Pattern Recognition Letters xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Semantic retrieval of events from indoor surveillance video databases

Chengcui Zhang *, Xin Chen, Liping Zhou, Wei-Bang Chen

Department of Computer and Information Sciences, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Semantic video retrieval
Relevance feedback
Spatiotemporal modeling

ABSTRACT

With the existence of "semantic gap" between the machine-readable low level features (e.g. visual features in terms of colors and textures) and high level human concepts, it is inherently hard for the machine to automatically identify and retrieve events from videos according to their semantics by merely reading pixels and frames. This paper proposes a human-centered framework for mining and retrieving events and applies it to indoor surveillance video databases. The goal is to locate video sequences containing events of interest to the user of the surveillance video database. This framework starts by tracking objects. Since surveillance videos cannot be easily segmented, the Common Appearance Intervals (CAIs) are used to segment videos, which have the flavor of shots in movies. The video segmentation provides an efficient indexing schema for the retrieval. The trajectories obtained are thus spatiotemporal in nature, based on which features are extracted for the construction of event models. In the retrieval phase, the database user interacts with the machine and provides "feedbacks" to the retrieval results. The proposed learning algorithm learns from the spatiotemporal data, the event model as well as the "feedbacks" and returns the refined results to the user. Specifically, the learning algorithm is a Coupled Hidden Markov Model (CHMM), which models the interactions of objects in CAIs and recognizes hidden patterns among them. This iterative learning and retrieval process contributes to the bridging of the "semantic gap", and the experimental results show the effectiveness of the proposed framework by demonstrating the increase of retrieval accuracy through iterations and comparing with other methods.

1. Introduction

In building an intelligent monitoring system, a large amount of surveillance videos are collected via surveillance cameras and stored in the database. Sequential browsing of such videos from the database is time consuming and tedious for the user, and thus cannot take full advantage of the rich information contained in the video data. The goal of this paper is to present a framework that incorporates various aspects of an intelligent surveillance system – object tracking, video segmentation and indexing, and human-centered automatic semantic retrieval of events, with the main focus on event retrieval.

In our previous work (Chen et al., 2003), we proposed an object segmentation and tracking algorithm for surveillance videos, from which object-level information such as the bounding boxes and the centroids can be obtained and stored in the database for future queries. For indexing purposes, videos can be segmented into shots. However, surveillance videos are composed of monotonously running frames. It is not feasible to apply existing shot detection methods, which can only detect shot boundaries by sharp scene changes such as in movies or sports videos. In L. Chen's CAI (Common Appearance Interval) model (Chen and Özsu, 2002), a video segmentation concept – Common Appearance Interval (CAI) is proposed, which has some flavor of a video shot in a movie. According to this concept, each video segment is endowed with some "semantic" meaning in terms of temporality and spatial relations. This concept is adopted in the proposed framework for surveillance video segmentation.

After trajectory tracking and segmentation, the event retrieval is performed. There are many researches on automatically detecting events from videos. Recently, the focus has been on applying stochastic signal models on this problem. Good success has been reported on using Hidden Markov models (Kettnaker, 2003; Petkovic and Jonker, 2001; Robertson and Reid, 2005). The choice of a HMM seems appropriate since it offers dynamic time warping and Bayesian semantics, which can be applied to recognize patterns in such spatiotemporal data as object trajectories. In the surveillance videos captured by security cameras, there is usually a large number of moving (e.g. a human) and static objects. To recognize events from them, we need to analyze the interactions among these objects. Du et al. (2006) proposed a Bayesian Network based approach to recognize interactions. In (Oliver et al., 2000; Brewer et al., 2006), the Coupled Hidden Markov Model (CHMM) is used for modeling human object interactions. The work in (Oliver

* Corresponding author. Fax: +1 205 934 5473.
E-mail addresses: zhang@cis.uab.edu (C. Zhang), chenxin@cis.uab.edu (X. Chen), zlp@cis.uab.edu (L. Zhou), wbc0522@cis.uab.edu (W.-B. Chen).

0167-8655/$ - see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.patrec.2009.05.004

et al., 2000; Brewer et al., 2006) analyzes the relative positions of two people in the video and models such macro interactions as two people "approach and meet". In our proposed framework, we will use a Coupled Hidden Markov Model (CHMM) to model interactions among objects in the video and to recognize normal and abnormal behaviors. For this purpose, the CHMM in the proposed work can model both macro and micro interactions between two people such as two people fighting. Different from other related work, the proposed work targets at events that have peculiar semantic meanings (e.g., "fighting"), which the users of the retrieval system are interested in. Therefore, only differentiating macro human interactions such as "meet and split", "meet and walk together", or "approach and meet" is not sufficient to meet the user's needs. In this paper, we further model the detailed spatiotemporal interactions (i.e., micro interactions) between two objects such as fighting. This will allow us to separate fighting from handshaking.

The proposed framework strives to meet the huge challenge of managing and retrieving video sequences according to their semantic meanings. This is challenging due to the fact that a machine does not have the equal ability in deducing semantic concepts from low level features as a human does. Such low level features can be as simple as pixel intensities of video frames, or more advanced ones such as textures of video frames. CHMM is a supervised statistical machine learning algorithm. By analyzing only the low level features, no matter how sophisticated the algorithm is, there is still a "semantic gap", which is a gap between the low level features and high level human concepts. Therefore, a human needs to provide some guidance to the learning algorithm (i.e. to teach the system). As in traditional machine learning, CHMM can accomplish this through constructing training set from the expert's prior knowledge. However, semantic video retrieval is different from a traditional data mining task. It is difficult to obtain a proper training set for each "relevant" class before the query, due to the scarcity of "relevant" samples and the uncertainty of users' interest. This is especially true in large video databases, where multiple "relevant" and "irrelevant" classes exist according to the different interests of different users (Nakazato et al., 2003), and the data in each "relevant' class may only constitute a very small portion of the entire database. For example, in "query-by-example" for video retrieval, the user may submit a query by giving a video example, which shows two people "meet and fight". However, without further information, it is uncertain what the user is really looking for – is he more interested in video sequences that contain "two people meet", or those that involve scenes of "two people fight"? In another word, it is not clear if the user is more interested in the macro interactions of the two objects or their micro interactions. If the user is interested in "two people meet" and does not care what they do after they meet, then video sequences that contain people "meet and fight", "meet and handshake", "meet and talk" are all relevant. On the other hand, if the user is interested in "fighting" scenes, then "fight and chase", "fight and run", "fight and fall down" are all relevant. Therefore, we need a customized search engine that can provide retrieval results according to individual users' preferences.

To solve this problem, we adopt a technique called "Relevance Feedback" (Rui et al., 1997) in the proposed semantic retrieval framework. When the framework returns the initial query results to the user according to some heuristics, the user can provide feedbacks. The learning algorithm then gathers training samples and learns from these feedbacks, and returns the refined results to the user. This process goes through several iterations until the user is satisfied with the results. In another word, with "Relevance Feedback", the database user takes the initiative to train the learning algorithm and is rewarded by a set of better results according to his/her own interest. This cannot be accomplished through traditional data mining where the training is limited by the expert's knowledge. The role of "Relevance Feedback" in the proposed framework is therefore two-fold: (1) to reduce the "semantic gap" by guiding the system and (2) to progressively gather training samples and customize the learning and retrieval process.

