TRICTRAC Video Dataset: Public HDTV Synthetic Soccer Video Sequences With Ground Truth.

X. Desurmont[1], J-B. Hayet[2], J-F. Delaigle[1], J. Piater[2], B. Macq[3]

[1] Multitel ASBL, Parc Initialis – Avenue Copernic, 1, B-7000, Mons, Belgium
[2] Institut Montefiore, University of Liège, Building B28, B-4000 Liège, Belgium
[3] Communications and Remote Sensing Laboratory, UCL, Louvain-la-neuve,Belgium

Abstract. Object tracking in video sequences is an important task in many applications such as video surveillance, traffic monitoring, marketing and sport analysis. In order to enhance these technologies, an objective performance evaluation is needed. This evaluation requires to test the system with a given dataset and compare the output with the ground truth. One of the contributions of the TRICTRAC project is the supply to the video processing community of synthetic, high-definition video content of Pan-Tilt-Zoom (*PTZ*) cameras with 3D ground truth including the parameters of the cameras and the mobile objects. This paper presents this novel dataset.

1. INTRODUCTION

Testing video-analysis algorithms is very important in the academic and industrial communities. To enable performance evaluation, multiple steps are required. First, video sequences must be available. Secondly, the Video Content Analysis (VCA) technique to be evaluated must generate results on the test sequences. Thirdly, Ground Truth (GT) needs to be available. Then, the GT needs to be compared with the generated results, requiring an unambiguous definition of the metrics. Finally, the evaluation results are combined for each video sequence to be presented to the user. The Performance Evaluations of Tracking and Surveillance (PETS) workshop deals with these issues since 2000. It proposes datasets for surveillance, sports, smart meetings, etc. However, there are no datasets for Pan-Tilt-Zoom (PTZ) cameras with full camera ground truth available.

In this paper, we focus on the synthetic video datasets and their associated GT produced during the TRICTRAC [1] project. This is a scientific project that aims to build a tracking system for networks of multiple PTZ cameras. The three principle steps are to track objects in each video stream, to recover PTZ parameters in real time, and to merge the recovered camera and object parameters to build a 3D representation of the dynamic scene.

The paper is organized as follows: Section 2 discusses related work. Section 3 presents the modelling of the scene, objects, cameras and scenarios, and Section 4 explains the synthetic video and the ground truth format.

2. RELATED WORK

To enable proper benchmarking with respect to other algorithms, it makes sense to evaluate algorithms with standard video data. Moreover, carefully-prepared video test sequences are required for meaningful performance characterisation of VCA systems. They should be representative and contain both typical and worst-case scenarios. Apart from the video content, the characteristics of the image sensor, resolution and frame rate have impact on VCA performance. For this purpose, various test video datasets have been made publicly available via the Internet. They deal with different applications such as surveillance, biometry and multimedia.

The PETS[1] workshop series have introduced standard datasets in 2000 and during the following years. Thus, people can test tracking algorithms on the same datasets. Other publicly-available test datasets include the CAVIAR[2] [2] and the Terrascope [3] data.

The use of PTZ cameras in the test dataset was proposed for PETS 2005 with coastal surveillance. However, no ground truth of camera PTZ parameters was available. The TRICTRAC datasets presented here are unique because performance evaluation is possible not only for tracking in the 2D images but also for the positions in 3D and the PTZ camera parameters. The application of this dataset is soccer games. These were already addressed by the INMOVE[3] project but with fixed cameras (Fig. 1).

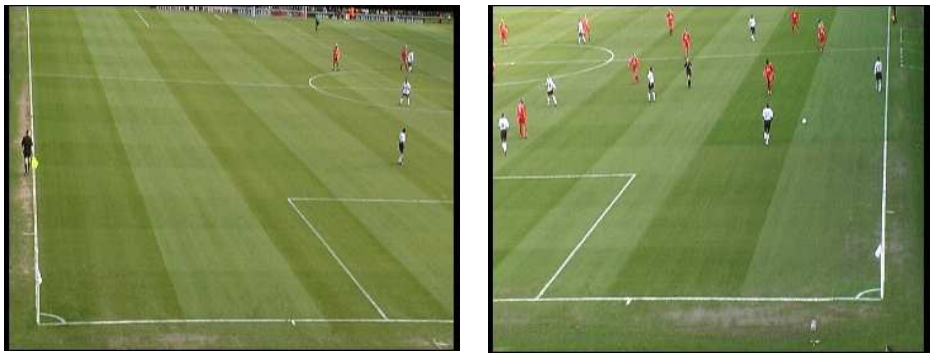

Fig. 1. Inmove PETS sequences.

