SEMANTIC-BASED TRAFFIC VIDEO RETRIEVAL USING ACTIVITY PATTERN ANALYSIS

Dan Xie[1], Weiming Hu[1], Tieniu Tan[1], Junyi Peng[2]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] Beijing University of Aeronautics and Astronautics, Beijing, China

ABSTRACT

In this paper, a semantic based retrieval framework for traffic video sequences is proposed. In order to estimate the low-level motion data, a cluster tracking algorithm is developed. A novel Hierarchical Self-Organizing Map is applied to learn the activity patterns. By using activity pattern analysis and semantic concepts assignment, a set of activity models are generated, which are used as the indexing key for accessing video clips and individual vehicles in the semantic level. The proposed retrieval framework supports various queries including query by keywords, query by sketch and multiple object queries.

1. INTRODUCTION

The use of vision-based camera systems in surveillance has grown exponentially in recent years, which produces large amount of video data stored for future use. In this context, the design of efficient indexing and retrieval techniques in video databases becomes an important issue. For surveillance video retrieval, object motion stands out as the best cue because it captures the rich dynamic content of video. Many retrieval systems have been developed by using the motion information. In [1], PCA is applied to reduce the dimensionality of trajectory data. In [2] each trajectory is segmented using wavelet decomposition and indexed based on velocity features. In [3], the motion model is set using polynomial curve fitting and used as the indexing key.

Although motion information has been widely used to represent the content of video clips, semantic gap is still existent between users and retrieval systems. Sometimes the decomposition and approximation of a trajectory cannot correctly represent original semantic meanings of it. The most common way to realize semantic retrieval is to annotate the video data with keywords, which are assumed to be defined by a human where extraction of content through image processing from video is hardly possible. Unfortunately, it is too expensive to go through manual annotation with large databases. As an effort to solve this problem, a surveillance video retrieval framework is proposed in this paper. After the motion trajectories are extracted, the activity patterns of moving targets are mined and a set of activity models are established to represent the general and abnormal behaviors of vehicles. Assigned with semantic concepts, the activity models are used as the indexing key for accessing the individual vehicle in the semantic level.

Multiple object tracking from natural video sequences is still a hard issue and an important research topic. Many algorithms [3][4][5] have been proposed in recent years. However, when focusing on the surveillance task in crowded traffic scenes with severe occlusions, most existing tracking algorithms fail because the computational complexity and cost increase dramatically. To address this problem, a robust fuzzy clustering based tracking algorithm is also proposed, which will be introduced in Section 2. Subsequently, the activity pattern analysis is introduced in Section 3 and the mechanism of indexing and retrieval is proposed in Section 4. Sections 5 and 6 give the experimental results and the conclusion.

2. CLUSTERING BASED TRACKING

The proposed tracking algorithm is based on the principle that a moving target always produces a cluster of pixels in the feature space and the distributions of the clusters change little between consecutive frames. The cluster centroids are shifted to fit the pixels' distributions in each subsequence frame by the synergy between a fuzzy clustering algorithm and a dynamic adaptation module. Figure 1 illustrates the framework of our tracking system.

After background subtraction, each foreground pixel is described by a feature vector *f* containing its coordinates *X*, *Y*, velocities V_x, V_y and color in the RGB space.

$$f = (X, Y, w_v \cdot V_x, w_v \cdot V_y, w_c \cdot R, w_c \cdot G, w_c \cdot B)$$

where the weighting factors w_c and w_v describe the relation between color and position and that between velocity and position. Velocities can be estimated using optical flow algorithm. Concentrating on the speed

0-7803-8554-3/04/$20.00 ©2004 IEEE.

requirements of the algorithm, we use a correlation method derived from [6][7]. Given a search region $R(P)$ of a foreground pixel *p(x,y)* in the current frame I_t, we define the similarity metric *D* for $p(x,y)$ and the pixel $\hat{p}(\hat{x},\hat{y})$ in previous frame I_{t-1} as:

$$D(p,\hat{p})=\sum_{i=-N}^{N}\sum_{j=-N}^{N}\left|p(x+i,y+j)-\hat{p}(\hat{x}+i,\hat{y}+j)\right|,\ \hat{p}\in R(p)$$

where *N* defines the radius of a neighborhood patch of a pixel. The optical flow of *p* can be estimate as $\arg\min_{\hat{p}\in R(p)}\left[D(p,\hat{p})\right]-p$.