In summary, the proposed framework tracks and analyzes spatiotemporal data from surveillance videos and retrieves events according to individual users' query interests. It systematically incorporates techniques from multimedia processing, spatiotemporal modeling, multimedia data mining, and information retrieval. In particular, the retrieval system is "human-centered" in that the user can interact with the retrieval system and the learning algorithm via Relevance Feedback (RF). The technique of RF is incorporated, with which the user provides feedback and the learning algorithm learns from it by depressing the "irrelevant" scenes and promoting the "relevant" scenes. Instead of pre-defined "expert" knowledge, individual user's subjective view guides the learning process. Although RF is a commonly used technique in Content-based Image Retrieval, to our best knowledge, it has only been incorporated in video retrieval using key-frame based approaches (Calistru et al., 2007), where the important spatiotemporal information is lost; or it has been used on the video sequences (Munesawang and Guan, 2005) represented by a sequence of frames without object tracking information. Key-frame extraction is not applicable in surveillance videos. Our work is therefore among the first effort to incorporate RF into a non-key-frame based video retrieval environment that uses object trajectories as the target of analysis. The proposed framework is especially useful in mining and retrieving events of interest from large surveillance video databases, where only raw data is stored. By using users' feedbacks, human knowledge is incorporated into such a database. In this study, abnormal events in indoor surveillance videos are modeled and retrieved. Specifically, the events of two people "fighting" and the events of "robbing and chasing" are tested. However, the framework can be easily tailored to the recognition of other abnormal interactions, if the appropriate event models are built for each type of interactions. Experimental results show the effectiveness of the proposed framework for the detection of "fighting" and "robbing and chasing" events.

The major contribution of the proposed work lies in: (1) an integrated video retrieval system is proposed which incorporates all aspects of an intelligent indoor surveillance video retrieval system – starting from the preprocessing phase i.e., object segmentation and tracking with the ultimate goal being learning and retrieval abnormal behavior in the videos. (2) Relevance Feedback is used in the whole learning and retrieval process to provide training data, acquire knowledge through user feedback, and guide the retrieval process.

In the rest of the paper, a literature review is provided in Section 2. Section 3 briefly introduces a semantic object extraction and tracking algorithm and the video segmentation. Section 4 exemplifies the event modeling. Section 5 presents the design details of the learning and retrieval process. Section 6 provides the experimental results. Section 7 concludes the paper.

2. Related work

In our previous work (Chen and Zhang, 2006), a framework for traffic accident retrieval from traffic surveillance video databases is constructed. The proposed framework in this paper is significantly different from that. The major difference lies in the query target, i.e. the type of video events we want to retrieve. The objective of this study is to retrieve user-interested events in indoor surveillance videos rather than traffic surveillance videos. Data means everything. There is basically no single best framework that can accommodate the different requirements incurred by the retrieval of different types of videos. For example, events of interest in these two types of videos are quite different and therefore require different event modeling techniques, which, in turn, implies the development of different learning and retrieval mechanisms. In (Chen and Zhang, 2006), traffic accidents usually feature the abnormal behavior of at least one involved vehicle. Although an accident may involve more than one vehicle, it is sufficient to just analyze the sudden behavioral change of each individual vehicle and use it as an indication of accident. Analyzing the trajectories of each pair of vehicles would be unnecessary. If two vehicles are moving normally, we usually do not care if they are driving toward the same or opposite direction as along as they are on separate lanes. Storing these pair-wise interactions would be a waste of resource since they do not reflect much semantic meaning of interest. However, things are totally different for indoor surveillance video retrieval, since one person's behavior may affect another and often it is the interaction between two subjects that we are interested. For example, two people walking in the hall way may change their directions and walk toward each other after they see each other. In another word, the interactions in indoor surveillance videos carry more semantic meanings and have much more varieties than that of traffic surveillance videos. Therefore, the interactions are instead the main focus of this study for the event retrieval from indoor surveillance video databases.

2.1. Event detection in videos

Numerous works exist in detecting and recognizing events in videos. A lot of studies in this area are based on the generic visual properties of frames. For example, change of histograms between two consecutive frames may indicate the transition between two scenes, or events can be represented through analyzing the frame histograms (Lavee et al., 2005). These works do not utilize the spatiotemporal information by tracking each semantic object in the video. As tracking can provide more accurate and detailed information about object behaviors in a video sequence, there are also some research works that use object trajectories as their basis for analysis. For example, Medioni et al. (2001) proposed an event detection system by defining some scenarios based on spatial and temporal properties of object trajectories. Events were detected by simply comparing with the pre-defined scenario models. The work in (Ersoy et al., 2004) focused on the event modeling based on object trajectories in the videos. There is no learning process involved in (Medioni et al., 2001; Ersoy et al., 2004).

Many other works exploit stochastic methods in learning and recognizing video events. Bobick et al. (1998) proposed a Coupled Hidden Markov Model (CHMM) and the associated stochastic grammars for recognizing activities. Similarly in (Petkovic and Jonker, 2001), a rule-based approach was used to set up event models and HMM was adopted for automatic learning. In (Robertson and Reid, 2005), the authors combined HMM, Bayes networks, and belief propagation to understand human behavior. HMM was also used in (Kettnaker, 2003) to detect intrusions. Our proposed work adapts a CHMM for detecting abnormal human interactions in the indoor surveillance videos.

Self Organization Map (SOM) has also been used in some works for event detection from videos. ~~Naftel and Khalid~~ Naftel and Khalid (2006) proposed to use SOM in clustering and classifying object trajectories, hence detecting abnormal object behavior. A similar idea was developed in (Qu et al., 2005), with a Parallel Adaptive SOM being applied. In (Naftel and Khalid, 2006; Qu et al., 2005), the input nodes are the coefficients of the modeled trajectories which are not real time series data since there is no temporal relation among these nodes. Our proposed learning framework is different from (Naftel and Khalid, 2006; Qu et al., 2005) in that the input are time series sequences with temporal constraints.

Other learning tools also being adopted include Petri-net as in (Ghanem et al., 2004), which is also a spatiotemporal modeling technique. However, it is not suitable for modeling object interactions as desired in the event-based video retrieval. There are also some domain-specific video retrieval research such as in soccer (Gong et al., 1995) and tennis games (Petkovic and Jonker, 2001). However, none of them considered the spatiotemporal interactions of objects.

2.2. Relevance feedback

In order to overcome the obstacle posed by the semantic gap between high-level concepts and low-level features, the concept of relevance feedback (RF) associated with Content-based Image Retrieval (CBIR) is first proposed in (Rui et al., 1997). In the past few years, the RF approach to image retrieval has been an active research field. This powerful technique has proven successful in many application areas. In addition, various ad hoc parameter estimation techniques have been proposed for the RF approaches. Most RF techniques in CBIR are based on the most popular vector model (Buckley et al., 1995; Rui and Huang, 1999; Rui et al., 1998; Salton and McGill, 1983) used in information retrieval (Ishikawa et al., 1998). The RF technique estimates the user's ideal query by using relevant and irrelevant examples (training samples) provided by the user. The fundamental goal of these techniques is to estimate the ideal query parameters accurately and robustly.

Most previous RF research has been based on query point movement or query re-weighting techniques (Ishikawa et al., 1998). The essential idea of query point movement is quite straightforward. It represents an attempt to move the estimation of the "ideal query point" towards relevant sample points and away from irrelevant sample points specified by the user in accordance with his/her subjective judgments. Rocchio's formula (Rocchio, 1971) is frequently used to iteratively update the estimation of the "ideal query point". The re-weighting techniques, however, take the user's query as the fixed "ideal query point" and attempt to estimate the best similarity metrics by adjusting the weight associated with each low-level feature component (Aksoy and Haralick, 2000; Chang and Hsu, 1999; Rui et al., 1998). The essence of this idea is to assign larger weights to more important dimensions and smaller weights to less important ones.

As the Relevance Feedback techniques in the abovementioned work are applied to content-based image analysis, we adjust it to fit the needs of semantic video retrieval in this paper.

