Space-Time Tubes and Motion Representation

Christos Diou, Anastasia Manta, and Anastasios Delopoulos

Multimedia Understanding Group,
Department of Electrical and Computer Engineering,
Aristotle University of Thessaloniki - Greece
diou@olympus.ee.auth.gr, manta@auth.gr, adelo@eng.auth.gr

Abstract. Space-time tubes, a feature that can be used for analysis of motion based on the observed moving points in a scene is introduced. Information provided by sensors is used to detect moving points and based on their connectivity, tubes enable a structured approach towards identifying moving objects and high level events. It is shown that using tubes in conjunction with domain knowledge can overcome errors caused by the inaccuracy or inadequacy of the original motion information. The detected high level events can then be mapped to small natural language descriptions of object motion in the scene.

1 Introduction

While video motion analysis is a broad subject that has been extensively studied, most of the established approaches appear to be insufficient when it comes to semantic analysis of video data. They either provide low level information that is primarily useful for coding purposes, or are highly dependent on image processing results that lack the accuracy required for identification of natural objects or events. Mobile object detection and tracking techniques have demonstrated satisfactory results (e.g., [5]), but often heavily rely on the robustness and effectiveness of the image processing algorithms applied; the use of common sense rules and domain knowledge is usually limited and implicit, integrated in the tracking algorithm.

In this paper we introduce space-time tubes as a general concept and discuss the ways they can be used to identify high level events related to natural object motion. One of the most important benefits is that low level processing is abandoned early in the event detection process, while results are mainly obtained using reasoning that can accomodate domain knowledge.

Since tubes have certain properties that can be directly mapped to events such as "Two objects meet", it is also possible to construct simple natural langugage descriptions of the events detected in a scene (see section 3).

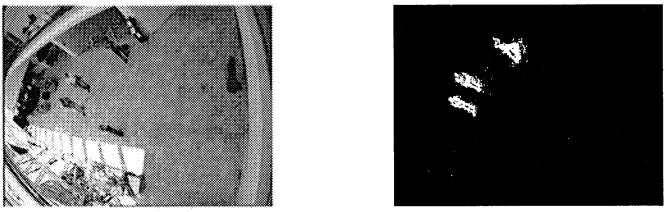
2 Space-Time Tubes

Assume that sensors detect motion in a scene, so that a binary motion mask $I_b(x, y, t)$ is provided:

Please use the following format when citing this chapter:
Diou, Christos, Manta, Anastasia, Delopoulos, Anastasios, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 583–590

$$I_b(x,y,t) = \begin{cases} 1 & \text{if motion is detected at point } (x,y) \text{ at time } t \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Note that moving points given by I_b correspond to the projection of moving objects on the sensor plane. When using a camera for example, the detected points result from processing frames that correspond to perspective projection at the camera projection plane and using the foreground extraction technique presented in [2] at a specific time (frame) is given in Figure 1(b).



(a) (b)

Fig. 1. A video scene and the corresponding binary mask for the moving points. The original image comes from the EC funded CAVIAR project [1].

The subset $S \subseteq \mathbb{R}^3$ of the moving points forms a topological space such that the function I_b defined above is the characteristic function of S. Moreover, every cross-section $S(t_0) \subseteq \mathbb{R}^2$ of S at time t_0 also defines a topological space, and the corresponding characteristic function is $I_b(x,y,t_0)$. An example is given in Figure 2, as obtained from a video sequence.

Any subset T of S that is connected and its cross-section $T \cap S(t_0)$ at t_0 is also connected for any t_0 is called a *tube*. A single connected component of S is called a *composite tube*, in the sense that it is formed by union of tubes. Note that for composite tubes and tubes, connectivity refers to $x-y-t$ space, while for their intersections at a specific time t, it refers to the $x-y$ space. Moreover, the above definitions allow tubes to have common elements. There are four main events that can be observed on tubes forming a composite tube:

1. *Start.* A tube starts at time t_0 if for every point $(x,y,t) \in T$, $t \geq t_0$. If tubes are maximal sets, i.e., they are the maximal sets that are connected and their cross-section $T \cap S(t_0)$ forms a connected component in $S(t_0)$, then their start points are the start points of the corresponding composite tube.
2. *Stop.* A tube stops at time t_0 if for every point $(x,y,t) \in T$, $t \leq t_0$. As with start points, if tubes are maximal sets, then their stop points are the stop points of the corresponding composite tube.
3. *Merge.* Two tubes T_1 and T_2 merge at point t_0 if their cross sections $T_1 \cap S(t)$ and $T_2 \cap S(t)$ are not connected for $t < t_0$ and are connected at t_0.

