HIERARCHICAL VIDEO INDEXING AND RETRIEVAL SYSTEM

Mohammed Yassine Kazi Tani*, Abdelghani Ghomari*, Lamia Dali Youcef**
*Université of Oran
Computer Science Department
Research on Industrial Informatics and Networks Laboratory (RIIR)
BP 1524, El-M'Naouer 31000 - Oran, Algeria
{yassine.kazi@gmail.com; ghomari65@yahoo.fr}
** Abou Bakr Belkaid University of Tlemcen
GEE Department
Systems and Technologies of Information and Communication Laboratory (STIC)
B.P 230, Chetouane- Tlemcen-
lamiadaliyoucef@mail.univ-tlemcen.dz

Abstract

In this paper we will improve a previous system named: Semantic Retrieval of Event from Indoor Surveillance Video Database by adding a hierarchical indexing approach. The aim of our work is to improve the initial result provided by the system and taking into account moving studies of objects in video documents of videosurveillance applications.

Key words: video document, indexing and retrieval video surveillance, indexing approach, Semi- automatic annotation.

1. Introduction

Nowadays, the existence of multiple sources of video capture (Phone, Videosurveillance…) attracted several researches in the field of modeling, indexing and retrieval video. The importance size of video documents requires new compressing methods to facilitate their use on the large network like Internet. For this, many standards of compression exists like: MPEG1, MPEG2, MPEG4 [2] and MPEG7 [3] that change the context of compression for standardizing the description of multimedia document content. MPEG21 [4] is also a standard of compression that describes the method of multimedia documents production and the consumption of their content.

The large scale of video databases used actually in many applications domains such as the videosurveillance require an efficient indexing system for videos retrieval.

For this purpose, there exist in the literature two approaches of indexing and retrieval video documents: the first one is based on textual annotations and the second one is based on the visual content (segmentation and analysis of the different structural units content) [7].

So, to overcome the problematic of indexation in video documents, a semi-automatic annotation technique exists which benefits of manual and automatic annotations advantages [5, 6].

In this paper, we try to improve a work called "Semantic retrieval of events from indoor surveillance video database" [18] by giving our point of view (the system is presented in the section 3 and 4 and its implementation is being done). At the same time, we try to highlight other points such as related works (Section 2) and the conclusion (Section 5).

2. Related Works

2.1 Video documents

2.1.1 Components of video documents

Content and container terms are essential to know in a multimedia document. The content is written on a text medium and the paper is the container [1]. Concerning video documents, the indexing is focused on the video sequences that can be a Plan or a Scene that compose the video documents (figure1).

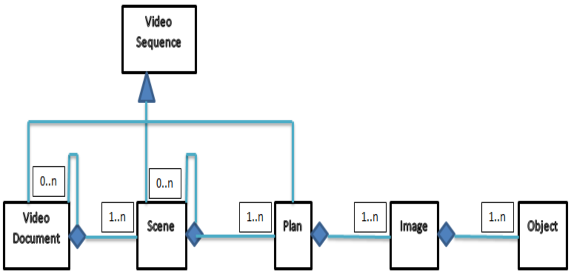

Figure 1 Structural unit of video documents

2.1.2 Characteristic of video documents

We can divide these features in three categories:

1. Media data: The video document itself, and the information about the compression format, the size of the video;
2. Metadata: The information about the video content, such as visual features (color, texture,…) and spatio-temporal characteristics;
3. Semantic data: This means the textual annotations that define the content of the video.

Thus, from these categories, the features of video documents are shown as follows [7] :

- Physical features: we can find the format (.Mpg), the type of compression (MPEG1, MPEG2, MPEG4 etc.), the size and the speed (number frames/second) NTSC (30 frames/second), the length and the name of video;

- Visual features: also called low-level features, like colors represented by a color scheme, texture (measure RGB values of a pixel relative to the other neighboring pixels), shapes and contour;

- Semantic features: also called high-level features, where we find the notion of annotations that define the content of the video.

2.1.3 Video documents indexing

The importance size of video documents manipulated in many critical applications like videosurveillance requires an efficient indexing system for videos retrieval.

The indexing process represents an operation that interprets, describes and characterizes a document or a part of a document for a future use.