3. Video segmentation and object tracking

In this section, the preprocessing of video data is briefly introduced. The first step is video segmentation. In each video segment, object tracking is performed and the obtained trajectory sequences are stored in the database.

3.1. Video segmentation

In a surveillance video database where a large amount of raw data is stored, it is essential to provide an efficient indexing schema for fast access. If the raw video clip is stored as it is, sequential browsing is inevitable when one wants to search for a segment of video sequence from the clip. A natural solution is to perform video segmentation and store the video segments as well as their meta-data in the database, which can be accessed by the query scheme in a more convenient and speedy way. As we stated in Section 1, common shot detection techniques cannot be applied to surveillance videos since these videos do not have changing backgrounds or clear-cut boundaries between different scenes. In

Please cite this article in press as: Zhang, C., et al. Semantic retrieval of events from indoor surveillance video databases. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.05.004

Chen's CAI model (Chen and Özsu, 2002), a concept called Common Appearance Interval (CAI) is defined to model an interval where a certain set of objects appear in the frame together. We incorporate this concept into our framework.

Fig. 1 illustrates the video segmentation schema used in the proposed framework. Videos are segmented into CAIs that are represented by the directed edges in Fig. 1. The two nodes connected by edges represent the starting and the ending frame of a CAI. An example of starting and ending frames is shown for CAI_2. The objects (i.e. human) are outlined by colored bounding boxes. When the object outlined by the yellow bounding box enters the scene, it signifies the ending of CAI_2 and the starting of CAI_3. In another word, a new CAI is generated whenever a new object enters the scene or an existing object leaves the scene. In this way, videos are indexed and stored in the database.

3.2. Automatic object tracking

With the segmented surveillance videos stored in the database, the next step is to perform object tracking on these videos. The propose work in this paper focuses on high-level vision and assumes that trajectories already exist. In the experiment, we use our previous work (Chen et al., 2003) to perform automatic tracking, in which an unsupervised segmentation method called the Simultaneous Partition and Class Parameter Estimation (SPCPE) algorithm, coupled with a background learning and subtraction method, is used to identify the objects in a video sequence. The technique of background learning and subtraction is used to enhance the basic SPCPE algorithm in order to better identify objects in surveillance videos. With this algorithm, we can obtain blobs of objects in each frame. We can further acquire the Minimal Bounding Boxes of the objects as well as the coordinates of each object blob's centroid, which are then used for tracking the positions of objects across video frames. The framework in (Chen et al., 2003) also has the ability to track moving objects (blobs) within successive video frames. By distinguishing the static objects from mobile objects in the frame, tracking information can be used to determine the trails/trajectories of objects.

With this framework, lots of spatiotemporal data is generated such as trajectories of moving objects. This provides a basis for video event mining and retrieval. In this paper, suitable spatiotemporal models for video data are built to further organize, index and retrieve these information.

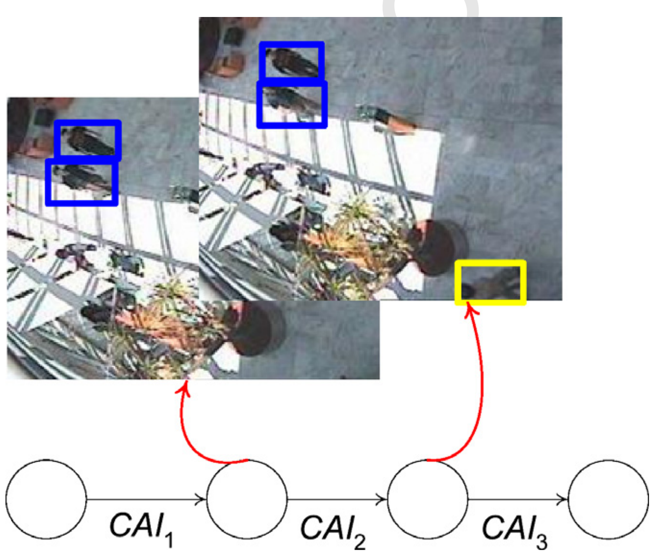

Fig. 1. Video segmentation with CAIs.

4. Event modeling

Various properties of objects along their trajectories can be extracted to build the models for specific event types. In this study, a spatiotemporal model is built for detecting abnormal behaviors in indoor surveillance videos. In the experiment, we used CAVIAR videos (CAVIAR: http://homepages.inf.ed.ac.uk/rbf/CAVIAR) taken in the lobby of a building in France and the videos we took in the lobby of Campbell Hall of University of Alabama at Birmingham (UAB).

After the video segmentation and object tracking, the spatiotemporal information of moving and static objects is obtained. In each CAI (Common Appearance Interval), pairs of object trajectories are studied, which will be referred to as Sequence Pair (SP) in this paper. It is observed that abnormal human interactions often involve the behavior of at least two people. By analyzing each SP, the events involve multiple people can also be detected. Therefore, the targets of learning are the interactive behavioral patterns of the two objects' trajectories in a SP. The focus of this study is on the interactions among people appearing in the video. For this purpose, some features of human behaviors are extracted from pairs of human trajectories. There are a lot of existing work on object tracking and interaction modeling (Sato and Aggarwal, 2004; Shi et al., 2006; Han et al., 2004; Efros et al., 2003). However, the emphasis of this paper is not to propose a sophisticated feature extraction algorithm for interaction modeling. Instead, the emphasis is on improving the retrieval accuracy through RF. Therefore, event modeling in the proposed work is not as sophisticated as those used in the above mentioned work. It largely involves the use of heuristics. The goal is to test that, based on the same event model, whether the proposed learning and retrieval system can effectively learn users' intent and improve the retrieval accuracy.

Normal human interactions include primitive ones such as "meet", "follow", and "walk together". Complex ones such as "meet and split" and "follow and reach and walk together" are usually composed of primitive interactions. For these macro human interactions, three properties are extracted: (1) *dist* – distances between two objects in the SP; (2) θ – degree of alignment of two objects, i.e., the signed angle between the motion vectors of two objects (illustrated in Fig. 2; $\vec{M}_1$ and $\vec{M}_2$ are the motion vectors of two objects at time t); (3) *vdiff* – change of velocities of the two objects between two consecutive frames.

In order to detect abnormal human interactions, another factor that needs to be taken into consideration is the magnitude of motion change of each object. This can be analyzed by the Optical Flow i.e., the pixel motions in the bounding boxes of objects. The basic idea is to find out the differences between one point in the current frame and the corresponding point it moves to in the next frame. Optical Flow can be used to describe the velocity and the direction of the motions in bounding boxes.

As mentioned in Section 1, to use Relevance Feedback, some heuristics need to be established in order to process the initial query. We observe that most of the human interactions in the

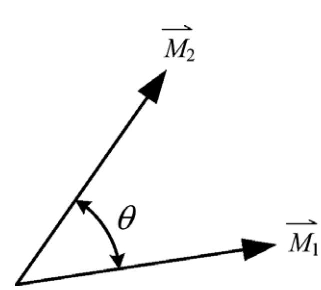

Fig. 2. The degree of alignment.

Please cite this article in press as: Zhang, C., et al. Semantic retrieval of events from indoor surveillance video databases. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.05.004

testing videos are normal such as two people meet with each other and talk. Some abnormal behaviors include two people "meeting and fighting" with each other or "robbing and chasing". For these abnormal human interactions, we build a heuristic model based on the observation that the sudden change of velocity and direction, the short distances between two objects, and the sharp change of motion energy may signify an abnormal human interaction. Therefore, at time t, the property vector of an object (human) can be represented as $\alpha_t = [vdiff_t, \theta_t, 1/dist_t, M_t]$. A series of such vectors $\alpha = [\alpha_1, \ldots, \alpha_n]$ represent the entire trajectory of an object in a SP. Each SP is therefore composed of two object sequences represented by the two series of property vectors $\alpha = [\alpha_1, \ldots, \alpha_n]$ and $\alpha' = [\alpha'_1, \ldots, \alpha'_n]$.

