Towards Automatic Annotation of Video Documents

Nabil Madrane and Morris Goldberg
Multimedia Department
EURECOM Institute
2229, Route des Crêtes, B.P. 193
06904 Sophia Antipolis Cedex France

Abstract

This paper presents a video indexing tool for sequences which contain moving persons, using a model-based dynamic scene analysis. We show how scene-specific knowledge can be used to guide the analysis process. Our approach permits to avoid some problems inherent to the segmentation of primitives. A model-based post-processing step is used to correct a posteriori the shape of the extracted objects. A scenario, describing the sequence in terms of basic events is proposed. These events constitute a first level of annotation and experiments show a good robustness regarding occlusions problems.

1 Introduction

Research in computer vision and scene analysis fields has shown the difficulty of the automatic annotation of video documents. Video data is inherently uninterpreted information in the sense that there currently are no general computational mechanisms for content-searching video data with the precision and semantic exactness of generalized textual search. By imposing specific restrictions onto the scene, the complexity of the problem can be reduced. One such restriction is the analysis of human body motion in a monocular sequence. This scenario is relevant for many applications.

In this contribution, an approach to automatically analyze human body motion from image sequences is presented. *Spatio-Temporal Indices* are extracted and describe the scene in terms of basic events. This is achieved by a model-based dynamic scene analysis combined with a high-level search algorithm.

2 3D human model

The 3D human model we use is composed of an articulated structure and a surface characterization.It is based on 10 rigid components whose relative position is described by articulation constraints. We can define the body's position and configuration in space by specifying its reference point and the orientation of each component in the articulated structure.

Concerning the articulation constraints (spatial constraints), we limit the amplitude of the 3 rotations around each articulation. In our analysis, we ignore the accuracy of the 3D structure of the individual components and instead exploit the articulation constraints.

Regarding the surface characterization, each surface patch of the 3D structure is associated with a vector of charateristic parameters. A connected set of surface patch with the same parameters is called a 3D region. In other words, we constitute a mapping of 3D regions on the model. Each component is composed by one or more 3D regions and each region is attached to a set of characteristic parameters.

3 Spatio-Temporal Indices

We define the *Spatio-Temporal Indices* as a collection of descriptors extracted from the raw video document. Its internal structure is described as follows :

- Frame number n.
- Set of 2D or 3D parameters.
- Description of the event.
- Pointer to the next *Spatio-Temporal Index.*

The field "Set of 2D or 3D parameters" constitute a description of a *Spatio-Temporal Index* in terms of location, orientation and motion vector. The 2D parameters are easy to extract but do not make full use of the 3D model. The 3D parameters are used to match the 3D human model with the frame number n of the video. They include the reference point and the orientation of each component in the articulated structure.

1051-4651/94 $04.00 © 1994 IEEE

However, recovering 3D parameters from 2D images is a difficult problem so we use also 2D parameters.

The field "Description of the event" provides a description of a *Spatio-Temporal Index* in terms of keywords or symbolic representations. It permits a higher level of abstraction. For example, the sentence "head turned to the right" is constituted of a list of basic keywords and describe a spatial event. A symbolic representation may be an icon or any graphic representation of an event.

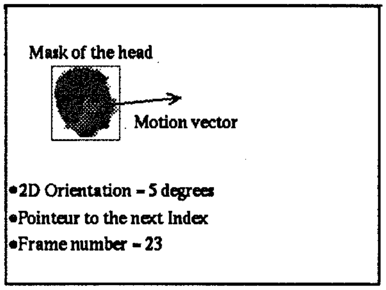

Figure 1: A *Spatio-Temporal Index* with 2D parameters

4 Generic Hypothesis

To work at a high level of abstraction, we must define the basic entity we use for the dynamic scene analysis : the *Generic Hypothesis.* It is represented by a simple matching between a subset of 2D regions located in a particular frame and a subset of 3D regions mapped on the 3D human model (Figure 2). This definition permits a model-based high level analysis while preserving a simple link with the extracted primitives.

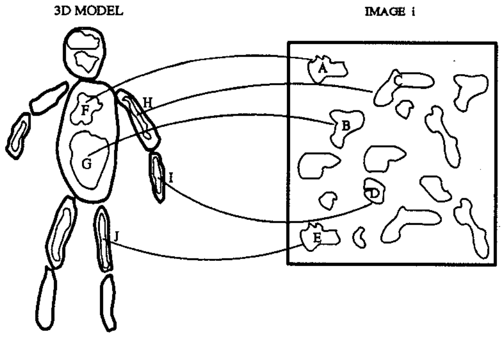

Figure 2: A Generic Hypothesis

To avoid problems of splitting-merging regions we allow a one-to-many mapping between 3D regions and 2D regions. We also allow unlabelled mapping to match an unknown region. In Figure 3, a *Generic Hypothesis* concerning the head is represented. The 3D region of the face is matched with two different 2D regions and there is one unlabelled region.