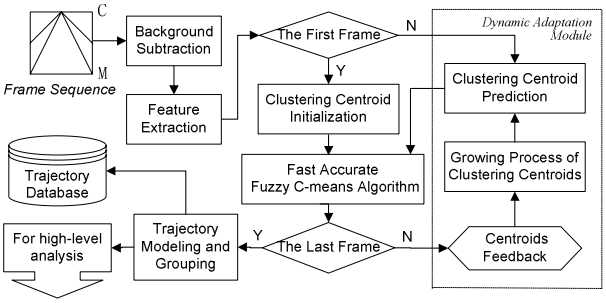

Figure 1. Framework of the cluster tracking system

In clustering analysis, a fast accurate fuzzy c-means algorithm [8] is employed, which can not only provide a data partition that is more meaningful and stable than hard clustering approaches (e.g., k-means) but also overcome the time consuming problem in traditional fuzzy c-means (FCM) algorithm. In the first phase of the fast FCM, data reduction is performed using sample quantization and aggregation. It creates a set of new samples X' representing a reduced-precision view of the original dataset X without adversely affecting clustering quality. Then the representative samples X' are clustered using a weighting FCM, in which the value of each cluster centroid is calculated by:

$$V_{ij}(t+1)=V_{ij}(t)+\frac{\sum_{l=1}^{M}R_{lj}(t)\cdot w_l\cdot(X'_{li}-V_{ij}(t))}{\sum_{l=1}^{M}R_{lj}(t)},\quad 1\le i\le N,\ 1\le j\le K$$

where *M* and *K* denote the numbers of representative samples and cluster centroids; *N* denotes the dimension of feature vectors. $R_{lj}(t)$ is the membership and w_l is the weight associated with each representative sample.

To produce a new set of initial prototypes for the next frame, all cluster centroids should be adapted according to a prediction algorithm, which preserves fixed linkages between pixel clusters (moving targets) and corresponding cluster centroids. In this way the complex matching and tracking processes are not required. To minimize any computing cost, a fast prediction algorithm, double exponential smoothing-based prediction (DESP) [10] is employed as equations (1) to (3),

$$\vec{Sc}_t=\gamma\vec{C}_t+(1-\gamma)\vec{Sc}_{t-1} \quad (1)$$

$$\vec{Sc}_t^{[2]}=\gamma\vec{Sc}_t+(1-\gamma)\vec{Sc}_{t-1}^{[2]} \quad (2)$$

$$\hat{\vec{C}}_{t+1}=\left[2+\frac{\gamma}{(1-\gamma)}\right]\vec{Sc}_t-\left[1+\frac{\gamma}{(1-\gamma)}\right]\vec{Sc}_t^{[2]} \quad (3)$$

where $\vec{C}_t=(x,y,v_x,v_y,r,g,b)$ denotes the value of a clustering centroid in current frame and the degree of exponential decay is determined by the parameter $\gamma\in[0,1)$. Furthermore, to depict the whole trajectories of vehicles, each individual in the set of cluster centroid should be born or erased according to the target-entering/leaving event. Once a video clip is processed, the centroids trajectories should be grouped into vehicle trajectories offline by a grouping module using a common motion constraint. The proposed tracking algorithm has good performance in crowded traffic scene with severe occlusions and clutter effect.

3. ACTIVITY PATTERN ANALYSIS

Given sufficient motion trajectories (generally thousands trajectories in hours tracking), we can analyze the activity patterns using a learning algorithm. An *activity* is described by a *Spatio-Temporal Trajectory (STT)* $T_{ST}=\{f_1,f_2,...,f_i,...,f_{n-1},f_n\}$, $f_i=(x_i,y_i,dx_i,dy_i)$, which contains not only the spatial information represented by the motion trajectory but also the temporal information represented by the velocities in each sample point in the trajectory. Therefore, if two targets moving at different speeds, their activities can be classified into two classes although their spatial trajectories are highly similar. The *Activity Models (AM)* are established through activity pattern learning algorithm, and are also described by STTs. An AM is the cluster center and the representation of a class of activities that have common motion characters and semantic meanings.

For activity pattern learning, a Hierarchical Self-Organizing Map (HSOM) [9] is developed. We introduce a novel hierarchical structure to SOM to overcome the convergence problem in high dimensional space. In traditional SOM, neurons in the output layer are relatively independent. In our method, a set of neurons that have certain intimate relationships constitutes a group that is defined as an internal net; all neurons are partitioned into several internal nets; and all internal nets constitute one external net. The relationships between neurons in an internal net are defined by neuron neighborhoods. An internal net can be treated as a "big neuron" and the

relationships between internal nets in the external net are defined by internal net neighborhoods. Both the neurons in an internal net and all the internal nets in the external net are trained according to the SOM learning procedure.