Although "meeting and fighting" and "robbing and chasing" are two different events, they belong to the same category. Both of them involve intense motion change when two objects are close. The difference is that "meeting and fighting" involve two people walk toward each other in normal speed and then are both engaged in the dramatic motion change i.e., "fighting". However, "robbing and chasing" involve one person's dramatic motion change i.e., "suddenly run toward another person and quickly grab that person's belongings" then both persons intense motion change i.e., "run fast toward the same direction." Therefore, the same event model can be applied to both events. The experimental results show that the proposed retrieval system can gradually learns the intent of the user through RF.

5. Event learning and retrieval

5.1. Coupled Hidden Markov Model

Hidden Markov Model (HMM) is a stochastic model that characterizes real-world signals. It is known for its ability to model processes that have structure in time since it automatically performs dynamic time warping. The HMM considers a system as being in one of the limited distinct states at any time. These states are connected by the transitions with the associated probabilities. These transitions convey a clear Bayesian semantics.

It is not uncommon that a real-world signal has multiple channels. In our application, if we model the trajectory of an object with the four-variant ($\alpha_t = [vdiff_t, \theta_t, 1/dist_t, M_t]$) sequence, each sequence (process) then has four channels. HMM can accommodate this by formulating multivariate p.d.f's on the outputs. However, this cannot meet our need for modeling multiple processes, since interactions between two people involve two multivariate processes. Therefore, the classic HMM structure is not suitable for this application. An extension of HMM – Coupled Hidden Markov Model (CHMM) (Brand, 1996), which has compositional states, is seemingly a better choice.

Fig. 3 shows the tree structure of a CHMM rolled out in time. A CHMM is appropriate for processes that influence each other asymmetrically and possibly causally. We use a two-chain CHMM for modeling the interactions between pairs of people in the surveillance video. The posterior of a two-chain CHMM is given below:

$$P(S|O) = \frac{P_{S_1}P_{O_1}P_{S'_1}P_{O'_1}}{P(O)} \prod_{i=2}^{l} P_{S_i|S_{i-1}} P_{S'_i|S'_{i-1}} P_{S'_i|S_{i-1}} P_{S_i|S'_{i-1}} P_{O_i} P_{O'_i}, \tag{1}$$

where s_i, s'_i, o_i and o'_i are the i^{th} state variables and observation outputs on the two chains of the CHMM. l is the length of the observation and thus the length of the state variable sequence. Brand (1996) solved this problem by N-head dynamic programming. For a two-chain CHMM, the associated dynamic programming problem is in principle $O(MN^4)$. However, by relaxing the assumption that every transition must be visited, Brand's algorithm (Brand, 1996) is shown to be $O(4MN^2)$.

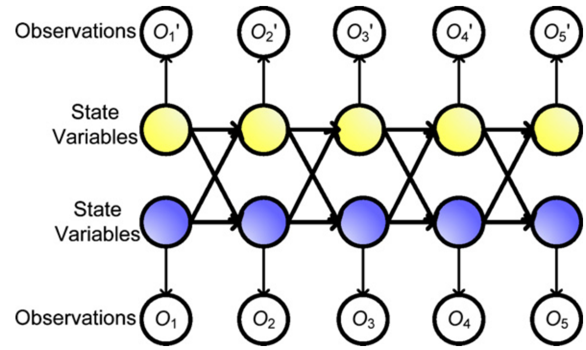

Fig. 3. CHMM rolled out in time.

When modeling the human interactions in our application, we have each chain model the behavior of one person. The influences of each person to the other are reflected in the cross transitions between two chains. Therefore, both the individual behaviors and the interactions between two persons are modeled in a single system.

5.2. Interactive event learning and retrieval

Prior to the learning and retrieval, pairs of human trajectories are collected. The trajectories are time series data in that their values change over time. The analysis of time series data shall not only focus on each individual data point separately but also look into the continuity within such kind of data. In time series models, there is a commonly used method called sliding window, which slides over the whole set of time series data to extract consecutive yet overlapped data sequences i.e. windows. This idea is also adopted in this framework. Fig. 4 shows an example of sliding window for time series data. In this example, a set of 6-tuple sequences is extracted from time series data by sliding a window of size 6 one step a time along the time axis t.

In the initial query, the user specifies an event of interest as the query target. The ultimate goal is to retrieve those video sequences that contain similar events. At this point, no relevance feedback information is provided by the user. Therefore, no training sample set is available to learn the pattern of user interested events. In order to provide an initial set of video sequences for the user to provide relevance feedback, for each object trajectory segment in the database, we calculate its relevance (or similarity score) to the target query event according to some event-specific search heuristics.

Suppose in one CAI, there are nTrajectory Pairs (TPs) and mSequence Pairs (SPs) of length $l$extracted from each TP by window sliding, with l being the window size. In the initial retrieval for "fighting" events, for each SP, at each time point there are two corresponding feature vectors $\alpha_t = [vdiff_t, \theta_t, 1/dist_t, M_t]$ and $\alpha'_t = [vdiff'_t, \theta'_t, 1/dist_t, M'_t]$. The relevance score of an SP is thus $\max_{t=1}^{l}(score(\alpha_t, \alpha'_t))$, where $score(\alpha_t, \alpha'_t) = \sqrt{(1/dist_t)^2 + vdiff_t^2 + vdiff_t'^2 + M_t^2 + M_t'^2}$. $\langle vdiff_t, vdiff'_t \rangle$ are the velocity changes and $\langle M_t, M'_t \rangle$ are the two object motion energies in that SP at time t, respectively. The degree of alignment, i.e.,θ_t

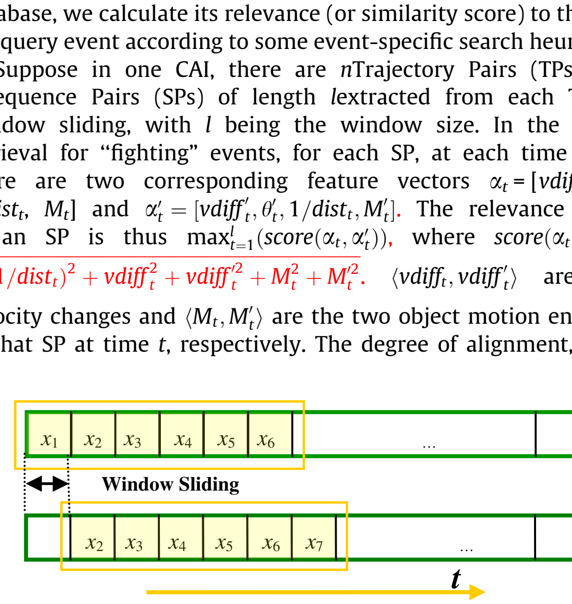

Fig. 4. An example of sliding window.

Please cite this article in press as: Zhang, C., et al. Semantic retrieval of events from indoor surveillance video databases. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.05.004

is not used in this computation since it mainly models interactions, which cannot be directly combined with individual behavioral features such as velocity changes. However, this feature will be used in CHMM as a separate channel for each interacting process. The retrieval results are returned in the descending order of each SP's relevance score. It is assumed that a big velocity change, a drastic change of motion, and a short distance between two people are indications for possible abnormal interactions such as fighting.