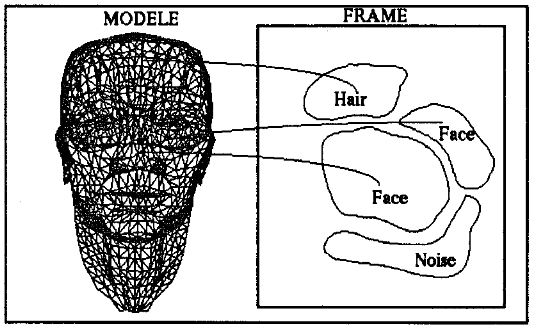

Figure 3: A Generic Hypothesis concerning the head

5 Analysis

We exploit scene-specific knowledge to guide the analysis process. This knowledge is formulated as a set of specific hypotheses that are verified throughout the analysis. The initial assumptions of our work are summarized in the following. The scene consists of a moving person in a complex background with possible occlusions. We use monocular video sequences and exploit color information. The components of the 3D human model exhibits rigid 3D motion.

The analysis of a raw video document is achieved by integrating six main steps :

1. Initialisation of *Generic Hypotheses.*
2. Creation of the *Search Space.*
3. Search of *Micro-Scenarios.*
4. Identification of a *Macro-Scenario.*
5. Post-processing of the *Macro-Scenario.*
6. Deducing *Spatio-Temporal Indices.*

Given one of the ten rigid components of the 3D human model, these steps are applied successively. This process yields a number of candidate interpretations of the motion of this component.

The first and second steps concern the initialisation phase. A set of good and plausible *Generic Hypotheses* is created and an appropriate *Search Space* is built. The third and fourth steps concern the model-based high level dynamic scene analysis. The spatial constraints of the 3D human model are used in conjunction with a search heuristic in order to identify a scenario describing the whole sequence. In the fifth step, we try to correct segmentation errors and improve the accuracy of the scenario. The last step concerns the description of the sequence at a high level of abstraction.

5.1 Initialisation of Generic Hypotheses

In the first step, we create a set of *Generic Hypotheses* for a particular frame of the video, based on the primitives previously extracted in some way. Using *a priori* information about the structure of the objects in the scene, we reduce the number of potential *Generic Hypotheses*, selecting and retaining only the most coherent. For instance, the *Generic Hypothesis* represented in Figure 5 is incoherent because of the relative position of the head and the leg. More precisely, we define a coherence criterion to filter all the most likely *Generic Hypotheses.* This criterion depends on the relative position of the 2D and 3D regions and on the similarity of the characteristic parameters between the 2D and 3D regions.

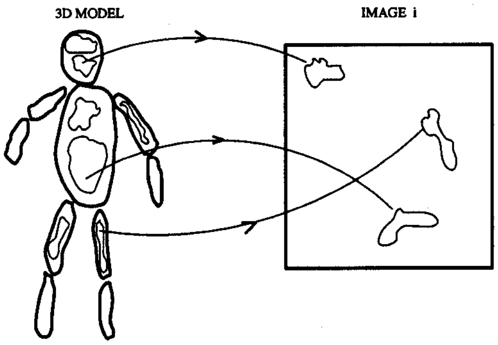

Figure 4: An incoherent Generic Hypothesis

5.2 Creation of the Search Space

The set of *Generic Hypotheses* created in the first step is not structured and therefore is not suitable for a high level analysis. In other terms, we must introduce a relation between two *Generic Hypotheses* in the same frame or in two consecutives frames. The second step allows to create such a hierarchy and store the *Generic Hypotheses* in a *Search Space.* This *Search Space* is a 3D graph as shown in Figure 4. Each plane represents one frame and the nodes are the *Generic Hypotheses.*

We make the distinction between two kinds of links. The intra-frames links represent a relation between two *Generic Hypotheses* in the same frame and symbolize a spatial constraint. The inter-frames links represent a relation between two *Generic Hypotheses* in two consecutive frames and thus symbolize a motion constraint.

The creation of the links is one of the most important steps in the analysis. To reduce the complexity of the search, we need to limit the number of links. This problem is adressed in the next step.

5.3 Creation of the Micro-Scenarios

A *Micro-Scenario* is a short time succession of *Generic Hypotheses* and is represented by a small path of inter-frames links as shown in Figure 4. Because of its short lenght, a *Micro-Scenario* is robust and easy to extract. Experiments show that a lenght of 3 or 5 frames is sufficient. We note *First*(S) and *Last*(S) the first and the last *Generic Hypotheses* of a *Micro-Scenario* S. $Frame(H)$ denotes the frame number associated with a *Generic Hypothesis* H.

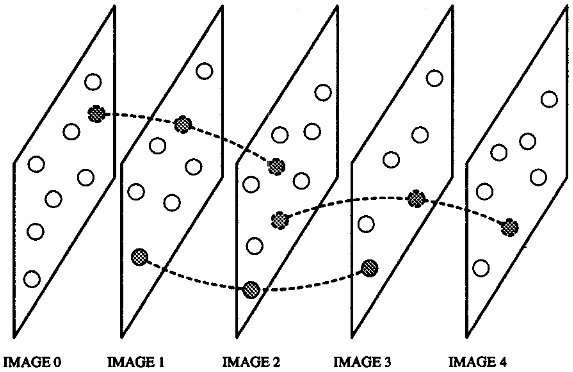

Figure 5: Some Micro-Scenarios

Micro-Scenarios are created by simple matching of adjacent *Generic Hypotheses* in the time domain. Two consecutive *Generic Hypotheses* are linked and integrated in the same *Micro-Scenario* if they are spatially close.

