural Science Foundation of China (61003130) and National Key Technologies R&D Program of China(2012BAH33F03)

References

[1] BISHOP C M. "Pattern recognition and machine learning", *New York: Springer-Verlag*, (2006).

[2] Bertini M, Bimbo A. "Ferracani A, et al Interactive Multi-user Video Retrieval Systems", *Multimedia Tools & Applications*, **62**,pp.111-137,(2013).

[3] CORTES C, Vapnik V. "Support-vector Networks", *Machine Learning*, **20**,pp. 273-297,(1995).

[4] Gao Wen. "The Key Issue in the Network Video Service", *Communications of the CCF*, **9**,pp.25-29, (2013).

[5] Laasri, El Hassan Ait, et al. "A fuzzy expert system for automatic seismic signal classification". *Expert Systems with Applications An International Journal*, **42**.pp.1013-1027, (2015).

[6] Liu, Xiao, et al. "Joint shot boundary detection and key frame extraction.", *International Conference on Pattern Recognition IEEE*, pp.2565-2568,(2012).

[7] Lin C, Tseng B. "Segmentation, Classification and Watermarking for Image / video Semantic Authentication", *IEEE Workshop on Multimedia Signal Processing*, pp. 359-362,(2002).

[8] LI S. "Markov Random Field Modeling in Image Analysis", *Advances in Computer Vision and Pattern Recognition*, (2009).

[9] Metze F, DING D, Younessian E, et al. "Beyond Audio and Video Retrieval:. Topic-oriented Multimedia Summarization", *International Journal of Multimedia Information Retrieval*, **2**,pp.131-144,(2013).

[10] Popoola O, Wang K. "Video-based Abnormal Human Behavior Recognition-A Review", *IEEE Transactions on Systems Man & Cybernetics Part C*, **42**,pp.865-878,(2012).

[11] QUINLAN J R. "Induction of decision trees", *Machine Learning*, **1** ,pp.81-106,(1986).

[12] Rabiner L, Juang B H. "An Introduction to Hidden Markov Models", *IEEE Assp Magazine*,**3**.pp.4-16,(1986).

[13] Ren Xi. "Research on Video Semantic Extraction", *Huazhong University of Science and Technology*, (2013).

[14] Wang Juan, Jiang Xinghao, Sun Tanfeng. "Review of video abstraction", *Journal of Image and Graphics*, **19**,pp.1685-1695,(2014).

[15] Jin, Kai, H. C. Feng, and T. Yang. "Multi-modality Video Scene Segmentation Algorithm with Semantic Concept.", *Journal of Chinese Computer Systems*,117,pp.2001-2014,(2014).

[16] Zhong.X, Li L. "Video semantic feature extraction model", *Computer intelligent Computing and Education Technology*,(2014).