According to the figure 2, the indexing process is divided into four major steps [5]:

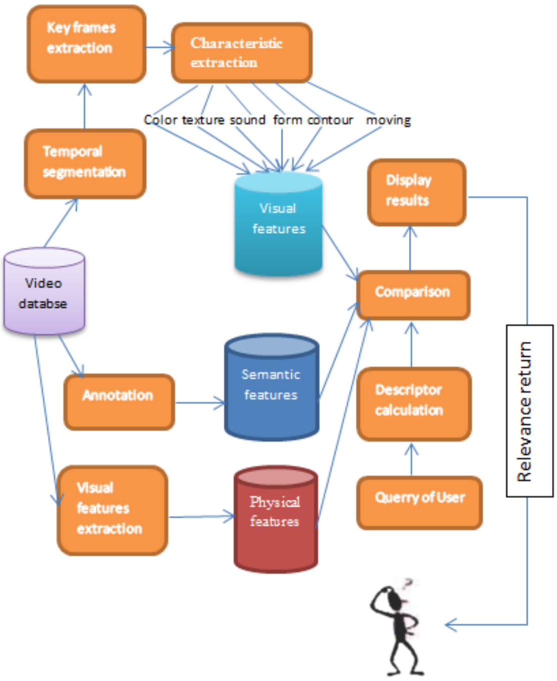

Figure 2 A video indexing and retrieval system

- Segmentation: It consists in dispatching the whole video into several parts “plan or scenes” as needed and especially to keep the same semantic aspect of the scene or plan in order to facilitate their indexing (figure 3**)**. As a result, several methods exist for video segmentation [14, 15, 16, 17], the difference from pixel to pixel, the comparison of color histograms, motion estimation, and so on….

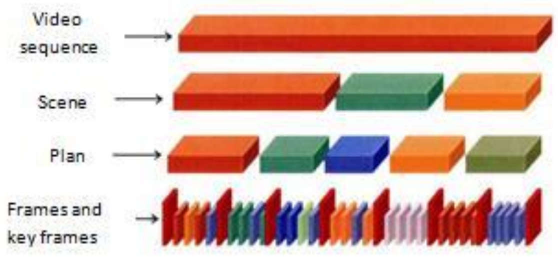

Figure 3 Temporal segmentation of video sequences

- Representation and classification

After the segmentation step, different types of features "Physics, Semantics, visual" also called digital signatures are extracted and assigned to different video sequences for two purposes: interpretation or description.

- Index creation

The most widely common methods for index creation are those based on the annotation with its different forms: *manual annotation, automatic annotation* and *semi-automatic annotation*. The annotation expresses two distinct aspects “description and interpretation”. In the description of the video, we can find all concepts that explain the video content (objects, people, places, events ...), and interpretation gives a point of view to explain a given sequence or any other part of the video [13].

- *Manual annotation*: is done by a human being who will annotate the various videos in the database with its own semantics and its own way of interpretation. The advantage of this method is to be accurate but when the database is very large, the annotation process is very heavy for the annotator as he/she will be obliged to browse the entire database to annotate it.
- *Automatic annotation*: Unlike manual annotation, automatic annotation is made by a machine that consists of extracting the different features of video and then spreading them in order to annotate other related videos that have the same features in the database. The automatic annotation has the advantage of annotating a large database, but its biggest flaw is that it is unable to give satisfactory results when the videos contain several objects, several movements, many people….
- Semi-*automatic annotation:* To overcome the problems presented by the both previous methods, semi-automatic annotation is based on the accuracy of manual annotation to annotate a part of its database and then use the advantage of automatic annotation in the purpose of annotating a very large database by comparing all video that have the similar visual features.

- Retrieval and interactivity

Retrieval step represents the final goal in the indexing process. Therefore, present retrieval systems [5] allow expressing the user query in four different ways:

1. *Retrieve by physical features*: the user can formulate his query by physical features such as modification date, size and number of images...

1. *Retrieve by semantic features*: the most common of all keywords retrieve represents the most used in the world as used for example by YouTube and allows the user to express his query based on keywords that represent the semantic of the video.