Black *et al* [4] proposed pseudo-synthetic video as an alternative to manual ground truthing. The main idea is to construct a comprehensive set of pseudo-synthetic video sequences from a real background and segmented moving objects. The main drawback is that the resulting videos may contain incoherent visual effects (e.g., ambient lighting conditions might be different for different rendered objects.)

---

1 PETS test sequences: http://www.visualsurveillance.org/

2 CAVIAR: Context Aware Vision using Image-based Active Recognition: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

3 Project IST INMOVE (IST-2001-37422): http://www.inmove.org

2. Scene and scenarios.

In this section, we present examples of synthetic images from the TRICTRAC dataset.

2.1 Scene

The scene is a soccer stadium (Fig. 2). It is composed of contextual objects (ground, goals, audience, lights) and mobile objects (players and the ball, video cameras). Cameras are located around the grass field and one is situated in a higher point of view (Fig. 3). One can see in Fig. 4 the different views of the cameras. Except for the last image, all views track the ball. For Camera 9 there are two possible fields of view (one of them is very wide to see the entire stadium). The 3D position of each camera is given in Table 1.

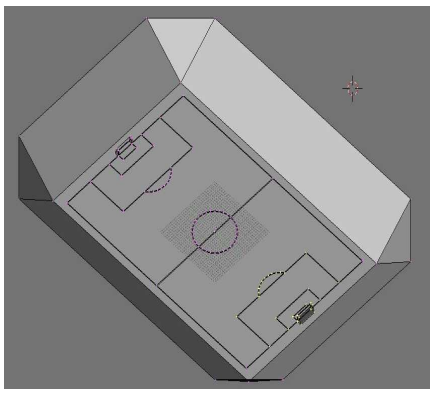

Fig. 2. Synthetic view of the stadium.

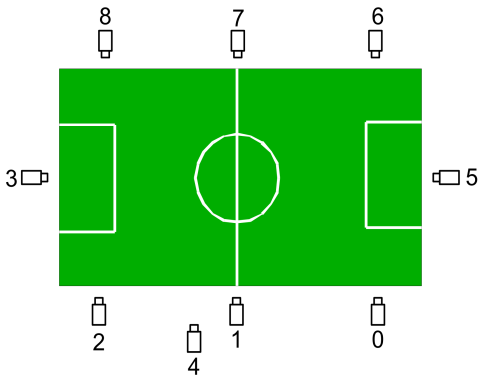

Fig. 3. Location of the cameras. Camera 9 (not shown) is located above the center of the stadium at an altitude of 100 m.

2.2 Scenarios

Soccer is one of the most popular sports in the world. It is a team sport played between two teams of eleven players each. It is a ball game played on a rectangular grass field with a goal at each end. The aim of the players is to score by putting the ball into the opposite team's goal. Only the goalkeepers can use their hands to propel the ball. In the video sequences we provide, scenarios are quite simple. Each sequence lasts around 20 seconds and ends with a goal scored by the red team. Each scenario is observed by several cameras that might be fixed or PTZ. Some key frames are available on the website.

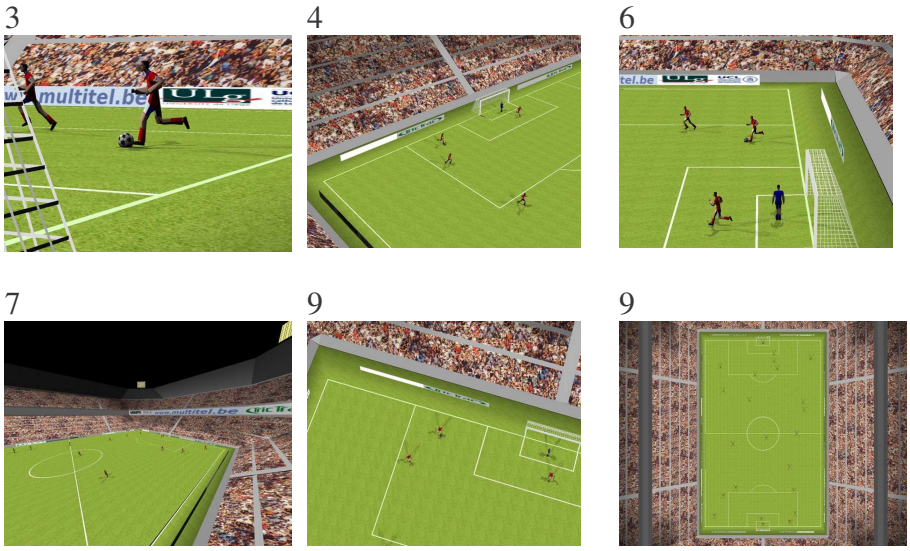

Fig. 4 . Views from different cameras.

Table 1. 3D position of cameras.