After the initial query, a certain number of SPs are presented to the user in the form of video sequences. In our experiment, the top 20 video sequences are returned for the user's feedback. The user identifies a returned sequence as "relevant" if it contains the event of his/her interest, or 'irrelevant' if otherwise. With this information at hand, a set of training samples can be collected. Each training sample is in the form of $\langle[\alpha_1, \alpha_2, \ldots, \alpha_l], [\alpha'_1, \alpha'_2, \ldots, \alpha'_l]\rangle$. α_i's and α'_i's are the feature vectors of two objects at consecutive time points. These training samples are then fed into the learning algorithm, which learns the best parameters for the CHMM. In the following iterations, these parameters are further refined with new training samples collected from users' feedbacks. In this iterative process, the user's query interest is obtained as user feedbacks and transferred to the learning algorithm, and the refined results are returned to the user for the subsequent run of the retrieval-feedback. It is shown in our experiment that, with this interactive learning technique, the retrieval results can be improved iteratively.

6. Experiments

6.1. System overview

The main functional units of the system include:

1. Preprocessing: The raw video is analyzed by segmenting videos into CAIs and tracking semantic objects (human) in them.
2. Trajectory modeling: In each CAI, trajectories are further modeled with the sliding window technique.
3. Event modeling: In this study, an event model for two people fighting is built, and the feature vectors of human objects at consecutive time points are extracted.
4. Initial retrieval: When the user submits a query, the system performs an initial query based on some heuristics specific to the event type, and returns the initial retrieval results to the user.
5. Interactive learning and retrieval: The user responds to the retrieval results by giving his/her feedbacks. The learning mechanism in the system learns from these feedbacks and refines the retrieval results in the next iteration. The whole process goes through several iterations until a satisfactory result is obtained.

Two sets of testing videos are used in the experiments. One is from the CAVIAR (CAVIAR: http://homepages.inf.ed.ac.uk/rbf/CAVIAR) videos taken in the lobby of a building. Another set is collected at the lobby of the Campbell Hall at the University of Alabama at Birmingham (UAB). Fig. 5 shows the interface for the user to provide feedback information. The user specifies an event of interest as the query target. Ideally, there should be several event categories for the user to choose, e.g., "meet and talk", "chasing", etc. Since only "fighting" events are modeled and tested in this paper, the interface does not show these query options to the user. The top 20 video sequences are returned to the user at each iteration. The user can play the retrieved video sequences by clicking the 'play' button and view the trajectories of problematic people objects. A retrieved example in CAVIAR videos is provided in Fig. 6. An example of two people "meet, talk, and walk way together with each other" in UAB videos is shown in Fig. 7. If the user thinks the marked trajectories in a particular sequence are what he is looking for, that video sequence will be selected and marked as 'relevant'. As shown in Figs. 5 and 6 sequences are labeled "relevant" in a query for the event of two people fighting.

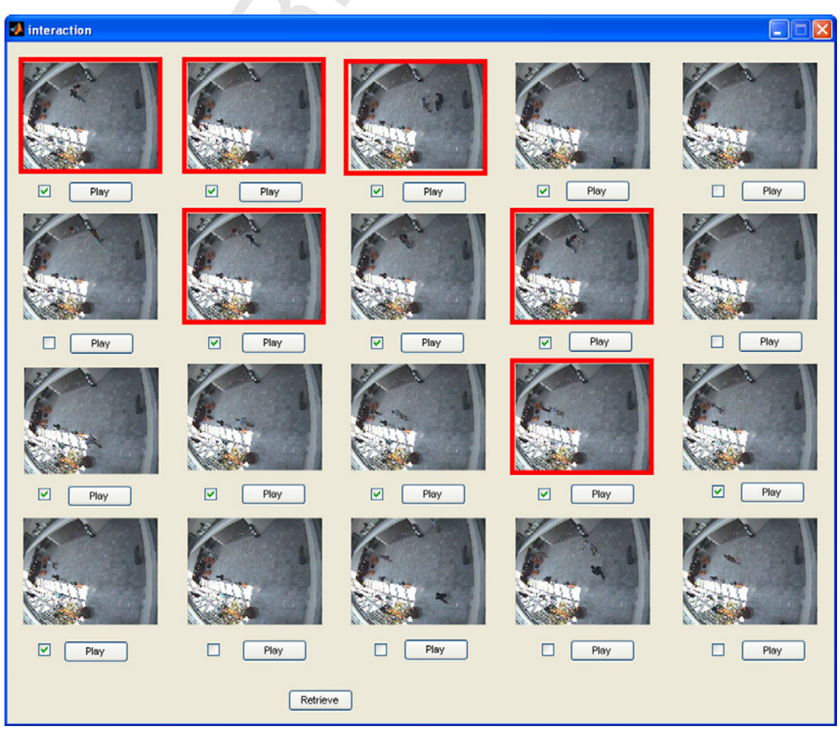

Fig. 5. The user interaction interface.

Fig. 6. An example of two people "meet and fight with each other" (CAVIAR).

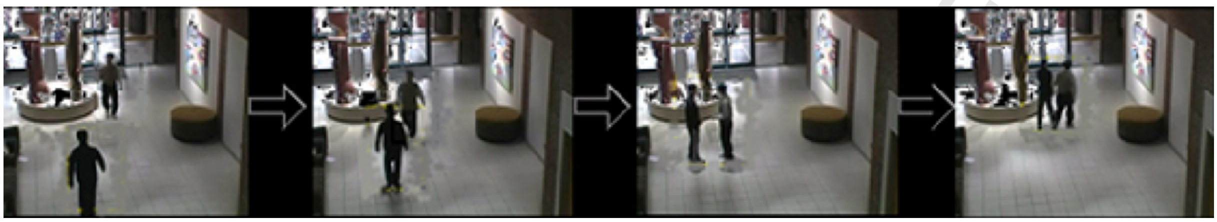

Fig. 7. An example of two people "meet, talk, and walk away together" (UAB).

6.2. System performance

In this study, abnormal human interactions are modeled for indoor surveillance video retrieval. In particular, the retrieval of "meeting and fighting" events and "robbing and chasing" events are tested with the proposed framework. For CAVIAR video sets, ten video clips containing human interactions are extracted. For the "UAB" video, 28 video segments containing human interactions are obtained. The majority of people interactions in these videos are normal such as "meet and walk together", "meet, walk together and split", "meet, split, and a third guy appears", "split", and "a crowd meet and split". These normal interactions are similar to the "meeting and fighting" or "robbing and chasing" interactions since all of them involve "two people get together and/or split". The slight difference lies in the drastic change of behaviors of individual people. Therefore, although they are similar in terms of macro interactions, we are able to differentiate them in terms of micro interactions. This is accomplished through the spatiotemporal modeling (i.e., extracting and indexing features) of "meeting and fighting" and "robbing and chasing" events. Besides normal human interactions, the CAVIAR videos contain only "meeting and fighting" events which "UAB" videos contain both "meeting and fighting" and "robbing and chasing" events. These video clips were taken at a frame rate of 25 frms/sec. The window size is 100, i.e. 100 points (frames) in a window. With a step size of 20 for window sliding, there are altogether 299 sequences (100 frames each) from the CAVIAR videos and 331 sequences from the "UAB" videos stored in the database. After the initial retrieval, the first training set obtained via user-provided feedback is used to determine the number of states in CHMM. Through ten-fold cross validation, the number of states is determined to be 3 in our case.

