Semantic Event Retrieval from Surveillance Video Databases

Xin Chen and Chengcui Zhang
Department of Computer and Information Sciences, University of Alabama at Birmingham, Birmingham, Alabama, 35294 USA
{chenxin, zhang}@cis.uab.edu

Abstract

This paper proposes a framework for retrieving semantic video events from indoor surveillance video databases. The goal is to locate video sequences containing events of interest to the user. This framework starts by tracking objects and segmenting videos into Common Appearance Intervals (CAIs). The spatiotemporal trajectories are obtained, based on which features are extracted for the construction of semantic event models. In the retrieval, the database user interacts with the machine and provides "feedbacks" to the retrieval result. The learning component learns from the spatiotemporal data, the semantic event model as well as the "feedback" and returns the refined result to the user. Specifically, the learning algorithm is developed based on 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.

1. Introduction

In an intelligent surveillance system, a large amount of surveillance videos are collected via surveillance cameras and stored in the database. 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 automatic semantic event retrieval, with the main focus on semantic event retrieval.

In our previous work [4], we proposed an object tracking algorithm, 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 games. In L. Chen's CAI model [3], 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. We tailor this concept to the specific needs of video segmentation in the proposed framework.

There are many existing works on automatically detecting semantic 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 [8]. In [5], a classification framework based on trajectories is proposed to retrieve traffic accidents by analyzing the sudden behavioral change of each individual vehicle. Interactions between and among vehicles are not studied in [5] because analyzing the behavior of individual vehicles is sufficient for the purpose of accident detection. However, 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 semantic events from them, there is often a need to analyze the interactions among these objects. In [2], the Coupled Hidden Markov Model (CHMM) is used for modeling human object interactions. The work in [2] 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 classify semantic events at a fine/micro level such as separating fighting from handshaking events.

978-0-7695-3454-1/08 $25.00 © 2008 IEEE
DOI 10.1109/ISM.2008.82

It is a challenge to manage and retrieve video sequences according to their semantic meanings. This is due to the fact that a machine does not have an equal ability in extracting semantic concepts from low level features as a human does. By analyzing only the low level features, no matter how sophisticated the algorithm is, there is still a "semantic gap" between the low level features and high level human concepts. Therefore, the machine needs the user's guidance in order to learn his real need/interest. 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 pre-define a comprehensive set of training sets for all "relevant" classes before the query, due to the scarcity of "relevant" samples and the uncertainty of users' interest. This is especially true in large multimedia databases, where multiple "relevant" and "irrelevant" classes exist according to the different preferences of different users [7], and the data in each "relevant' class may only constitute a very small portion of the entire database.

To solve this problem, we adopt a technique called "Relevance Feedback" [9] in the proposed semantic retrieval framework. When the framework returns the initial 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. 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. As the Relevance Feedback technique has been commonly adopted in content-based image retrieval, it is seldom used in video retrieval except for key-frame based video retrieval [11]. We adjust it to fit the needs of semantic video retrieval in this paper.

Section 2 briefly introduces a semantic object extraction and tracking algorithm and the video segmentation. Section 3 exemplifies the semantic event modeling. Section 4 presents the design details of the learning and retrieval process. Section 5 provides the experimental results. Section 6 concludes the paper.

2. Video segmentation and object tracking

2.1. 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 L. Chen's CAI model [3], 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.

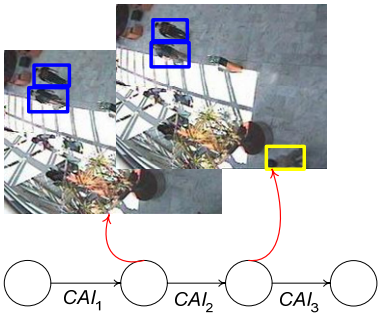

Figure 1. Video Segmentation with CAIs

Figure 1 illustrates the video segmentation schema used in the proposed framework. 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.