Camera id	X (m)	Y (m)	Z (m)
camera0	-40	15	-50
camera1	-40	15	0
camera2	-40	15	50
camera3	0	1.5	57
camera4	-50	30	10
camera5	0	1.5	-57
camera6	40	15	-50
camera7	40	15	0
camera8	50	15	50
camera9	0	100	0

3. VIDEO SEQUENCES AND GROUND TRUTH

3.1 Video Material

The imagery is 1400x1050 progressive scan. One can either convert it to 720p or 1040p. It is compressed in JPEG format (about 200 KB per image) and the frame rate is 25 fps. Images are also available in 640x480 and in 320x200 in avi video files. Table 2 summarizes the data. The whole dataset, in high quality, amounts to about 8 GB.

All data is publicly available on the Web and can be downloaded from http://www.multitel.be/trictrac/?mod=3. If you publish results using the data, please

acknowledge the data as coming from the TRICTRAC project, found at URL: http://www.multitel.be/trictrac.

3.2 Rendering

The rendering of the images is done via the Ogre 3D[4] library (Ver 1.04) in C++ that is LGPL. Ogre 3D is available on Windows and Linux. The main program can operate in real time for small images.

We used this library from our own C++ code that implements basic behaviour of soccer players who try to score. The display is rendered with OpenGL on accelerated video graphics hardware, and is then uploaded to the central memory in RGB format to be compressed into JPEG format. The rendering uses shadow effects (see Fig. 5) and there is no anti-aliasing processing in the 1400x1050 images.

Table 2. Scenarios, cameras and size of each video sequence

Sequence	Scenario	Camera	Fov in degrees	Jpegs 640x480	Jpegs 1400x1050 .tar	Avi file	Ground truth
2	1	4	15	97 MB	464 MB	8 MB	5 MB
3	1	3	15	133 MB	508 MB	10 MB	6 MB
4	1	6	15	99 MB	390 MB	8 MB	5 MB
5	1	7	100	95 MB	507 MB	10 MB	5 MB
6	1	4	100	127 MB	555 MB	10 MB	7 MB
7	2	3	100	123 MB	646 MB	10 MB	7 MB
8	3	4	30	139 MB	723 MB	10 MB	7 MB
9	3	6	15	105 MB	440 MB	8 MB	6 MB
10	3	7	100	108 MB	565 MB	9 MB	6 MB
11	3	3	15	140 MB	536 MB	10 MB	6 MB
12	3	9	15	124 MB	615 MB	10 MB	7 MB
13	3	9	65	120 MB	650 MB	7 MB	5 MB

Fig. 5. View of shadows around players.

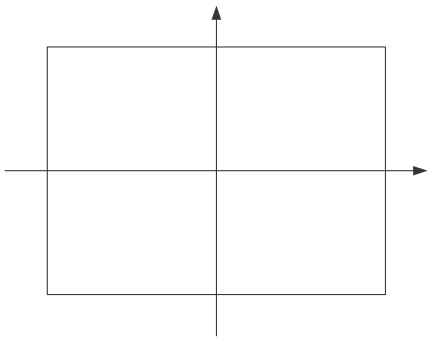

Fig. 6. Relative 2D image coordinates.

[4] http://www.ogre3d.org/

3.1 Ground Truth

The GT of the dataset is created automatically during the rendering. The GT only describes mobile objects. It consists of the parameters of the active cameras, the positions in 3D and images of the players. A typical frame description is shown in Fig. 7.

The world coordinates are expressed in meters (the center of the stadium is (0,0,0), the ground plane is Y=0). The screen coordinates are expressed as relative positions, $-4/3 < x < 4/3$ and $-1 < y < 1$ as shown in Fig. 6. Thus, pixel is not Necessary Square.

The time is expressed as `yyyy/mm/dd@hh:mm:ss.microseconds`, and there is a bijection between time and frame number (every 40 milliseconds, there is a new frame).

```
<?xml version="1.0" encoding="UTF-8" ?>
<event_history>
      <event>
            <time>2005/07/08@17:46:57.200000</time>
            <name>camera</name>
            <parameters>
                  <activecamera>4</activecamera>
                  <projection_m>5.69682,0,0,0,0,7.59575
                  ,0,0,0,0,-0.99999,-1.99999,0,0,-
                  1,0</projection_m>
                  <view_m>0.390245,-2.90758e-
                  09,0.920711,28.7194,0.340928,0.928917
                  ,-0.144503,-12.2661,
                  0.855264,0.370288,0.362506,50.2468,0,
                  0,0,1</view_m>
            </parameters>
      </event>
      <event>
            <time>2005/07/08@17:46:57.200000</time>
            <name>tracking</name>
            <parameters>
                  <id>0</id>
                  <class>person</class>
                  <worldposition>19.2547,0,-
            39.8366</worldposition>
                  <screenposition>0.0416068,0.00513261<
            /screenposition>
            </parameters>
      </event>
…/…
</event_history>
```

Fig. 7. XML format of the GT of TRICTRAC Dataset.