2. *Retrieve by visual features*: This type of retrieve is performed by inserting a video key by the user with which the system performs a comparison of low-level features with existing videos in the database.

3. *Retrieve by features combination*: This is the type of research that gives more satisfaction as far as the accuracy is based on the three types of features (physique, visual and semantic).

Interaction represents the dialogue interface between the user and the indexing system which expresses queries with different existing types.

2.2 SHIATSU (Tagging and Retrieving Video without Worries)

SHIATSU [6] is a semi-automatic system that covers the problems due to the use of only textual annotation like the semantic gap. The existence of synonyms (indexed by a synonym of the keyword in the query formulated by the user), homonymy/polysemy (two synonyms’ words). The architecture of SHIATSU system is based on three ideas:

1. *The hierarchical annotation*: makes two levels of indexing and starting with the sequences that make up the video and then proceed after that to indexing the entire video with summarizing the different indexes sequences;
2. *The similarity-based labeling*: Assign previously existing indexes to different videos that have the same visual features;
3. *The indexing and retrieval based on multidimensional taxonomy*: A system which implies the existence of several dimensions (root retrieval).

To perform indexing videos, SHIATSU is based on:

2.2.1 Shot detection

In order to separate the video sequences, SHIATSU is based on the balance approach. It exploits the color histogram and the object border for comparing two successive frames.

Color histogram HSV (hue, Saturation and Value): the distance between two consecutive frames k and k +1 « d_{HSV} (k, k+1) » is defined by:

$$d_{HSV}(k, k+1) = \frac{1}{6N} \sum_i |h_k[i] - h_{k+1}[i]|$$

N is the number of pixels, h_k represents the histogram of the image k.

The resulting distance is compared to a threshold θ_{HSV} :

$$\theta_{HSV} = \frac{\beta_{HSV}}{M/f} \sum_{i=M-M/f+1}^{M} L_{HSV} \text{ (i)}$$

β_HSV is a sensitivity parameter (by default, it is at 1), M is the total number of images in the video, f is the frame rate in the video and L_{HSV}represents the list of ascending values HSV away from all consecutive sequences.

ECR “Edge Change Ratio”: Change report between two frames k, k+1 is calculated as follows:

$$\text{ECR (k, k+1)} = \max\left(\frac{X_k^{out}}{\sigma_k} \frac{X_{k+1}^{in}}{\sigma_{k+1}}\right)$$

σ_k is the number of pixels edge and k, X_k^{out} , X_{k+1}^{in} represent respectively pixels of existing and new edges in the frames k and k +1.

The change ratio is compared to a threshold θ_{ECR} :

$$\theta_{ECR} = \frac{\beta_{ECR}}{2M/f} \sum_{i=M-2M/f+1}^{M} L_{ECR} \text{ (i)}$$

β_{ECR} is a sensitivity parameter (by default, it is 1) L_{ECR} and represents order list crossing ECR values.

Whenever the two values (d_{HSV} (k, k +1)) and ECR (k, k +1)) exceed their thresholds, there will be a cut to separate the two video sequences consecutively.

2.2.2 Indexing video

There are two levels of video indexing, sequences indexing and then the entire video indexing, because the system SHIATSU [6] is based on hierarchical annotation.

Sequences indexing: is based on different key frames that compose it. The process of selecting key frames can be done in three different ways:

- Select the first frame of each sequence;
- Select the first, the middle and the last frame of each sequence;
- Select a depending number on the sequence length L (s), N (k) = C. L(s)/ f, where C is a constant.

We can define the indexing process video sequences as follows:

After extracting key frames of each video sequence, each of them will pass through the extractor of visual features to extract color and texture. These features are then used by the annotation module to search from the database the frames that have the same features. Indexes are then proposed for this key frame and this is repeated for all key frames sequences and we take only the terms which recur most in the majority of key frames.

In the end, the user can choose the indexes proposed by the system or introduce its own indexes.

Hierarchical indexing: In order to index the whole video, we proceed as follows: We first compute the relevance of each sequence "S" length in relation to the whole video.

$$W(s) = \frac{L(s)}{L(v)}$$

Then, we calculate the rank R (t) of each sequence index.

$$R(t) = \frac{1}{N_s} \sum_s W(s)\, A(t, s)$$

N_s is the total number of video sequences and A (t, s) is the relevance of the index "t" in relation to the sequence "S" (A (t, s) = 0 when the sequence "S" has not the index "t").

