Action Recognition using Time of Flight Cameras

Serban Oprisescu, Constantin Burlacu, Vasile Buzuloiu
Image Processing and Analysis Laboratory
University POLITEHNICA of Bucuresti
Bucuresti, Romania
soprisescu (at) imag.pub.ro

Abstract—The paper deals with the issue of action recognition as an application of the new 3D time-of-flight (ToF) camera, exploiting the special ability of the device to measure distances. Segmentation of moving people is straightforward from the distance information and subsequent steps of the processing chain follow in a classical way. We describe the first results on action recognition using ToF camera distance images for a simple task of deciding actions of a single person. The total variation of a function reveals as a very useful feature in these applications.

Keywords-component: ToF cameras, action recognition, video and image processing

I. INTRODUCTION

Action recognition is one of the main applications of computer vision and at the same time one of the most difficult to realize due to the complexity of tasks and also to time constraints when we think to real-time recognition. Of course, it supposes action categorization, motion analysis and before these complex off and on-line operations, the reliable segmentation of the objects of interest in the movie. One can find excerpts of references to our subject in various papers among which we mention [1] and [2].

We have put as our task to develop action recognition applications in the framework of using Time of Flight (ToF) cameras. This is a very important constraint, which must radically simplify the segmentation / object extraction from the scene.

ToF cameras [3] work with their own light source - an infrared amplitude modulated light which, reflected from the objects in the scene, arrives back on the camera sensor array - and concomitantly delivers two (grey level) images: one is the usual amplitude image and the second is the distance image. In spite of the fact that with nowadays ToF-camera one can develop applications where even submillimeter movements could be sensed, the distance error could be high (0.5m) depending on the scene as well as on the camera optics and geometry. That is why at the beginning we focused on finding the way to correct these distance errors ([3][4]). Now we are sure that the next generation of ToF cameras will deliver correct distance images and so the extraction of the object of interest (usually a person) from the image/movie will become a simple task (of segmentation by thresholding).

The paper is organized as follows: §II describes the ToF camera principle; §III gives a brief presentation of some recent achievements in methodologies as well as in concrete systems for action recognition; §IV describes our chain of tasks/operations involved in action recognition; §V reports on our work for various tasks and §VI contains conclusions and directions of future work. In a previous paper [5] we tackled the problem of recognition of combined actions.

II. TIME OF FLIGHT CAMERAS

In Fig. 1 the principle of a ToF camera is drawn. The infrared light modulated with e.g. 20MHz is sent toward the scene and, after reflections from objects, returns on the sensor array of the camera where a phase detector delivers at the pixel i a (quasi-)sinusoidal signal:

\[ I(i) = a(i) \cos(\phi(i)) \]

whose amplitude \( a(i) \) is proportional to the object reflectivity and the phase \( \phi(i) \) is proportional to the distance \( d(i) \) of the object to the camera, \( \phi(i) = 4\pi fd(i)/c \).

In this formula \( f \) stands for the modulation frequency (e.g. 20MHz) and \( c \) for the speed of light. It is clear that the range \([0,2\pi]\) is covered by a well defined range in distance (in our case \( c/2f = 3.108/(4.107) = 7.5 \text{ m} \)).

This ToF phase-measurement principle is used by several manufacturers of ToF-cameras, e.g. PMDTec/ifm electronics (www.pmdtec.com), Mesa Imaging (www.mesa-imaging.ch) and Canesta (www.canesta.com). We used in our experiments a SR3000 Mesa Imaging ToF camera, having an image resolution of 176x144 pixels, capable of maximum 29 frames per second.

Figure 1. The principle of the Time of Flight camera
III. STATE OF THE ART

Action recognition means human motion analysis. General reviews are e.g. [6] and [7]. A pioneer in the domain is J.K. Aggarwal. An important point is that a system must be designed to accomplish some precise tasks. In [8] the human motions are categorized in: single person motion, interactions between people or activity involving objects, in order to build a system that detects human climbing fences. The algorithm has two major parts: human body representation and time series analysis. An extended star-skeleton representation is used for extracting the predicates to form the feature vector for every frame. The resulting time series is encoded via a hidden Markov model trained to obtain the state sequence.