2.2. Automatic object tracking

With the segmented surveillance videos stored in the database, the next step is to perform object tracking on these videos. In our previous work [4], 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. With this algorithm, we can obtain blobs of objects in each frame and further acquire the Minimal Bounding Boxes of the objects as well as the coordinates of each object blob's centroid, which is used for tracking the positions of objects across continuous video frames. The framework in [4] also has the ability to track moving objects within

successive video frames. By distinguishing the static from mobile objects in the frame, tracking information can be used to determine the trails of objects.

By tracking each moving object in the video, a series of object centroids on successive frames are recorded. We can approximate the trajectory of the object by using the least-square curve fitting. A k^{th} degree polynomial for the curve is:

$$y = a_0 + a_1x + \ldots + a_kx^k \quad (1)$$

The fitted curve represents a rough shape of the moving trajectory. It can be described by only a few polynomial coefficients. The first derivative of a polynomial curve is a tangent vector representing the velocities of that object at different time. With trajectory modeling, we have the option to store only coefficients in the database instead of the whole set of centroids which contains redundant information and is not suitable for event retrieval.

3. Semantic event modeling

In this study, a spatiotemporal model is built for detecting abnormal behaviors in indoor surveillance videos. In the experiment, we used CAVIAR videos [1] taken in the lobby of a building in France.

After the video segmentation and object tracking, the spatiotemporal information of moving and static objects is obtained. In each CAI, pairs of object trajectories are studied, which will be referred to as Sequence Pair (SP) in this paper. It is observed that the abnormal human interactions 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.

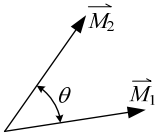

Figure 2. The degree of alignment

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 Figure 2; $\overrightarrow{M_1}$ and $\overrightarrow{M_2}$ 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. Optical Flow can be used to describe the velocity and the direction of the motions in bounding boxes. 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. The problem is then to find out the corresponding points between two consecutive frames. The Optical Flow technique is based on the assumption that images are made of patches whose intensities change smoothly. Therefore, for a pixel in the image, its surrounding pixels have similar intensities. Suppose a pixel's intensity is $I(x_0, y_0, t_0)$ in the current frame and it moves to (x_1, y_1) at time $t_1=t_0+\partial t$ with $x_1 = x_0+\partial x$, and $y_1= y_0+\partial y$. For the surrounding pixels (x, y) in the local patch, we want to solve the following problem and get ∂x and ∂y:

$$\min(\sum_{(x,y)}(I(x,y,t)-I(x+\partial x,y+\partial y,t+\partial t))^2) \quad (2)$$

The Optical Flow of point (x_0, y_0) is then:

$$(F_{x_0},F_{y_0})=(\frac{\partial x}{\partial t},\frac{\partial y}{\partial t}) \quad (3)$$

We compute the motion energy of an object in its bounding box at time *t* as the mean flow norm:

$$M=\frac{1}{N(t)}\sum_{(x,y)}\left\|(F_x,F_y)\right\|=\frac{1}{N(t)}\sum_{(x,y)}\sqrt{F_x^{\,2}+F_y^{\,2}} \quad (4)$$

where, *N* is the total number of pixels in the bounding box at time *t*. We use the motion energy *M* of each object within its bounding box as another property of the object.

As mentioned in Section 1, to use Relevance Feedback, some heuristics need to be established in order to conduct the initial query. For those abnormal behaviors such as two people fighting with each other, 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 α_t = [$vdiff_t$, θ_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 $\alpha'=[\alpha_1',\ldots,\alpha_n']$.

4. Semantic event learning and retrieval

4.1. Coupled Hidden Markov Model

Hidden Markov Model (HMM) is a stochastic model known for its ability to model processes that have structure in time since it automatically performs dynamic time warping. HMM can be denoted as $\gamma=\{T, A, B, \pi\}$. T is the number of states. A is the transition probability distribution matrix. B is the observation probability distribution. π is the initial state distribution. HMMs observe Markov Property in that future states are independent of the past states given the current state. HMM can be used to answer questions such as "what is the best model to describe how a given observation sequence comes out." The answer to this type of question is the training process. This is what we need to accomplish in the proposed framework. Formally, Equation 5 denotes the problem for getting posterior i.e., performing classification.