In the end, we take the top 10 R (t) as an index of the whole video sequence.

2.2.3 Retrieval method

In the literature, we can find the most used retrieval system such as SHIATSU [6] that offers three ways:

- KS (keyword retriever) ;
- FS (frame retriever) ;
- KFS (keyword and frame retriever): that represents the retrieve by combining the two previous methods.

2.4 Semantic retrieval of events from indoor surveillance video database

Here, we focus our interest to the main goal of this work [18] which guides users to find required sequences in database of video surveillance. At this end, several steps are required:

- *Preprocessing*: the raw video is analyzed by segmenting videos into CAIs [19] and tracking semantic objects (human) in them.
- *Trajectory modeling*: in each CAI, trajectories are further modeled with the sliding window technique.
- *Event modeling*: In this study, an event model for two people fighting is built, and the feature vectors of human objects at consecutive time point are extracted.
- *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 result to the user.

- *Interactive learning and retrieval*: the user responds to the retrieval results by giving his/her feedbacks and refines the retrieval results in the next iterations until a satisfactory result is obtained.

The CAVIAR [20] videos database is used and the results of this framework are shown in the following graph (figure 4):

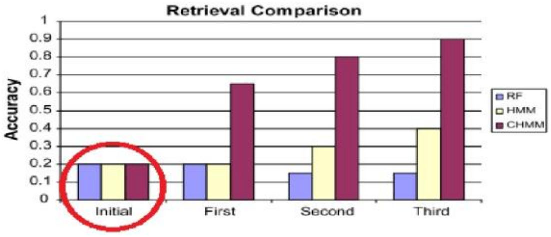

Figure 4 Accuracies of "meeting and fighting" events across iterations

From this graph, we can see the accuracy of the initial results returned to the user and this accuracy of 0.2/1 is very low. For this purpose, our approach is based on a hierarchical indexing to improve the initial results returned to the user.

3. Our hierarchical video indexing and retrieval approach

After having seen and analyzed the graphs resulting from the experiments done by the work of "Semantic retrieval of event from indoor surveillance video database" [18], we have noticed that during the initial iteration, there was a little relevance in the result returned by the indexing system according to the query of the user. Therefore, there was a continuous need to do the relevant feedback "RF" in order to improve the final result. For this purpose, our approach (figure 5) is to improve the initial iteration accuracy. This is possible when we include a hierarchical indexing [6] in the step of "event modeling". So, the proposed approach is as follows:

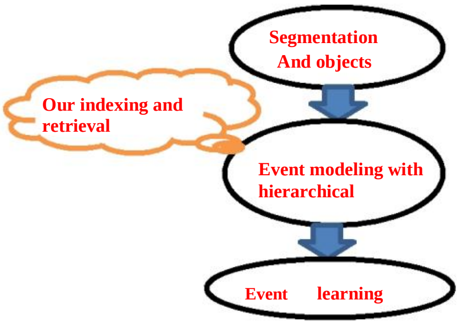

Figure 5 Our video indexing and retrieval approach

3.1 Video segmentation and objects tracking

In this step, we used the CAIs technique "Common Appearance Interval" for segmentation [19] (figure 6):

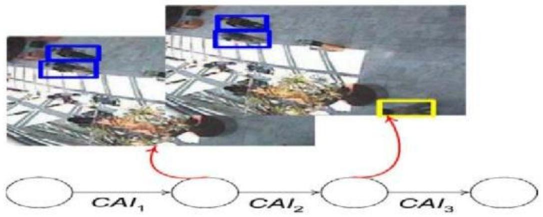

Figure 6 Video segmentation with CAIs

As for object tracking, a method called Simultaneous partition and Class Parameter Estimation (SPCPE) associated with Background learning and Subtraction methods are used.

3.2 Event modeling

In order to improve the indexing process, we propose a hierarchical annotation thanks to indexing the different CAIs (normal or abnormal human interaction) before annotating the whole video sequence.