Four rounds of user relevance feedback are performed - Initial (no feedback), First, Second, and Third. In each iteration, the top 20 video sequences are returned to the user. To evaluate the retrieval performance of the proposed video retrieval system, we use the measure of accuracy for such purpose. In particular, the accuracy rates within different scopes, i.e. the percentage of relevant video sequences within the top 5, 10, 15 and 20 returned video sequences are calculated. In the area of Content-Based Image Retrieval (CBIR), the measure of accuracy has been widely used instead of precision-recall for performance evaluation and comparison. Such examples can be easily found in most of the recent works in CBIR (Su et al., 2003). The reasons for using accuracy for multimedia data retrieval lie in two aspects. (1) Multimedia retrieval systems are designed to return only a few relevant images/videos, where the user only browses the top few images; thus, precision is emphasized over recall. (2) As the size of image database grows, manually separating the collection into relevant and irrelevant sets becomes infeasible, which in turn prevents the accurate evaluation of recall. Although we do not have the ground-truth to calculate precision and recall, we can give a rough estimate of that by using the number of video clips that contain fighting. In CAVIAR videos, there are 10 video clips with only 4 clips containing fighting events. In UAB videos, there are 28 video clips with 15 of them containing fighting events. It is also worth mentioning that the framework retrieves sequence pairs which are extracted by sliding a window inside a CAI. In total we have 630 such sequence pairs with each of them containing two trajectory sequences of 100 frames. Even in the video clips that have fighting events, among all the sequence pairs extracted from the video, there are still some that do not contain fighting events. Specifically, for each video clip that contains fighting events, our calculation shows that, on average, approximately 50% of the video content actually contains fighting events. Therefore, a rough estimate of the fighting sequence pairs in the two test databases is 31 and 124, respectively.

In order to test the robustness of the proposed event model, we compare the features currently being used in this study (represented as feature set F_1) with another set of features (represented as feature set F_2) proposed in Ribeiro and Santos-Victor's work for human activity modeling and feature selection (Ribeiro and Santos-Victor, 2005). This set of features (F_2) includes speed/velocity ratio, motion energy, and relative velocity. The velocity ratio is the ratio between the average speed and the norm of the average velocity (Ribeiro and Santos-Victor, 2005) and is used to describe how irregular the motion actually is. When the value approaches 1, it means that the object always moves in the same direction along a straight line. When the value is close to 0, the object moves irregularly in various directions. The change of relative velocity between two objects is used to signify the interaction pattern of two objects. In general, a normal interaction between two objects tends to produce a constant relative speed. For example, when two persons walk toward each other and meet together, the relative speed of the two persons has little change. In other words, the variance of the relative speed is close to 0. In contrast, during an abnormal interaction, such as two people "meeting and fighting" or "robbing and chasing", the relative speed is more likely to change over time. One example is the "robbing and chasing" event. When the "robbing and chasing" event is happening, the relative speed of two

Please cite this article in press as: Zhang, C., et al. Semantic retrieval of events from indoor surveillance video databases. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.05.004

Table 1
Retrieval results comparison between two different sets of features.

	Initial		First		Second		Third	
	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2
CAVIAR	20%	25%	65%	60%	80%	65%	90%	75%
UAB Meet and Fight	25%	30%	60%	65%	75%	80%	85%	90%
UAB Rob and Chase	50%	30%	50%	40%	80%	45%	90%	60%
Average	31.67%	28.33%	58.33%	55%	78.33%	63.33%	88.33%	75%

objects will increase more rapidly compared to other normal interactions. It is proved through experiments that all these features have good performance in classifying walking, fighting, and running events (Ribeiro and Santos-Victor, 2005). We test both sets of features in our retrieval framework and present their retrieval accuracies (the percent of relevant video sequences among the top 20 retrieved sequences) for the three video sets in Table 1. F_1 represents the features used in the proposed framework. F_2 is the set of features for comparison. For "UAB Meet and Fight" video, F_2 performs better than F_1. However, the average performance of F_1 is better than F_2. It is worth mentioning that in both cases (F_1 and F_2), the retrieval accuracy increases across all iterations monotonically, indicating the robustness of the proposed framework.

From Figs. 8 and 9, we can see that the retrieval accuracies of "meeting and fighting" events increase steadily across multiple iterations with the incorporation of the user's feedback. For example, in the second iteration, the total accuracy for CAVIAR videos has already reached 80% i.e. 16 out of 20 returned sequences are regarded as "relevant" by the user. If the user is still not satisfied with the results and wants to continue the process, he/she is able to find 18 relevant sequences after the third iteration, making the total retrieval accuracy 90%. Notice that after the second iteration, the accuracy among the top 15 returned results has reached 100%. For the UAB videos, its accuracy has also reached 75% after the second iteration and the overall retrieval accuracy increases to 85% in the third iteration. Fig. 10 illustrates the retrieval accuracies of "robbing and chasing" events in "UAB" videos. The accuracy increases across iterations and reaches 90% in the third iteration.

In our experimental design, the proposed framework is compared with the HMM and the traditional weighted relevance feedback method, using different feature sets (F_1 and F_2, respectively). For the HMM, each SP is represented by a series of seven-feature vectors $\langle 1/dist_t, \theta_t, \theta'_t, vdiff_t, vdiff'_t, M_t, M'_t \rangle$. It models each SP as a 7-channel sequence instead of two multi-channel sequences as in CHMM.

In the weighted relevance feedback method, each feature component in the feature vector α_t has its associated weight. The initial round of retrieval is the same as that of the proposed framework. That is to say, the initial weights of the features $\langle 1/dist_t, vdiff_t,$

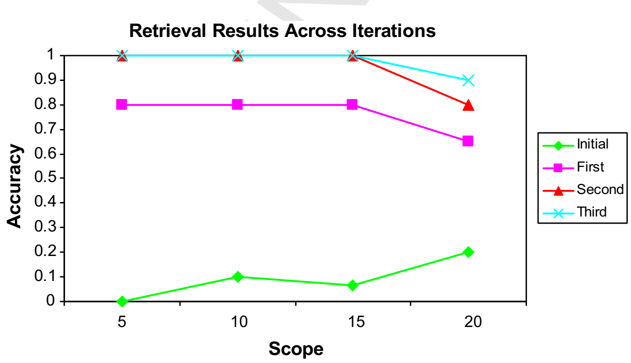

Fig. 8. Retrieval accuracies of "meeting and fighting" events across four iterations for CAVIAR videos.

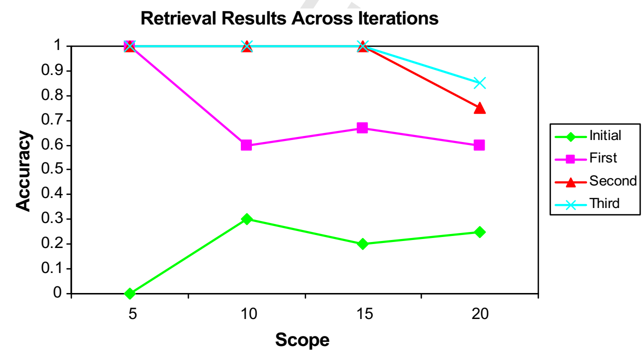

Fig. 9. Retrieval accuracies of "meeting and fighting" events across four iterations for UAB videos.

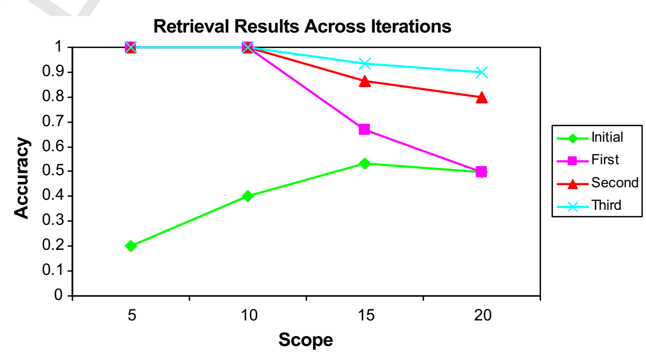

Fig. 10. Retrieval accuracies of "robbing and chasing" events across four iterations for UAB videos.