The set of *Micro-Scenarios* is sparse and the *Search Space* comprises a number of discontinuities or areas without *Micro-Scenarios* called *Holes.* We distinguish two kinds of *Holes.* A *First Order Hole* is a *Hole* between two *Micro-Scenarios* S_1 and S_2 with $Frame(Last(S_1)) = Frame(First(S_2))$. A *Second Order Hole* is a *Hole* between two *Micro-Scenarios* S_1 and S_2 with $Frame(Last(S_1)) < Frame(First(S_2))$.

In brief, *Micro-Scenario* expresses local and robust knowledge in the sequence as "Something here looks like the head and is moving from this point to this point between frames 10 and 13" for example. It is composed of a spatial knowledge (*Generic Hypotheses* concerning the head) and a temporal knowledge (motion of the head between frames 10 and 13).

5.4 Identification of a Macro Scenario

A *Macro-Scenario* is a long and complete time succession of *Generic Hypotheses*, describing a video cut in terms of consecutive events. It is represented by a long path of inter-frames links.

We grapple with the problem of the identification of a *Macro-Scenario* by navigation in the *Search Space*, searching the most coherent path from the first to the last frame. The idea is to find a global path from the first to the last image of a video cut which go down through the sparse set of *Micro-Scenarios* in a coherent way. Each time we go down a *Micro-Scenario* a pseudo-distance is used to determine the next *Micro-Scenario* to use. Then we solve the *Hole* problem to make a jump between the two *Micro-Scenarios.*

5.5 Post-processing

Because of the poor quality of the extracted primitives (regions), a post-processing step is necessary to correct the segmentation. This model-based correction is achieved at two levels. Firstly, a local and more accurate segmentation inside the bounding box of the component is performed. The background is locally removed and the outline of the component is improved. Secondly, the one-to-many mappings are transformed (if possible) into one-to-one mapping by merging 2D regions matched to the same 3D region.

5.6 Deducing Spatio-Temporal Indices

Deducing *Spatio-Temporal Indices* means that an appropriate method must be used to calculate the 2D or 3D parameters described in Section 3 and interpret them in terms of keywords or symbolic representations. Recovering 2D parameters from a *Macro-Scenario* is easy. The 2D orientation of a given component is approximated by the principal axis of the mask. First and second order moments may be extracted to describe the shape more precisely. A 2D motion vector is calculated between two consecutive frames.

Concerning the interpretation of 2D/3D parameters in terms of keywords, a simple partition in the parameters space is built. Each partition is associated to a keyword.

6 Conclusion and future work

As an initial effort towards the automatic annotation of video documents, high level analysis mechanisms combined with traditionnal scene analysis algorithms have been developed. We showed how to analyze human body motion and nearly automatically extract *Spatio-Temporal Indices* from a video sequence. Our approach, based on a model-based search heuristic, permits to avoid some problems inherent to the segmentation of primitives (splitting-merging techniques, poor quality of regions). A model-based post-processing step is used to correct a posteriori the shape of the extracted objects. The proposed *Macro-Scenario* describes the sequence in terms of basic events and constitute a first level of annotation.

We have implemented all the mechanisms described in this paper in a prototype package designed as a possible video indexing tool. At present, this tool is based on a graphic interface and a kernel of video analysis. Experiments have demonstrated a satisfactory accuracy and robustness with regards to occlusions and noise. We shall now consider other important aspects in future work, including the development of a complete computer tool for video indexing. In particular, we are investigating various approaches to developing an automatic presentation system based on image synthesis.

References

[1] Huang, T.S., *"Modeling, Analysis, and Visualisation of Nonrigid Object Motion", Proc. of International Conf. on Pattern Recognition*, Vol. 1, pp. 361-364, Atlantic City, NJ, June 1990.

[2] Johannson, G., *"Perception of Motion and Changing Form", Scandinavian J. Psychology*, Vol. 5, pp. 181-208, 1964.

[3] O'Rourke, J. and N.L. Badler, *"Model-Based Image Analysis of Human Motion Using Constraint Propagation", IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-2, No. 6, pp. 522-536, 1980.

[4] J. K. Aggarwal and N. Nandhakumar, *"On the computation of motion from sequences of images - A review", Proc. IEEE*, vol. 76, no. 8, pp. 917-935, Aug. 1988.

[5] A. Shio and J. Sklansky, *"Segmentation of People in Motion", Proc. IEEE*, vol. 76, no. 8, pp. 325-332, 1991.

[7] T. Broida, R. Chellappa, *"Kinematics and structure of a rigid object from a sequence of noisy images : a batch approach", Proc IEEE Conf. Comput. Vision Pattern. Recog.*, Miami, FL (1986), pp. 176-182.