First, for annotating the different CAIs, we need to extract the three properties for normal human interaction:

- Dist: distances between two objects in the SP (Sequence Pair);
- Θ : degree of alignment of two objects (i.e. M1 and M2 are the motion vectors of two objects at time t) (figure 7);

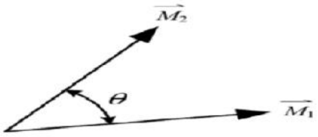

Figure 7 The degree of alignment

- Vdiff: changes of velocities of the two objects between two consecutive frames.

In addition, another propriety that is the magnitude of motion change of each object which can be analyzed by Optical Flow needs to be taken into account for abnormal human interactions "meeting and fighting" or "robbing and chasing".

After annotating the different CAIs, the indexing process is improved thanks to a hierarchical indexing which indexes the whole video. We proceed as follows:

We first compute the relevance of each CAIs "C" length in relation to the whole video.

$$W(c) = \frac{L(c)}{L(v)}$$

L(v) represents the length of the whole video. Then, we calculate the rank R (t) of each CAIs index.

$$R(t) = \frac{1}{N_c} \sum_c W(c)\ A(t, c)$$

N_c : is the total number of CAIs and A (t, c) is the relevance of the index "t" in relation to the CAIs "C" (A(t, c) = 0 when the CAIs "C" has not the index "t"). And last, we take the R (t) that has the most occurrences as an index of the whole video sequence.

3.3 Event learning and retrieval

In this step, we keep the same learning algorithm CHMM "*Coupled Hidden Markov Model*" [18] and we also use the relevant feedback after the initial query of the user if necessary.

In our approach, we think that it would be the least possible necessary to use the relevant feedback "RF" and the result of initial query will be performed.

4. Our video indexing and retrieval system

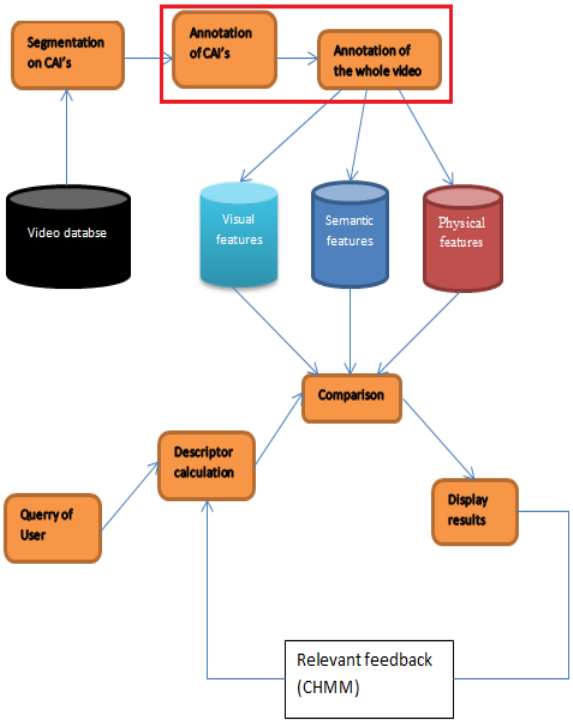

Figure 8 Our video indexing and retrieval systems

The system working proceeds as follows: we segment the videos from video database (Figure 8) using the (CAIs) technique. Then, to improve the initial result feedback of the system cited in [18], we annotate the different CAIs segments to get a set of index that contribute to annotate the whole video. This process represents the hierarchical indexing.
Hereafter, we store the result of the annotations in three sets of databases: visual, semantic and physical.
When a user send a query to the system, the descriptor processing's module extract the different characteristics of this query and forward them to the comparison module. This module makes similarities with the characteristics stored in the three set of databases and then displays adequate videos to the user.
The aim of our approach is to maximize the user satisfaction in the initial query to avoid relevant feedbacks of the CHHM algorithm.

5. Conclusion

In this paper, we discuss our proposed video indexing and retrieval approach by explaining the hierarchical indexing technique to improve the initial results obtained in [18]. Our video indexing and retrieval system is under development in the RIIR Laboratory and will be experimented by using the CAVIAR video database in the future.

References

[1] " Assistance Intelligente a la RI", book chapter: Indexation multimédia, Rédigé par Bruno Bachimont.