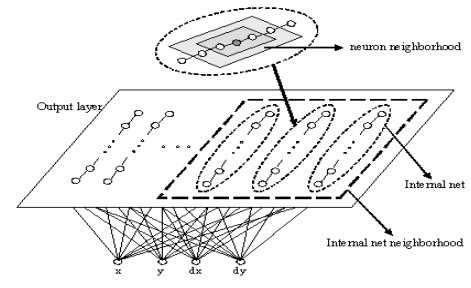

Figure 2. Network structure of HSOM

In our system, all the output neurons are linked to *l* lines. Each line corresponds to an AM and forms an internal net. The relationship between all internal nets forms the external net. The neural network is learned from a sequence of corresponding movements among neurons, and gradually organizes towards an optimal solution in which the distribution of output neurons is consistent with the distribution of flow vectors $f_i = (x_i, y_i, dx_i, dy_i)$ in training samples, and the distribution of the neuron lines (internal nets) is consistent with the distribution of activities (STTs). Compared with the existing neural network structures that are used to learn patterns of trajectories, our network structure has smaller scale and faster learning speed, and is thus more effective.

4. INDEXING AND RETRIEVAL

The activity pattern analysis performs a classification of the activity samples and produces a set of AMs, which can be used as the indexing key for accessing the individual activity of a vehicle in video databases.

4.1. Semantic indexing

In our retrieval framework, an AM represents the common motion characters of a class of similar activities. The keywords that describe these common motion characters can be assigned to the AM manually in order to achieve a semantic level indexing. The individual activity automatically inherits all the semantic descriptors of the AM which it belongs to. Tables 1 and 2 give the data structure of *Activity Descriptor (AD)* and *Activity Model Descriptor (AMD)* which are stored in company with the original video clips in databases. The structure of semantic indexing and retrieval is illustrated in Figure 3.

When a new video clip is inputted to the databases, the activities of moving targets are extracted. According to the similarity metric of STT, these activities are classified to certain activity classes and sorted to the ACT_List of the corresponding AMs. It should be mentioned that the AMs can be adjusted dynamically when there are some activities cannot be classified to existing models. If the minimum distance of STT of a new activity to all models exceeds a threshold T_{normal}, the motion of this target should be seen as a *temporary abnormal activity* and represented by the Abnormal Activity Model (AAM). The HSOM is used periodically to classify the activities in AAM. Let N_i denote the number of activities in class *i* in AAM. If N_i exceeds a threshold in time *t*, the activities in class *i* should be seen as normal activities. A new AM is then established for class *i* and added to the collection of existing AMs. Semantic descriptors are also assigned to it. Sometimes two AMs will converge into one model. We examine all neighboring AMs after every hour's running to determine whether to merge them. With these strategies our system has the ability to capture the slow and long-term changes of target behaviors in a complex scene.

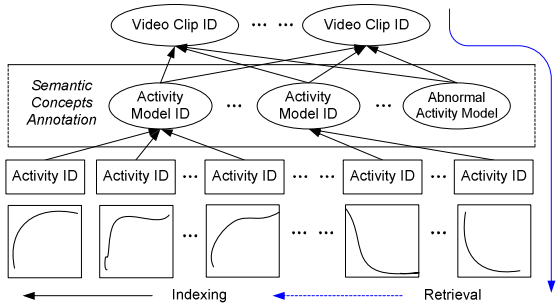

Figure 3. Structure of semantic indexing

Table 1. Activity Descriptor (AD)

Components	Value
ACT_ID	ID of Activity
VIDEO_ID	ID of Video Clip
Birth_Time	Frame number
Death_Time	Frame number
STT (Spatio-Temporal Trajectory)	$T_{ST} = \{f_1, f_2, ..., f_i, ..., f_{n-1}, f_n\}$ $f_i = (x_i, y_i, dx_i, dy_i)$
Obj_Color	Object color (R,G,B)
Obj_Size	Object size (height, width)

Table 2. Activity Model Descriptor (AMD)