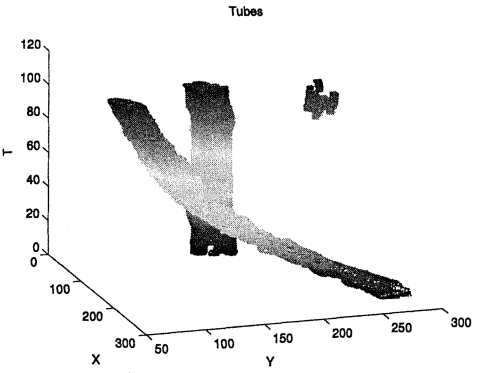


Fig. 2. The points detected in the $x-y-t$ space, as obtained by applying a foreground detection algorithm on a video sequence.

4. *Split.* Similar to merge, two tubes T_1 and T_2 split at point t_0 if their cross sections $T_1 \cap S(t)$ and $T_2 \cap S(t)$ are connected for $t < t_0$ and are not connected at t_0.

Merge and split points of tubes that are maximal sets are also merge and split points of the corresponding composite tube.

Tube segments are tubes that form a partition of a composite tube such that each one of them starts at a start, merge or just after a split event and stops at a stop, split or just before a merge event. Moreover, their cross section $T_s \cap S(t)$ at any time t forms a connected component of $S(t)$ (hence two tube segments can only be connected at their start or stop points). Given a composite tube, a number of possible tubes can be constructed. If we allow tubes to start or stop at any point (i.e., not restrict tubes to start or stop whenever a tube event occurs), there are infinite possibilities. However only one partition of tube segments can be constructed.

All of the above can be better explained using Figure 3, that shows sketches of what a projection of a composite tube on the $x - t$ plane might look like.

The above definitions depend on the topological properties of a given set of points S, however tubes also have certain geometric properties that are of interest, namely tube *centroid*, *area*, *velocity* and *duration*. For a tube T these four properties are functions of time.

The centroid can provide an approximation of the trajectory that the tube followed and for each time t equals the centroid of the set $S(t) \cap T$ of points. A similar property would be the tube *skeleton* that can be extracted via the use of a skeletonization algorithm, however this tends to be a computationally intensive process, compared to centroid calculation. A tube's area at time t is simply the area the tube occupies at that time. As far as velocity is concerned, both x and y axis velocity components v_x and v_y are the same for all points of a tube for a specific time t and are given by the tube's gradients $v_x(t) = \frac{dx}{dt}$

(a) (b) (c)

Fig. 3. (a) Sketch of composite tube, marking the events that occur. (b) Sketch of two possible tubes for the original composite tube, (c) Sketch of the corresponding tube segments.

and $v_y(t) = \frac{dy}{dt}$ for a specific point (x, y) (e.g., the centroid). Finally, a tube's duration is given by the difference $d = t_f - t_0$ of the stop and start times of the tube.

All the definitions above can easily be extended in the discrete case, if the sensor signal obtained is digital.

3 Mapping Natural Objects to Tubes

Tubes, or certain tube properties can be used to describe motion of natural objects in the observed scene i.e., find their trajectories and velocities and even lead to verbal descriptions of natural events such as "An object A entered the scene and moves fast" or "objects A and B meet".

Consider the ideal case, with sensor information provided being completely accurate and the binary mask I_b free of errors. Then, we can make the assumption that each moving natural object generates exactly one tube. A composite tube is generated when more than one objects move and their projections at the sensor plane meet at least once. Hence, given a composite tube that has a single tube segment (no merge or split) we can unambiguously determine the motion of the corresponding moving object. This is not the case in general, however: In Figure 3(a) there are multiple tubes that can form the initial composite tube and one example is given in Figure 3(b). If no restrictions are posed on the start/stop points and the area a tube can occupy, there is an infinity of possible natural events that would produce the same composite tube.

In most cases of practical interest a tube starts at an event of the corresponding composite tube i.e., a start, merge or split and stops at a stop, merge or split. We can therefore find a finite number of possible tubes for a given composite tube, a problem similar to finding all the connected subgraphs of a graph where each node will correspond to a tube event and each link to a tube segment. Additionally, subgraphs that are mapped to temporally concurrent tube segments are rejected e.g., TS_1 and TS_2 of Figure 3(c) cannot form a single tube (the corresponding natural object would be at two places at the same time).

