SPATIO-TEMPORAL BLOCK MODEL FOR VIDEO INDEXATION ASSISTANCE

Alain Simac-Lejeune
Listic, Université de Savoie and Gipsa-lab / Université Joseph Fourier, Grenoble, France

Michèle Rombaut
Gipsa-lab, Université Joseph Fourier, 961 rue de la Houille Blanche, BP 46, F-38402 Grenoble Cedex, France

Patrick Lambert
Listic, Université de Savoie, BP 80439 74944 Annecy-le-Vieux Cedex, France

Keywords: Video indexing, Spatio-temporal blocks, Assistance system, Questions/Answers approach.

Abstract: In the video indexing framework, we have developed an assistance system for the user to define a new concept as semantic index according to the features automatically extracted from the video. Because the manual indexing is a long and tedious task, we propose to focus the attention of the user on pre selected prototypes that a priori correspond to the concept. The proposed system is decomposed in three steps. In the first one, some basic spatio-temporal blocks are extracted from the video, a particular block is associated to a particular property of one feature. In the second step, a *Question/Answer* system allows the user to define links between basic blocks in order to define concept block models. And finally, some concept blocks are extracted and proposed as prototypes of the concepts. In this paper, we present the two first steps, particularly the block structure, illustrated by an example of video indexing that corresponds to the concept *running* in athletic videos.

1 INTRODUCTION

In the image or video indexing framework, the automatic indexation task generally requires a preliminary learning task in order to link the index to the features extracted from the video. This learning task may be realized in two different ways. The first one consists in using a learning data base already annotated by users, but this manual annotation task is long and tedious despite some attempts to reduce the task length using a collaborative annotation (Ayache and Quénot, 2008). In this case the knowledge is indirectly introduced through the annotated data. The second way is the direct use of expert knowledge, but the methodology to extract this expert knowledge is not easy as generally experts are not specialists in image processing. An example of such an approach is detailed in (Valet et al., 2003) where the authors use a fuzzy rule system to translate user expertise. In any case, the user task is essential. The solution proposed in this paper is an hybrid solution. It consists to develop a system to assist the user to define new index and to annotate the learning data base.

In some way, the proposed approach is inspired by the text retrieval methods, and by the visual words (visual vocabulary) proposed by Sivic and Zisserman (Sivic and Zisserman, 2003) who first proposed quantizing local image descriptors for the sake of rapidly indexing video frames. They showed that local descriptors extracted from interest points could be mapped to visual words by computing prototypical descriptors with k-means clustering, in order to make faster the retrieval of frames containing the same words. Csurka et al. (Csurka et al., 2004) first proposed to use quantized local descriptors for the purpose of object categorization. Some image descriptors are mapped to a bag-of-words histogram corresponding to the frequency of each word. Then, the categories are learned using this vector representation. These proposed approaches are always funded

Simac-Lejeune A., Rombaut M. and Lambert P..
SPATIO-TEMPORAL BLOCK MODEL FOR VIDEO INDEXATION ASSISTANCE.
DOI: 10.5220/0003098904750480
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 475-480
ISBN: 978-989-8425-28-7

on a learning phase where the user must manually index the frames and video. Based on the same type of data architecture, the goal of the proposed system is to manage this learning phase.

In this paper, we present a new model of representation adapted to the video indexation. The proposed idea is to help the user to define concepts (i.e high level index). In a first step, some "basic spatio-temporal blocks" (section 3) are extracted from a set of low level features. Then the user is interrogated, with a set of adapted questions, in order to define some links between these basic blocks and the searched concepts. In order to make easier the definition of new index, only a set of "natural" questions and answers are proposed to the user. From these answers, "concept blocks" are defined (section 4). Finally, some video prototypes that seem correspond to concepts are proposed to the user who validates or not. The performances of this system are studied on the specific *running* concept within a set of 100 shots (section 5).