$$P(O|\gamma)=\sum_{Q} P(O|S,\gamma)P(Q|\gamma) \quad (5)$$

where O is the observation sequence; S is the state variable sequence; and γ is the model. Instead of enumerating all possible state variable sequences, this can be solved by a Forward-Backward procedure. For Question 3, there is in fact no optimal way of estimating the model parameters (A, B, π) given the observation sequences as training data. An iterative procedure such as Baum-Welch method (or equivalently a EM method [6]) can be used to choose parameters for model γ such that $P(O|\gamma)$ is locally maximized.

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 (α_t = [$vdiff_t$, θ_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) [4], which has compositional states, is seemingly a better choice.

Figure 3 is 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. In particular, we have each chain model the behavior of one person. The influences of each person to the other is reflected in the cross transitions between two chains. Therefore, the two persons both have their internal behavioral model as well as their interactions with another person.

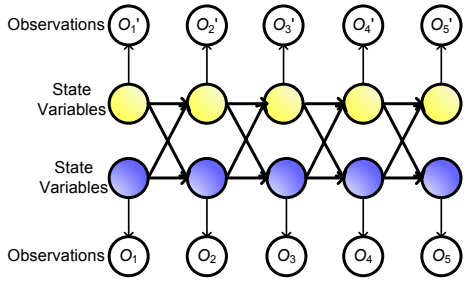

Figure 3. CHMM rolled out in time.

4.2. 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. 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 method is also adopted in this framework.

In the initial query, 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. 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 n Trajectory Pairs (TPs) and m Sequence 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 α_t = [$vdiff_t$, θ_t, $1/dist_t$, M_t] and α_t' = [$vdiff_t'$, θ_t', $1/dist_t$, M_t']. The relevance score of an SP is thus $\max_{t=1}^{l}(score(\alpha_t,\alpha_t'))$, where $score(\alpha_t,\alpha_t')=\sqrt{(1/dist_t)^2+vdiff_t^2+vdiff_t'^2+M_t^2+M_t'^2}$ · $<vdiff_t, vdiff_t'>$ are the velocity changes and $<M_t, M_t'>$ are the two object motion energies in that SP at time t, respectively. The degree of alignment, i.e., θ_t 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. As shown in Figure 4, 6 sequences are labeled "relevant" in a query for the event of two people fighting.

Figure 4. The user interaction interface

With this information at hand, a set of training samples can be collected. Each training sample is in the form of $<[\alpha_1, \alpha_2, \ldots, \alpha_l], [\alpha'_1, \alpha'_2, \ldots, \alpha'_l]>$. 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.

5. Experimental results

In this study, abnormal human interactions are modeled for indoor surveillance video retrieval. In particular, the retrieval of "fighting" events is tested with the proposed framework. The 10 testing videos are from the CAVIAR [1] video taken in the lobby of a building. 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 "fighting" 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 of a "fighting" event. 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 from the CAVIAR videos. 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 [10] 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.

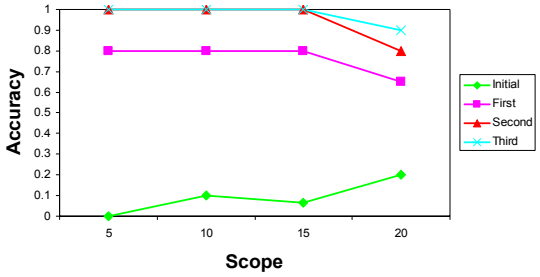

Figure 5. Retrieval accuracies across four iterations

From Figure 5, we can see that the retrieval accuracies of the proposed framework increases steadily across iterations with the incorporation of the user's feedback. In the second iteration, the total accuracy 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%.

In our experimental design, the proposed framework is compared with the HMM and the traditional weighted relevance feedback method. For the HMM, each SP is represented by a series of seven-

feature vectors $<1/dist_t, \theta_t, \theta_t', vdiff_t, vdiff_t', M_t, M_t'>$. It models each SP as a 7-channel sequence instead of two multi-channel sequences as in CHMM.