More recently, some authors [9] claim advantages for a new approach (“actions as space-time shapes”). Despite good results achieved in action recognition by “traditional approaches” they impute them either difficulties in estimating optical flow or difficulties in employing feature tracking. They cite papers which show it is useful to analyze actions by looking at a video sequence as a space-time volume. On the other hand studies in the field of object recognition in 2D images have demonstrated that silhouettes contain sufficient information to identify the object. Another recent paper [10] uses an approach to analyze 2D shapes based on a solution to a Poisson equation which results in a scalar field that allows extracting many useful properties of a shape including aspect ratio of different parts, in a simple way. In [9] it is this method for the analysis of 2D shapes which is generalized for volumetric space-time shapes induced by human actions. An important argument for this approach in the view of the authors is that “unlike images, where extraction of a silhouette might be a difficult segmentation problem, the extraction of a space-time shape from a video sequence can be simple in many scenarios”.

Another recent paper of interest for recognition of complex human activities [11] develops a model of activity which consists of pieces relatively easily learned which are combined together within a model of composition. At present we shall not enter in the analysis of such complex activities. It is enough to say that the use of ToF cameras simplifies the tasks for any previous models and we foresee future studies to integrate our work in more complex models.

IV. A SIMPLE CHAIN OF TASKS

At a first glance it is almost obvious that the first step in a chain of tasks for action recognition must be segmentation or silhouette extraction from each image of a video sequence, which in our framework (using ToF cameras) is simple, robust and fast. We shall stress that ToF cameras don’t have an unlimited view space (e.g. 0.5-7.5m), but this is not an essential constraint - even using normal camera one reaches some limits of functionality due to other reasons. Further more, one can use the distance images or the amplitude images; using the later one has more possibilities in choosing the key points.

The second step is to simplify as much as possible the “space-time shape” obtained by temporal concatenation of 2D silhouettes extracted in the segmentation phase. At present, by simplicity reasons, we only experimented the simplification at the 2D level, choosing key points in each silhouette, such that they can be easily tracked along the segmented video sequence. Each silhouette is reduced to a few points either by skeletonization and retention of some extreme points or by the retention of either some salient points (in this case using amplitude images) or some geometrical specific points (like centre of gravity etc.). Of course in this step, the choice of the key points is essential and we shall report elsewhere the conclusions of various experiments. Among the methods used to detect the candidates as salient points in the amplitude image, we used steerable filters [12]. The result of the second step is replacing the space-time shape with the trajectories of the key points.

In fact, in the approach above one is also confronted to the tracking of the key points. In some cases this becomes a relatively straightforward and simple task; it is also a difficult one in others, and again, we shall not debate these ones here.

So, suppose we have the trajectories of the key points as vector functions of time \( r_i(t) \) \( i = 1,...,N \) where \( r_i = (x_i, y_i, z_i) \) is the position of the key points in the 3D space of the camera. Obviously we have \( 3N \) scalar functions of time for \( N \) key points. These functions which stand for absolute positions could be advantageously replaced by other: e.g. if \( r_i \) stands for the centre of gravity of the silhouette and we prefer to record the relative movements to it of the various points of the silhouette (as more expressive compared to absolute movements) we shall replace \( r_i \) by \( r_i - r_0 \), etc.

The next step in simplifying the description of the action is to replace the trajectories with some features of them. Let any feature be characterized by a scalar. Then, with \( m \) features for a trajectory, the action will be characterized by \( 3Nm \) scalars i.e. a point in a \( 3Nm \)-dimensional space.

A very simplified scheme of this model could have \( m=2 \) (e.g. mean value and total variation of the function) 2 instead of 3 (that means motion in 2D space) and \( N=5 \) (not all extreme points of the skeleton and perhaps, the centre of gravity); we get already 20. Using enough training sequences one could reach the segmentation of the 20-dim space in regions for quite many actions. Of course, a wise selection of the key points together with a pre-processing of the trajectories can lead to essential improvements of the performances of a real system.

V. RESULTS

We used for our experiments the SR-3000 ToF camera. At each frame of the movie we have the amplitude gray-level image (Fig. 2a) and also the distance image like in Fig. 2b.

The first step in the processing chain (the segmentation of the silhouette of the person) is straightforward and is done by thresholding the distance image; the result is in Fig. 3a. The segmentation works on any non-uniform background (usually the background is farther than the person). We applied also some morphological filtering, as a preprocessing for the next step - skeletonization. For computing the silhouette’s skeleton we used an optimized algorithm [13].

The experiments reported here are done for the simple case of a single person in the scene, for decision among simple actions (i.e. not activities composed from more actions successive in time).
We have to decide between six different actions and the result can be either one of them or a response “unknown”; these 6 actions are: walk, carry, run, bent, jump and box.