2 GENERAL ARCHITECTURE OF THE GLOBAL SYSTEM

The general architecture is composed of a three main steps with one preliminary step:

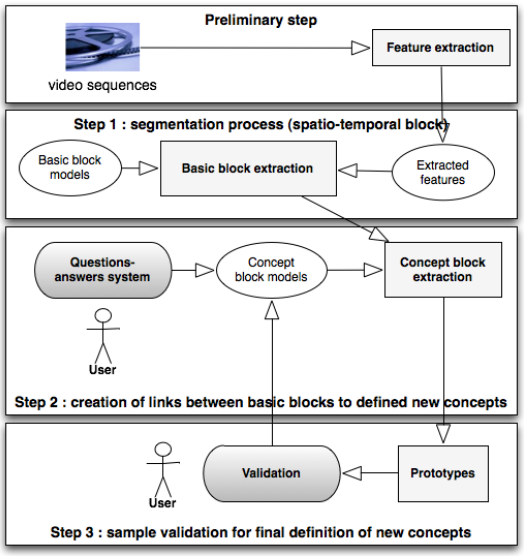

Figure 1: General process to define new concepts.

Feature Extraction. (Figure 1 - Preliminary step) Low level features are extracted from videos. The extractors used are among the most classical: interest points (Harris and Stephens, 1988),(Laptev and Lindeberg, 2003), optical flow (Bouguet, 2000), Hough detector (Duda and Hart, 1972), dominant motion (Odobez and Bouthemy, 1995) and dominant color.

Basic Spatio-temporal Block Extraction. (Figure 1 - Step 1) The basic spatio-temporal blocks are built using the features provided by the preliminary step and some basic block models. These models are *a priori* defined according to three main characteristics:

- a continuous temporal interval,
- a property of a specific feature (typically a feature value or a feature value interval),
- a spatial shape where the above property is verified. This shape could be a part of the image or the entire image.

Then, for each video, the basic spatio-temporal block extraction can be seen as an instantiation of the block models according to the extracted features.

Concept Block Extraction. (Figure 1 - Step 2) A concept is a high level index which has a semantic interpretation. It corresponds to a combination of different blocks, which can be basic blocks or other concept blocks. As for the previous step, this extraction requires the definition of concept block models. In order to build these models, it is necessary to use an expert knowledge given by a user. As the user is not necessarily familiar with image processing and is not able to explicitly express the links between these concept blocks, the basic blocks and the concept, a *Question/Answer* system is used in order to get this expertise and simultaneously build the concept block model. Then, instantiations of these models are performed by searching in videos the occurrences of these concept blocks, and consequently occurrences of concepts. This step is presented in section 4.

Sample Validation. (Figure 1 - Step 3) Each founded concept, which is regarded as a concept sample prototype, is proposed to the user who may accept or refuse it.

In this paper, we focus our attention on the extraction of the basic spatio-temporal and concept blocks.

3 BASIC SPATIO-TEMPORAL BLOCK

The first step of the video processing concerns the segmentation of the video in spatio-temporal objects

or blocks which are a sequence of images associated to a particular attribute. In such a block, the concerned attribute verifies a particular property, for instance it belongs to a particular interval. Thus, before video processing, a set of block models is defined. It can be seen as a visual vocabulary where each block is a space-time word.

3.1 Definition of Basic Block Models

A block model is defined through an attribute associated to a property. The property can be a value (the attribute has a specific value) or an interval (the attribute is within an interval - the most frequent situation). As a consequence, for a given attribute, there are several models of blocks corresponding to different values or intervals. For instance, we define the compactness $c \in [0,1]$ of an object as $c = \frac{minimum(width,height)}{maximum(width,height)}$. For this attribute, we propose to define three blocks models: low compactness (0 to 0.4), average compactness (between 0.4 and 0.65) and high compactness (between 0.6 and 1). Until now, the different ranges or values and the number of blocks are defined by expertise. At the end of this pre-processing phase, we obtain a database containing all the models of the basic block models. These models are relatively generic for different types of application, because they are only linked to the attributes and they have no particular semantic meaning. At the end of this modeling phase, we obtain a database containing all the basic block models. With the 40 attributes, 120 basic block models are defined. This number has to be compared to the size of visual word vocabulary used in static image indexation, which is typically equal to a few thousands.
The characteristics of all these models are stored in a database, that is: name, model type (value or interval), level (sequence, image or object) and finally the attribute values related to this block. Then, this structure can dynamically create queries to detect the occurrence of blocks in different indexed sequences.

3.2 Extraction of Basic Blocks

The extraction of the spatio-temporal blocks consists in looking for occurrences of model blocks within the videos. In other words, it consists in founding images or block of images within which an attribute verifies a model property. A block is defined by a model composed of an attribute and a property, an initial and a final image. For one model, a lot of basic blocks can be extracted. Conversely, an image or block of images can belong to different models. If the attribute of a block model corresponds to a moving object, an extracted basic block will correspond to a sequence of images containing the object and verifying the property. As all information is stored in databases and are classified by sequence, by block and increasing number of frames, the block extraction is realized using a query mechanism. To eliminate too small blocks and merge too close blocks, a morphological filtering is performed.