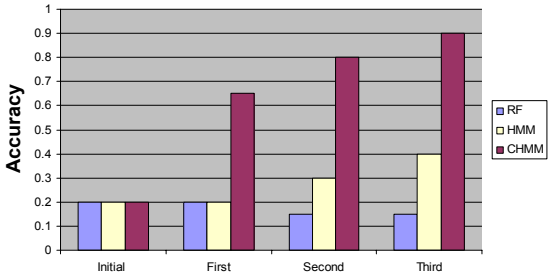

Figure 6. Compare the accuracies across iterations

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 $<1/dist_t, vdiff_t, vdiff_t', M_t, M_t'>$ 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. Each weight is further normalized by its percentage in the total weight.

Figure 6 compares the accuracies among the top 20 returned sequences across four iterations. "RF" is the weighted relevance feedback method aforementioned. "HMM" is the Hidden Markov Model, which has only one chain. "CHMM" is the proposed framework. It is obvious that the proposed framework outperforms the weighted relevance feedback as well as the HMM based method. 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.

6. Conclusions

In this paper, a human-centered semantic video retrieval platform is proposed. Given a set of videos, the semantic objects are tracked and the corresponding trajectories are modeled. Heuristic spatiotemporal event models are then constructed. The goal is to automatically retrieve abnormal human interactions in indoor surveillance videos. For learning and retrieval, the Coupled Hidden Markov Model is adapted to fit the specific needs of event identification and retrieval for in-door surveillance video data. The platform shows its effectiveness as demonstrated by our experiments on indoor surveillance videos. In the learning and retrieval phase, with the top returned 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 semantic 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.

Acknowledgement

The research of Dr. Chengcui Zhang is supported in part by NSF DBI-0649894.

References

[1] CAVIAR: Context Aware Vision using Image-based Active Recognition. (2004, Jan. 10). [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR.

[2] N. Brewer, N. Liu, O. D. Vel, and T. Caelli, "Using Coupled Hidden Markov Models to Model Suspect Interactions in Digital Forensic Analysis " in Proc. of the International Workshop on Integrating AI and Data Mining (AIDM'06), 2006.

[3] L. Chen and M. T. Özsu, "Modeling of Video Objects in a Video Database," in Proc. of the IEEE International Conference on Multimedia, Lausanne, Switzerland, 2002.

[4] S.-C. Chen, M.-L. Shyu, S. Peeta, and C. Zhang, "Learning-Based Spatio-Temporal Vehicle Tracking and Indexing for Transportation Multimedia Database Systems," IEEE Trans. on Intelligent Transportation Systems, vol. 4, no. 3, pp. 154-167, 2003.

[5] X. Chen and C. Zhang, "An Interactive Semantic Video Mining and Retrieval Platform-Application in Transportation Surveillance Video for Incident Detection," in Proc. of the IEEE International Conference on Data Mining, Hong Kong, China, 2006.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Roy. Stat. Soc., vol. 39, no. 1, pp. 1-38, 1977.

[7] M. Nakazato, C. Dagli, and T. S. Huang, "Evaluating Group-based Relevance Feedback for Content-based Image Retrieval," in Proc. of the IEEE International Conference on Image Processing (ICIP'03), Spain, 2003.

[8] N. M. Robertson and I. D. Reid, "Behavior Understanding in Video: A Combined Method," in Proc. of the Tenth IEEE International Conference on Computer Vision (ICCV'05), 2005.

[9] Y. Rui, T. S. Huang, and S. Mehrotra, "Content-based Image Retrieval with Relevance Feedback in MARS," in Proc. of the International Conference on Image Processing, 1997.

[10] Z. Su, H. Zhang, and S. L. S. Ma, "Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning," IEEE Trans. Image Processing, vol. 12, no. 8, pp. 924–937, 2003.

[11] C. Calistru, C. Ribeiro, G. David, I. Rodrigues, and G. Laboreiro, "INESC, Porto at TRECVID 2007: Automatic and Interactive Video Search," TRECVID 2007.