[2] LEE, H .Standard coding for MPEG1, MPEG2 and advanced coding for MPEG4. [En ligne] Rapport EE8205, 6 juin 1997, 15 p. Disponible sur : http://citeseer.nj.nec.com/lee97standard.html

[3] International Organisation for Standardisation. Overview of the MPEG7 Standard (version 6.0). ISO/IEC/JTC1/SC29/WG11 N4509. December 2001 , Pattaya, 90 p . Disponible sur : http://mpeg-industry.com/mp7a/w4980_mp7_overview1.html

[4] BORMANS, J., HILL, K. MPEG21 Overview. [En ligne] ISO/IEC JTC1/SC29/WG11/N4318. Juillet 2001, Sydney. Disponible sur : http://ipsi.fhg.de/delite/Projects/MPEG7/Documents/mpeg21-Overview4318.htm

[5] Un système pour l'annotation semi-automatique des vidéos et application à l'indexation, université du Québec, Aout 2009.

[6] M. Patella . C. Romani, l. Bartoloni. "SHIATSU: tagging and retrieving video without worries", Springer Science+business Media, LLC, 2011.

[7] Contribution aux techniques orientées objets de gestion des séquences vidéo pour les serveurs web, Mihaela SCUTURUCI, 2002.

[8] CHAN, S.S.M., WU, Y., LI, Q., ZHUANG, Y. A Hybrid Approach to Video Retrieval in a generic video Management and Application Processing Framework. [En ligne] Proceedings of the Second IEEE International Conference on Multimedia and Expo (ICME'01). August 22-25, 2001, Tokyo, Japan. Disponible sur: http://citeseer.nj.nec.com/chan01hybrid.html

[9] F SOUVANNAVONG, « Indexation et recherche de plan Vidéo par le contenu Sémantique », Thèse sur le traitement de signal et des images, Ecole Nationale Supérieure des Télécommunications, Paris, pp. 141, juin 2005.

[10] S LEFEVRE, J. HOLLER, N. VINCENT, « Segmentation Temporelle de Séquences d'images en Couleurs » Laboratoire d'Informatique, Université de Tours, France.

[11] A. HANJALIC, R. L. LAGENDIJK, and J. BIEMOND, "Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems", IEEE Transactions on Circuits and Systems for video Technology, pp. 580-588, juin 1999.

[12] R, BRUNELLI, O. MICH, and C. M. MODENA, "A Survey on the Automatic Indexing of Video Data", Journal of Visual Communication and Image Representation, ITC-irst, I-38050 Povo, Trento, Italy, pp. 78-112, Juin 1999.

[13] A. SALWAY, "Video Annotation: the Role of Specialist Text", Thesis, Departement of Computing, School of Electronic Engineering, Information technology and Mathématics, University of surrey, Guildford, United Kingdom, pp. 188, December 1999.

[14] P. WU, "A Semi-automatic Approach to Detect Highlights for Home Video Annotation", IEEE International Conference on Acousctics, Speech, ans Signal Processing, Montreal, Quebec, Canada, vol. 5, pp. 957 – 960, Mai 2004.

[15] IIARIA Bartoloni, Marco Patella, Corado Romani, SHIATSU, tagging and retrieving videos without worries, 2011

[16] Jacobs A, Miene A, Ioannidis GT, Herzog O (2004) Automatic shot boundary detection combining color, edge, and motion feautures of adjacent frames. In: TRECVID 2004, Gaithersburg, MD, pp 197-206.

[17] Qu Z, Liu Y, Ren L, Chen Y, Zheng R(2009) A method of shot detection based on color and edges features. In: SWS 2009, Lanzhou, China, pp 1-4.

[18] Semantic retrieval of event from indoor surveillance video database, Chengcui Zhang *, Xin Chen, Liping Zhou, Wei-Bang Chen, Journal homepage: www.elsevier.com/locate/patrec , available online 18 May 2009.

[19] L. Chen, M.T Ozsu, "Modeling of video Objects in a video database". IEEE Conference on Multimedia, Lausanne, Switzeland pp , 2002.

[20] Caviar video database, http://homepages.inf.ed.ac.uk/rbf/CAVIAR.