4 CONCEPT BLOCK MODEL

The indexation process is generally realized in a manual way because it refers to index with a high semantic level (denoted as concepts). For instance, a concept can be the *running action* of a personage. It is difficult to make this indexing process automatic, because it needs to link the low level attributes extracted from the video to the concepts that have a semantic interpretation. We assume that these links must be defined only with the help of the user. But before explaining how we propose to build these links, we define the concept block models.

4.1 Notion of Concept Block Models

A concept block model has a high-level semantic meaning which fits the user need in terms of indexation task. In this paper, we propose to build these models by combining models previously defined: basic block models or other concept block models.

The construction of such block models is performed by using simples combination rules: simultaneity of blocks (logical AND operator), presence of at least one block (logical OR operator), presence of only one block among two blocks, (logical XOR operator), succession of blocks (sequentiality), and alternation of blocks which is composed of several succession of blocks (periodicity). Initially, these operators are sufficient to define a number of concepts. Thereafter, it will be interesting to establish additional operators.

4.2 Learning Concept Block Models

The problem is to build relevant links between blocks. In such a situation, a classical approach consists in using a neural network or a supervised classification associated to a set of learning data corresponding to already indexing data (Burgener, 2006). But this manual indexing task, realized by users, is very tedious and very long. Another classical approach is to use the knowledge of an expert who explicitly defines the searched links. But in this case, he must be an expert

of image processing as well as an exert of the domain present in the video. Furthermore, the software translation of the user's expertise is not trivial. This is why we suggest in this paper a new approach that models the knowledge usually managed by the user without real assimilation. For this purpose, we propose to use an assistance system based on a *Question/Answer* process. The questions are addressed to a user who is not an expert in image and video processing but who is an expert in the specific domain present in the video. For each question, a set of answers are proposed and these answers are used to define the links between blocks, and finally the concept block model.

The *Question/Answer* structuring is a well known problem. The usual approach is to build a tree where nodes correspond to *Question/Answer* that lead to other questions, and where leaves correspond to *Question/Answer* that lead to the block models previously defined.

In the case of concept block model definition, the aim is to find most information based on extracted attributes, i.e. based on the block models defined previously. In other words, the system looks for all the leaves of the tree that correspond to the concept.

The system we have developed is composed of a list of *Question/Answer*. Each question can be activated or deactivated. If it is activated, it will be asked to the user. A list of possible answers is associated to each question. According to the concept he wants to define, the user chooses one of the answers. Each answer can be associated to i) one or more basic block models, ii) to combination rules between block models or iii) to the activation or inhibition of other questions.

As the links and the basic extracted blocks are mostly related to spatio-temporal situations, the main theme of the application is the movement. The current proposal contains 50 questions. All potential questions are ordered from the most generic (Environmental issues and context, type of shooting (camera)) to the most specific (Presence of moving or static persons, characteristics of the observed motion). We show in section 5 the scenario obtained when the user wants to define the *running* concept.

4.3 Extraction of "Concept" Blocks

Two approaches can be proposed to extract the concept blocks from the video. The first one consists in defining the model in a first step, then according to this model, building search requests in the block database previously defined in the second step. In this paper, we propose a second approach. The idea is to manage a dynamic list of spatio-temporal blocks or prototypes that potentially correspond to the concept. During the *Question/Answer* process, a request in the block database is done after each answer in order to select the prototype candidates in this dynamic list. The list is reduced to the candidates after each answer and the number of prototypes is transmitted to the user. The *Question/Answer* system stops when there is no more questions allowed from the list, or when the number of prototypes in the dynamic list is considered as enough reduced for the user. Finally, the prototypes are filtered and are stored in the basis with the others blocks.

In the third phase evoked in figure 1 but not described in this paper, the prototypes of the concept are presented to the user who can accept them as representative of the concept or not.

5 PERFORMANCES OF THE SYSTEM

In terms of video indexing, we are interested by the dynamic aspects. That is why the attributes extracted are movement oriented, and the proposed questions in the *Question/Answer* process essentially concerns the moving object and human behavior. In order to illustrate the system behavior, we propose to define the concept *running* for a personage.