In these first experiments we used skeletonization for obtaining in a simple way some key points. Namely, the simplest way to identify key points in this setting is to take the termination points of the skeleton plus the object centroid. To automatically identify the body part to which a termination key point belongs, we propose a skeleton radius descending sort, which will ensure that the head is always the first key point (in fact the second, because we start with the centroid). But, because the limbs have similar radius in the images, we implemented an initialization based on $y$ coordinate of the other four key points (usually the hands are upper, at least in our possible set of actions). This rule is plotted in Fig. 3b.

The tracking of our key points is a relatively simple task (with a search in the neighborhood of the previous position) and so we get the key point trajectories. In fact further we only use the trajectory of the centroid and the trajectories of the relative motion of the other key points against the centroid (it has more significance in deciding the type of action: for instance, the hands and legs will have a go-and-return type of motion relative to the centre of the object and the head will remain fixed if the subject is walking or running etc.).

The trajectories cannot be used efficiently as they are (even if there are works on this approach [14]) and we define several features computed from them e.g.:

1) The variation of the function:

$$\Delta f = f_{\text{Max}} - f_{\text{Min}}$$  \hspace{1cm} (1)

2) The total variation $\nu$ of the function:

$$\nu = \sum_{k=2}^{N} |f(k) - f(k-1)|$$

where $N$ is the number of frames, $k$ is the frame index and $f$ is the key point coordinate, on $x$, $y$ or $z$.

3) The real mean speed:

$$S_r = \frac{1}{N} \sum_{k=2}^{N} \frac{f(k) - f(k-1)}{t}$$

where $t$ is the time between two frames. $S_r$ is the mean of instantaneous speeds for each frame.

4) The absolute mean speed (the mean of the absolute value of each instantaneous speed):

$$S_a = \frac{1}{N} \sum_{k=2}^{N} |f(k) - f(k-1)|$$

From the first two features or from the last two one can easily distinguish between a motion in one direction and go-and-return one but also more subtle differences like waltzing or other type of a dance. The above formulas are valid for any projection of the trajectory: $x(k)$, $y(k)$ or $z(k)$ but in the examples we present here we only used $x(k)$ and $y(k)$. The distinction between the six actions named above, using the features 1 and 2 or 3 and 4 is straightforward. In fact, instead of a multidimensional picture we define simple rules for decision (see below) which are equivalent to the partition of the hyperspace of features.

In Fig. 4 we reproduced one frame of each movie for the 6 actions. In Fig 5 the graphics corresponding to $x(k)$ of the centroid or $y(k)$, are drawn. In Fig. 6 the graphics corresponding to $x'(k) = x(k) - x_{\text{centroid}}(k)$ are drawn for one of the key point distinct of the centroid. If the multidimensional point is put in the hyperspace it is obvious that we have a perfect separation. For this simple case, instead of defining the partition of the hyperspace directly, we perceive it in a simpler mode by some decision rules like in the decision tree (Fig 7).
Sax and Say are absolute mean speeds on x and y and Sx is the mean speed on x for the centroid; SxH is the absolute mean speed of a hand; TvsHx is the total variation of the speed of a hand on x; Vy is the total variation of the centroid on y. T1,…, T14 are thresholds computed in a learning phase (e.g. on a walk sequence). If the action is not recognized the label "?" = unknown is given.

In Table 1 is the confusion matrix computed by testing the recognition algorithm on 60 sequences, 10 sequences per action.

### Table 1. Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>walk</th>
<th>carry</th>
<th>run</th>
<th>bent</th>
<th>jump</th>
<th>box</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>walk</td>
<td>0.9</td>
<td>0.0</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>carry</td>
<td>0</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>run</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bent</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>jump</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>box</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### VI. Conclusions and Future Work

The framework of our approach is mainly limited by the restrictions imposed by the ToF camera i.e. limiting the volume of space of our scenes. Our approach is characterized for this incipient stage by the use of key points, their trajectories and some appropriately chosen features of these trajectories. Depending on the concrete application, one can develop a true real time system. We shall continue extended studies to compare the efficiency of our approach with others from the point of view of speed, robustness, error probabilities and versatility. In this phase we did not touch the problem of self learning the actions. We stress the usefulness of the total variation of a function as a type of features.

### ACKNOWLEDGMENTS

This research was accomplished in the framework of the European project "Action Recognition and Tracking based on Time-of-Flight Sensors" (ARTTS), funded by the European Commission (contract no. IST-34107) - the 6th Framework Program. This publication reflects only the views of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained herein.

### REFERENCES