Components	Value
AM_ID	ID of Activity Model
ACT_List	A List of Activities
STT (Spatio-Temporal Trajectory)	$T_{ST} = \{f_1, f_2, ..., f_i, ..., f_{n-1}, f_n\}$ $f_i = (x_i, y_i, dx_i, dy_i)$
Semantic_Descriptors	Keywords {turn left; low speed; north ahead; peccancy…}

4.2. Applicable query types

A. Query by keywords

A keyword-based query such as "show me a blue car running from south to north in a high speed" is performed in two stages. A keyword matching process is first used to find out the AM(s) whose descriptor matches the concepts of "from south to north" and "high speed". Then we retrieve the best matched activities in the ACT_List of this (these) AM(s) with color and size features. Multiple object queries are also supported by searching the best matching activities with a temporal restriction (e.g., simultaneity).

B. Query by sketch

It should be noted that there are activities and queries that cannot be expressed by keywords, such as, "Show me the video clips of a vehicle moving *in this way*." To solve this problem, a sketch-based query is also possible by matching the stored spatial trajectories and the sketch drawn by user with a time warping method.

5. EXPERIMENTAL RESULTS

The potential of the proposed video retrieval framework was verified on a variety of video sequences of natural traffic scenes. The RGB image size is fixed to 320×240 pixels. The experimental results of cluster tracking and activity pattern analysis are shown in figure 4, where the segmentation and tracking result in one frame is shown in (a) and all the trajectories extracted are shown in (b). 28 activity models (denoted by white lines) are generated and illustrated in (c). The correct rate of segmentation and tracking in the testing sequences is 97.4%. Our tracker runs at the speed of 5-10 fps with a moderate number of vehicles present in the scene.

(a) (b) (c)

Figure 4. Results of tracking and activity pattern analysis

(a) (b) (c)

Figure 5. Results of keyword-based queries

Figure 5 shows the retrieval results in semantic level. If a test query only contains the keyword of "turn left", the vehicle activities illustrated in (a), (b) and (c) are all best matches. When the test query contains more descriptive keywords such as "turned left and then ran towards west", (b) is eliminated from the candidates. Further, if the static characters of target ("a white car") are also specified, only (c) is retrieved as the final result.

6. CONCLUSION

In this paper, a semantic based retrieval framework for traffic video sequences has been proposed. We have developed a clustering based tracking algorithm to extract motion data of moving targets. Combining the activity pattern analysis and semantic indexing, our retrieval framework provides a semantic level query interface. Future research will focus on introducing user interactions and knowledge building process into our framework.

Acknowledgement: This work is partly supported by NSFC (Grant No. 60105002, 60373046, 60335010, 60121302), the National 863 High-Tech R&D Program of China (Grant No. 2002AA117010-11).

7. REFERENCES

[1] F. I. Bashir, A. A. Khokhar and D. Schonfeld, "Segmented Trajectory Based Indexing and Retrieval of Video Data", Proc. ICIP 2003

[2] W. Chen, S. F. Chang, "Motion Trajectory Matching of Video Objects", IS&T/SPIE, San Jose, CA, January 2000.

[3] Y. K. Jung, K. W. Lee, and Y. S. Ho, "Content-Based Event Retrieval using semantic scene interpretation for automated traffic surveillance", IEEE Trans. Intelligent Transportation Systems, vol. 2, no. 3, 2001.

[4] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic monitoring and accident detection at intersections", IEEE Trans. Intelligent Transportation Systems, 1(2), 2000

[5] J. M. Ferryman (ed.), Proceedings of the first IEEE workshop on performance evaluation of tracking and surveillance (PETS'2000), Grenoble, March, 2000

[6] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition", Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, 1998

[7] B. Maurin, O. Masoud, and N. Papanikolopoulos, "Monitoring Crowded Traffic Scenes", IEEE Conf. on Intelligent Transportation Systems, Singapore, 2002

[8] S. Eschrich, J. Ke, L. O. Hall, and D. B. Goldgof, "Fast accurate fuzzy clustering through data reduction", IEEE Trans. on Fuzzy Systems, vol. 11, no. 2, pp. 262-270, 2003.

[9] Weiming Hu, Dan Xie, Tieniu Tan, "A Hierarchical Self-organizing Approach for Learning the Patterns of Motion Trajectories", IEEE Trans. Neural Networks, Vol. 15, No. 1, January 2004.

[10] J. Joseph and J. LaViola, "Double exponential smoothing: an alternative to Kalman filter-based predictive tracking", Proc. Immersive Projection Technology and Virtual Environments 2003, ACM Press, pp. 199-206, 2003