5.1 Sequence Database

The video database is composed of: 100 shots (animation, TV, films, sports) with 20 of *running* and 80 of others without *running* that correspond to 12500 images, and 620 seconds of video. All information obtained on these 100 sequences represent about 0.6 MB. This represents about 50 bytes per image. The system is implemented in C++ with the use of OpenCV Library.

5.2 Basic Block Extraction

The first steps of the process are the feature extraction and basic block extraction as represented figure 1. These steps are achieved by preprocessing or offline process. The feature extraction is performed with an average speed of 10 frames per second. The basic block models use 6 attributes that correspond to 22 basic block models. The extraction of images corresponding to each basic block model is carried out fully automatically using a query builder requiring 30 seconds. The filtering concerns 2200 operations (two stages) for 100 shots and 22 basic block models. This step is less than 2 minutes. Finally, we obtain 8972

blocks, an average of 90 blocks per sequence with a maximum of 172 and a minimum of 20. The difference between min and max is mainly due to the fact that the attribute interval values are crisp. So variations around these values induce lot of different blocks. The 6 attributes (22 block models - each attributes generates two or three basic blocks) used for our application concern: moving object compactness, moving object orientation (horizontal or vertical), camera (static or moving), number of STIP, number of STIP by quadrant of the moving object.

5.3 Questions-answers system

We present in this section an example of scenario that corresponds to the step 2 of the figure 1. It is assumed the user wants to defined the concept *running*. The goal is to select prototypes in the videodatabase presented section 5.1 that potentially correspond to this concept.

A sample question with the answers provided and selected bricks are :
Question: `Is the personnage standing ?`
yes always, yes sometimes, no, impossible to answer
Answer: `yes always`
Corresponding Block: attribute `compactness` with property `low` AND attribute `orientation` with property `vertical`

5.4 Performances

The video database has been processed with the concept block models, and concept blocks have been extracted. A shot is considered as true if a concept block is included in. The results have been compared to the ground truth manually indexed. The results are the following:

Table 1: Precision and recall for the detection of the *running* concept on 100 shots.

	Real	True	False	Precision	Recall
Running	20	18	7	72%	90%
Others	80	73	2	97%	91%

In the database of 100 shots, the system extracted 25 shots of running where 18 shots are real running and 7 are false detection. The recall index is better than the precision index. This is due to the fact that the shots are eliminated from the list of prototypes all along the *Question/Answer* process. If attributes are not enough numerous and relevant for the concept, lot of shots remain after the *Question/Answer* process is finished. However, this step restricts the number of shots for the third step of validation by the user as represented figure 1. In this example, the user must manually index only 25 shots instead of 100 shots.

6 CONCLUSIONS AND PERSPECTIVES

The proposed system allows the user to limit this work of manual indexing for the construction of a learning data base. The *Question/Answer* principle is built assuming that the user is not a specialist of image processing, but is a specialist of the application. Until now, the system addresses specific applications of personage behavior analysis in the video. The extracted attributes are chosen to be relevant, but other attributes can be added for other types of application. For the same reasons, questions and answers are also oriented according to this application. As a future work, in order to extend the question/answer set, or to adapt the basic block models, an adaptive system corresponding to the third step of the figure 1 could use the user's opinion on the prototypes and ask the user to propose new question more relevant for his application.

ACKNOWLEDGEMENTS

We thank the Rhône-Alpes region for their support with LIMA project.

REFERENCES

Ayache, S. and Quénot, G. (2008). LIG and LIRIS at TRECVID 2008: High Level Feature Extraction and Collaborative Annotation. In *TRECVID Workshop*, Gaithersburg, MD, USA.

Bouguet, J.-Y. (2000). Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm.

Burgener, R. (2006). Artificial neural network guessing method and game - European Patent EP 1710735 (A1).

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*.

Duda, R. O. and Hart, P. (1972). Use of the Hough transformation to detect lines and curves in pictures. *ACM*, 15:11–15.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *The Fourth Alvey Vision Conference*, pages 147–151.

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. *ICCV'03*, pages 432–439.

Odobez, J. and Bouthemy, P. (1995). Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. *Proceedings of the International Conference on Computer Vision*.

Valet, L., Mauris, G., Bolon, P., and Keskes, N. (2003). A fuzzy rule-based interactive fusion system for seismic data analysis. 4(2):123–133.