$vdiff'_t, M_t, M'_t \rangle$ are all 1s and the L2 norm of these features is computed as the relevance score. θ_t and θ'_t are ignored for the reason aforementioned. With the user's relevance feedback, the feature vectors of all relevant SPs are gathered. The inverse of the standard deviation of each feature is computed and used as the updated weight for this feature in the next round. In our experiment, we found that some large weights can introduce bias in computing relevance scores and hence affect the retrieval accuracy. Therefore, it is necessary to normalize these weights. We first tried to linearly normalize these weights to the range of [0 1]. However, the problem with this method is that a weight of zero will always eliminate the corresponding feature. We then tried another method i.e., the percentage of each weight among the total weight is used as its normalized weight. In our experiment, it is found that the latter outperforms both the linear normalization and no normalization at all.

Figs. 11 and 12 compare the retrieval accuracies of "meeting and fighting" events among the top 20 returned video sequences across four iterations. Fig. 13 compares the retrieval accuracies of "robbing and chasing" events among the top 20 returned video sequences across four iterations. "RF" is the weighted relevance feedback method aforementioned. "HMM" is the Hidden Markov

Please cite this article in press as: Zhang, C., et al. Semantic retrieval of events from indoor surveillance video databases. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.05.004

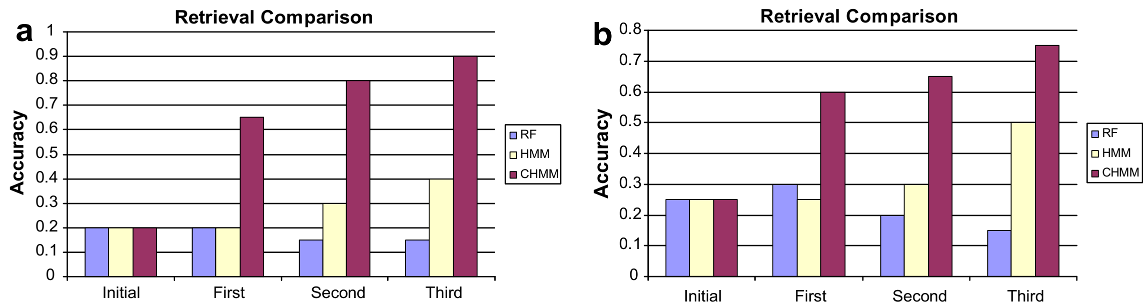

Fig. 11. Compare the accuracies of "meeting and fighting" events across iterations for CAVIAR videos: (a) the result of using F_1; (b) the result of using F_2.

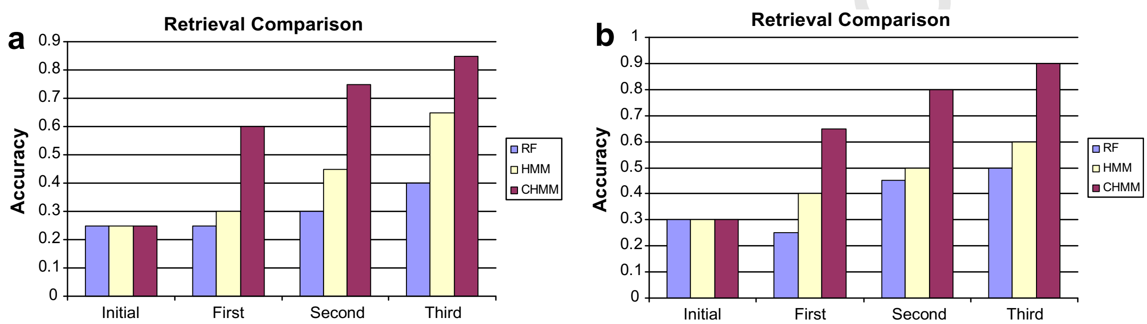

Fig. 12. Compare the accuracies of "meeting and fighting" events across iterations for UAB videos: (a) the result of using F_1; (b) the result of using F_2.

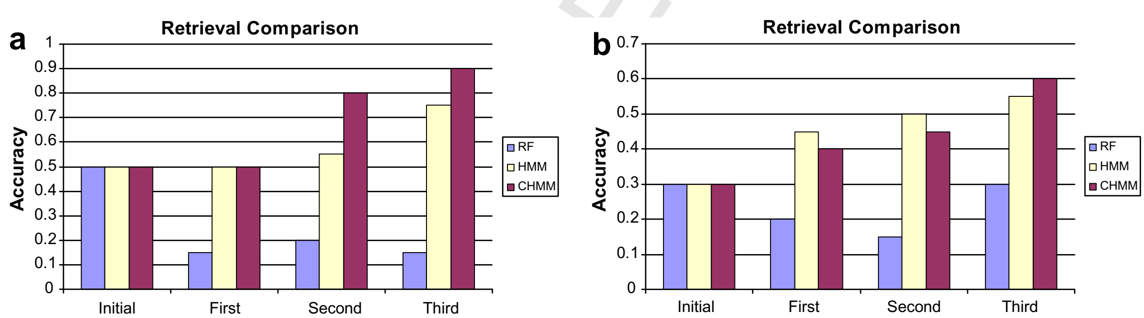

Fig. 13. Compare the accuracies of "robbing and chasing" events across iterations for UAB videos: (a) the result of using F_1; (b) the result of using F_2.

Model, which has only one chain. "CHMM" is the proposed framework. It is observed that the overall performance of the proposed framework is better than that of the weighted relevance feedback as well as the HMM based method for both video sets. Although the accuracies of "CHMM" using F_2 (Fig. 13b) in the initial, first, and second iterations are not as good as "HMM", "CHMM" outperforms "HMM" in the third iteration. This is due to the fact that the heuristic used in the initial retrieval does not consider interactions between two objects. Instead, the features of two objects are combined into one single feature vector such that a SP is regarded as one multiple-channel sequence in both 'RF' and 'HMM' methods. Since the initial retrieval for weighted RF, HMM and CHMM use the same heuristic, by comparing the results in the subsequent iterations of users' relevance feedback, it is clear that CHMM is more effective in recognizing patterns of interactions than either the weighted RF or the HMM. In another word, although the HMM and the classic RF methods (feature re-weighting) can model single signal well (Kettnaker, 2003; Petkovic and Jonker, 2001; Robertson and Reid, 2005; Rui et al., 1997), they are not suitable for modeling interactions of two signals.

A typical kind of false positive for 'fighting' is when two people are running, therefore with dramatic motion change. The event model for fighting has 'distance' factor in it. But it does not regulate that 'fighting' happens when two people have 'short distance' and at the same time 'big motion change'. The above comparison results show that it is through the study of the interaction process of two people with CHMM that these false positives can be reduced.

7. Conclusions and future work

In this paper, a human-centered semantic video retrieval platform is proposed. Given a set of raw videos, the semantic objects are tracked and the corresponding trajectories are modeled and stored in the database. Some spatiotemporal event models are then constructed. The goal is to automatically detect and retrieve abnormal human interactions in indoor surveillance videos. For the learning and retrieval, the Couple Hidden Markov Model (CHMM) is adapted to fit the specific needs of event identification and retrieval for indoor surveillance video data. The platform shows its effectiveness as demonstrated by our experimental results on two set of indoor surveillance videos. In the learning and retrieval phase, with the top returned video sequences in each iteration, the user provides feedback to the relevance of each video sequence.