Hence, for a given error-free composite tube this approach leads to a number of possible events. Tube features such as area or velocity may be used to determine the event that is most likely to have happened, but in general more features such as color or texture will need to be obtained in order to rank these events while the result will also depend on the application domain. This approach has the following advantages: (i) Tubes are used as a feature that can be employed by definitions of natural events in a knowledge base [3]. (ii) The possible events that are examined are restricted in number and identified. (iii) Processing with other features is optional but can greatly increase the accuracy of the inference process and enable balancing between complexity and validity of the results as developed in [3, 4] (iv) If other feature extraction algorithms are applied, tubes can provide the region of interest.

By assigning verbs to events (e.g., "meet" for merge, "part" for split), and designating each natural object with an alphanumeric label, it is possible to map the detected natural events to natural language descriptions as described in [6]. If objects are known or identified then their labels are replaced by their name or property.

4 Ambiguities due to Errors

In real-life applications sensor information will often be inaccurate and will lead to errors, due to imperfections of the devices and algorithms used. Additionally, there exist certain errors that are introduced when dealing with two-dimensional signals that describe three-dimensional scenes (e.g., occlusion). Certainly, the assumption that each moving object generates exactly one tube is not valid in that case and a tube preprocessing stage must be introduced before proceeding to examination of natural events.

There are three main errors that can be observed in tubes with respect to natural objects:

1. *Temporal discontinuity.* A single natural object may generate more than one tubes due to occlusion or other factors. An example is given in Figure 4(b), where two tubes are generated and correspond to a single tube given in Figure 4(a).
2. *Spatial discontinuity.* A single natural object generates two or more concurrent tubes, because parts of the object were detected as different moving objects. An example of the combination of this and the previous error is given in Figure 4(c).
3. *Noise.* A tube is generated where no moving object exists. This is common in algorithms that determine the motion mask of a video sequence, where sudden changes in the lighting conditions lead to detection of regions that do not correspond to moving objects in the original scene.

A tube preprocessing stage can be introduced to compensate for these errors based on how objects are expected to behave in the given application domain.

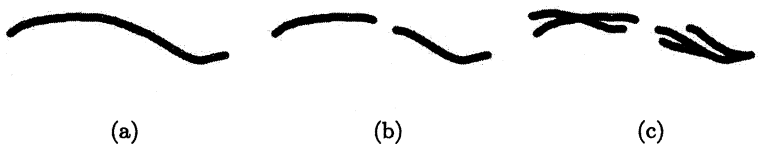


Fig. 4. Tube errors. (a) Correct tube. (b) Temporal discontinuity (c) Spatial and temporal discontinuity.

Generally, each composite tube or set of neighboring composite tubes is transformed into composite tubes that have "lower resolutions". A weight is assigned to each of them, designating the degree up to which they approximate the ideal scenario of section 3. Its value is determined using a set of metrics on tubes e.g., the distance or displacement vector between a stop and a consequtive start event (so as to identify temporal discontinuities) and a correpsonding fuzzy membership function [7]. For example, the tubes of Figure 4(c) can be transformed into the tubes of Figures 4(b) and 4(a).

For each transformed composite tube, there exists a different set of natural events, as in section 3. Based on the weight of each transformed composite tube a certainty value is assigned to each of the natural events, denoting our degree of belief that this event is what actually happened. Additional features can then be used to increase or reduce this certainty value and rank the possible outcomes.

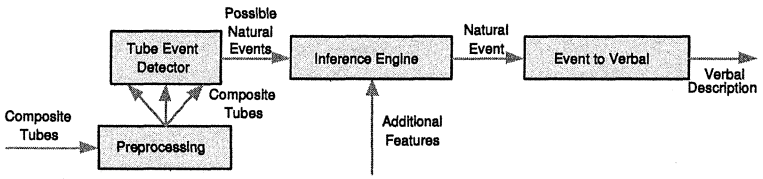


Fig. 5. The stages of identifying and assigning verbal descriptions to natural events using tubes.

Figure 5 presents a simplified block diagram that summarizes the use of tubes in extracting verbal descriptions based on motion in a scene. Note that there are two main sources of ambiguity in this process: (i) Image processing errors that are dealt with in the preprocessing stage and (ii) the one-to-many mapping of a composite tube to natural events. In the latter case the inference engine uses domain information as well as additional features to reason about the event that is most likely to have occurred.

5 Experiments

In order to evaluate the use of tubes in motion analysis and detection of natural events, a number of experiments were conducted using video sequences obtained from a static camera, mainly based on the datasets provided by [1]. The foreground mask is extracted using two techniques, the first one [8] is fast but rather error prone when there are sudden changes in the lighting conditions while the second [2] is more accurate but computationally intensive. In both cases filtering and morphological operations served as an initial processing stage.