The learning algorithm then refines the retrieval results with the user's feedbacks. This platform successfully incorporates the Relevance Feedback technique in retrieving events from video data, which is a well studied topic in Content-Based Image Retrieval but needs significant extensions (e.g. the modeling and incorporation of spatiotemporal characteristics) when applied to video data retrieval.

In the future work, more general event models will be constructed and tested with the proposed platform. More videos containing other types of events will be collected to test the framework. With users' feedbacks stored in the database log, we will also equip the system with the ability for long-term learning. In this way, future queries can benefit from the knowledge gathered from previous queries.

Acknowledgement

The work of Chengcui Zhang was supported in part by NSF DBI-0649894 and SBE-0245090 and the UAB ADVANCE program of the Office for the Advancement of Women in Science and Engineering.

References

Aksoy, S., Haralick, R.M., 2000. A weighted distance approach to relevance feedback. In: Proc. Internat. Conf. on Pattern Recognition.
Bobick, A.F., Pentland, A.P., Poggio, T., 1998. VSAM at the MIT Media Laboratory and CBCL: Learning and understanding action in video imagery PI report 1998. In: Proc. DARPA Image Understanding Workshop.
Brand, M., 1996. Coupled hidden markov models for modeling interacting processes. Neural Comput..
Brewer, N., Liu, N., Vel, O.D., Caelli, T., 2006. Using coupled hidden Markov models to model suspect interactions in digital forensic analysis. In: Proc. Internat. Workshop on Integrating AI and Data Mining (AIDM'06).
Buckley, C., Singhal, A., Miltra, M., 1995. New retrieval approaches using SMART:TREC4. In: Proc. Text Retrieval Conf., Sponsored by National Institute of Standard and Technology and Advanced Research Projects Agency.
Calistru, C., Ribeiro, C., David, G., Rodrigues, I., Laboreiro, G., 2007. INESC, Porto at TRECVID 2007: Automatic and Interactive Video Search. TRECVID 2007.
CAVIAR: Context Aware Vision using Image-based Active Recognition (2004, January 10). Available: <http://homepages.inf.ed.ac.uk/rbf/CAVIAR>.
Chang, C.-H., Hsu, C.-C., 1999. Enabling concept-based relevance feedback for information retrieval on the WWW. IEEE Trans. Knowledge Data Eng. 11 (4), 595–609.
Chen, L., Özsu, M.T., 2002. Modeling of video objects in a video database. In: Proc. IEEE Internat. Conf. on Multimedia, Lausanne, Switzerland.
Chen, X., Zhang, C., 2006. An interactive semantic video mining and retrieval platform – application in transportation surveillance video for incident detection. In: Proc. IEEE Internat. Conf. on Data Mining, Hong Kong, China.
Chen, S.-C., Shyu, M.-L., Peeta, S., Zhang, C., 2003. Learning-based spatio-temporal vehicle tracking and indexing for transportation multimedia database system. IEEE Trans. Intell. Trans. Systems 4 (3), 154–167.
Du, Y., Chen, F., Xu, W., Li, Y., 2006. Recognizing interaction activities using dynamic bayesian network. In: Proc. 18th Internat. Conf. on Pattern Recognition (ICPR'06), vol. 1, pp. 618–621.
Efros, A.A., Berg, A.C., Mori, G., Malik, J., 2003. Recognizing action in a distance. ICCV, 726–733.
Ersoy, I., Bunyak, F., Subramanya, S.R., 2004. A framework for trajectory based visual event retrieval. In: Proc. Internat. Conf. on Information Technology: Coding and Computing (ITCC'04).
Ghanem, N., DeMenthon, D., Doermann, D., Davis, L., 2004. Representation and recognition of events in surveillance video using petri nets. In: Proc. Computer Vision and Pattern Recognition Workshop (CVPRW'04).
Gong, Y., Sin, L.T., Chuan, C.H., Zhang, H.-J., Sakauchi, M., 1995. Automatic parsing of TV soccer programs. In: Proc. IEEE Internat. Conf. on Multimedia Computing and Systems, Washington, DC.
Han, M., Xu, W., Tao, H., Gong, Y., 2004. An algorithm for multiple object trajectory tracking. CVPR 1, 864–871.
Ishikawa, Y., Subramanya, R., Faloutsos, C., 1998. Mindreader: Query databases through multiple examples. In: Proc. 24th Internat. Conf. on Very Large Databases.
Kettnaker, V.M., 2003. Time-dependent HMMs for visual intrusion detection. In: Proc. Computer Vision and Pattern Recognition Workshop.
Lavee, G., Khan, L., Thuraisingham, B., 2005. A framework for a video analysis tool for suspicious event detection. In: Proc. Workshop on Multimedia Data Mining in Conjunction with KDD2005, Chicago, IL, USA.
Medioni, G., Cohen, I., Bremond, F., Hongeng, S., Nevatia, R., 2001. Event detection and analysis from video streams. IEEE Trans. Pattern Anal. Machine Intell. 23 (8), 873–879.
Munesawang, P., Guan, L., 2005. Adaptive video indexing and automatic/semi-automatic relevance feedback. IEEE Trans. Circuits Systems Video Technol. 15 (8), 1032–1046.
Naftel, A., Khalid, S., 2006. Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space. Multimedia Systems 12 (1), 227–238.
Nakazato, M., Dagli, C., Huang, T.S., 2003. Evaluating group-based relevance feedback for content-based image retrieval. In: Proc. IEEE Internat. Conf. on Image Processing (ICIP'03), Spain.
Oliver, N.M., Rosario, B., Pentland, A.P., 2000. A Bayesian computer vision system for modeling human interactions. IEEE Trans. Pattern Anal. Machine Intell. 22 (8), 831–843.
Petkovic, M., Jonker, W., 2001. Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events. In: Proc. IEEE Internat. Workshop on Detection and Recognition of Events in Video, Vancouver, Canada.
Qu, W., Bashir, F.I., Graupe, D., Khokhar, A., Schonfeld, D., 2005. A motion trajectory based video retrieval system using parallel adaptive self organizing Maps. In: Proc. IEEE Internat. Joint Conf. in Neural Networks (IJCNN'05).
Ribeiro, P.C., Santos-Victor, J., 2005. Human activity recognition from video: Modeling, feature selection and classification architecture. In: Proc. 2005 Internat. Workshop on Human Activity Recognition and Modeling, Oxford, UK.
Robertson, N.M., Reid, I.D., 2005. Behavior understanding in video: A combined method. In: Proc. 10th IEEE Internat. Conf. on Computer Vision (ICCV'05).
Rocchio, J.J., 1971. Relevance Feedback in Information Retrieval. Prentice Hall Inc..
Rui, Y., Huang, T.S., 1999. A Novel relevance feedback technique in image retrieval. In: Proc. 7th ACM Internat. Conf. on Multimedia.
Rui, Y., Huang, T.S., Mehrotra, S., 1997. Content-based image retrieval with relevance feedback in MARS. In: Proc. Internat. Conf. on Image Processing.
Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S., 1998. Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Segmentation, Description, and Retrieval of Video Content 18(5), 644–655.
Salton, G., McGill, M.J., 1983. Introduction to Modern Information Retrieval. McGraw-Hill Book Company.
Sato, K., Aggarwal, J.K., 2004. Temporal Spatio-velocity transform and its application to tracking and interaction. Computer Vision and Image Understanding 92 (2), 100–128.
Shi, Y., Bobick, A.F., Essa, I.A., 2006. Learning temporal sequence model from partially labeled data. CVPR 2, 1631–1638.
Su, Z., Zhang, H., Ma, S.L.S., 2003. Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning. IEEE Trans. Image Process. 12 (8), 924–937.

Please cite this article in press as: Zhang, C., et al. Semantic retrieval of events from indoor surveillance video databases. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.05.004