A simple set of rules was used to transform and remove errors from the extracted composite tubes. These rules were based on the tube segments' duration and area as well as the distance between successive stop and start events. Tube segments with small duration or very small area were removed or merged with other tube segments depending on whether they formed a composite tube on their own. If such a tube segment was also a composite tube it was considered noise and was removed, otherwise it was merged with a tube segment with longer duration or larger area to avoid spatial discontinuities. Furthermore, composite tubes that were very close to each other were united to avoid temporal discontinuities.

Figure 6 shows an example taken from a 300-frame video sequence. Note that using rules to transform the original tubes removes the errors caused by inaccurate information obtained through image processing operations. In most experiments, the results from both foreground extraction algorithms were similar, even though the foreground mask provided by [2] was far more accurate.

6 Conclusions

Space-time tubes, a novel feature that can be used to analyze motion information, was presented and the stages required to obtain semantic-level natural language descriptions regarding events in the observed scene were outlined. Experiments that were carried out demonstrated how the use of tubes and tube processing can overcome image processing errors that would otherwise lead to false conclusions in event detection within video sequences. The main benefits of using tubes lie on the fact that information about natural events is obtained through knowledge based reasoning and rules, not based on raw sensor information or low level processing results that tend to be inaccurate. Furthermore, tubes can be used in conjunction with other features independently, thus allowing for smooth integration to a general reasoning framework.

References

1. EC Funded CAVIAR project/IST 2001 37540 `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`.

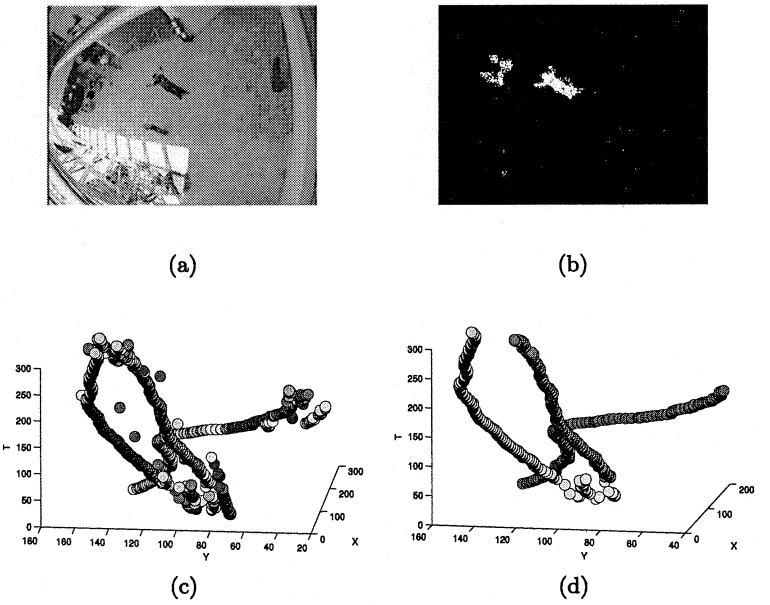


Fig. 6. (a) Original image (b) Foreground mask (c) Centroids of the tubes before processing. Different colors indicate tube segments. (d) Centroids of processed tubes.

2. Ahmed Elgammal, David Harwood, and Larry Davis. Non-parametric model for background subtraction. In *Proceedings of the 6th European Conference on Computer Vision*, pages 751–767, 2000.
3. M. Falelakis, C. Diou, A. Valsamidis, and A. Delopoulos. Complexity control in semantic identification. In *IEEE International Conference on Fuzzy Systems, Reno, Nevada, USA*, May 2005.
4. M. Falelakis, C. Diou, A. Valsamidis, and A. Delopoulos. Dynamic semantic identification with complexity constraints as a knapsack problem. In *IEEE International Conference on Fuzzy Systems, Reno, Nevada, USA*, May 2005.
5. Ismail Haritaoglu, David Harwood, and Larry Davis. w^4: Real-time surveillance of people and their activities. *IEEE Transactions on pattern analysis and machnine intelligence*, 22(8), August 2000.
6. Gerd Herzog and Peter Wazinski. Visual translator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review*, 8(2-3):175–187, March 1994.
7. George J. Klir and Bo Yuan. *Fuzzy Sets and Fuzzy Logic; Theory and Applications.* Prentice Hall, 1995.
8. Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on pattern analysis and machnine intelligence*, 19(7):780–785, July 1997.