The position of the object (in 2D and 3D) is on the ground plane (z=0). It is also the position of the human 3D model used by Ogre3D. Note that all the people in the scene are described in the ground truth data, even if they are not in the image. In this case, the "screenposition" is outside of the image. The following equations give the relations between the 3D positions and the 2D positions of the players.

<projection_m>$P_{0,0}$, $P_{0,1}$, $P_{0,2}$, $P_{0,3}$, $P_{1,0}$, $P_{1,1}$, $P_{1,2}$, $P_{1,3}$, $P_{2,0}$, $P_{2,1}$, $P_{2,2}$, $P_{2,3}$, $P_{3,0}$, $P_{3,1}$, $P_{3,2}$, $P_{3,3}$</projection_m>

$$Projection M = \begin{bmatrix} P_{0,0} & P_{0,1} & P_{0,2} & P_{0,3} \\ P_{1,0} & P_{1,1} & P_{1,2} & P_{1,3} \\ P_{2,0} & P_{1,1} & P_{2,2} & P_{2,3} \\ P_{3,0} & P_{3,1} & P_{3,2} & P_{3,3} \end{bmatrix}$$

<view_m> $V_{0,0}$, $V_{0,1}$, $V_{0,2}$, $V_{0,3}$, $V_{1,0}$, $V_{1,1}$, $V_{1,2}$, $V_{1,3}$, $V_{2,0}$, $V_{2,1}$, $V_{2,2}$, $V_{2,3}$, $V_{3,0}$, $V_{3,1}$, $V_{3,2}$, $V_{3,3}$</view_m>

$$ViewM = \begin{bmatrix} V_{0,0} & V_{0,1} & V_{0,2} & V_{0,3} \\ V_{1,0} & V_{1,1} & V_{1,2} & V_{1,3} \\ V_{2,0} & V_{1,1} & V_{2,2} & V_{2,3} \\ V_{3,0} & V_{3,1} & V_{3,2} & V_{3,3} \end{bmatrix}$$

<worldposition>X , Y, Z</worldposition>

$$WorldPosition = \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

$$screenPosition = \begin{bmatrix} 4/3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \times Projection M \times ViewM \times WorldPosition$$

Table 3. Publicly available video datasets (SC = static camera, PTZ = pan-tilt-zoom camera).

Dataset	Events	Views	Annotation	Compression	Sequences/ Seconds
VS-PETS FOOTBALL INMOVE	Outdoor people tracking in soccer match	3 SC	Yes for camera 3, XML	720x576, 25 fps, JPEG	5/380
TRICTRAC	Outdoor people tracking in soccer match, synthetic images	7 SC & PTZ	XML	1400x1050, 25 fps, JPEG	13/400

Fig. 8. Image with 3D positions of players projected into the image with the projection formula.

3. CONCLUSION AND FUTURE WORK

We have produced new sport video datasets and propose them as test data for performance evaluation and comparison of VCA systems. The novelty is the use of PTZ cameras with GT that contains both object positions and cameras parameters. We present the format of the images, the scene and scenario and the GT xml format. Table 3 summarizes the soccer datasets publicly available on the Web.

Our test sequences are synthetic and thus come with objective and accurate ground truth data. However, by the very nature of synthetic data, the realism of the enacted scenarios and the visual appearance of the renderings are somewhat limited. Thanks to the use of sophisticated rendering techniques, we believe nevertheless that the dynamic renderings are realistic enough to serve as a valid reference for comparison of VCA algorithms.

In future work, we will propose new sequences with new scenarios, including rendering of shadows or not, background or not, etc. We will also introduce representative evaluation metrics for these new PTZ data.

Acknowledgment: This work is supported by the Walloon Region within the scope of the TRICTRAC project. We thank Thomas Gallienne for his helpful work.

References:

[1] J.B. Hayet, T. Mathes, J. Czyz, J. Piater, J. Verly, and B. Macq. A Modular MultiCamera Framework for Team Sports Tracking. in Proc. of the IEEE Conf. on Advanced Video and Signal based Surveillance (AVSS'05). 2005

[2] B. Fisher, The PETS04 Surveillance Ground-Truth Data Sets, 6th International Workshop on Performance Evaluation for Tracking and Surveillance, 10 May 2004, Prague, Czech Republic.

[3] C. Jaynes, A. Kale, N. Sanders, E. Grossmann, The Terrascope Dataset: A Scripted Multi-Camera Indoor Video Surveillance Dataset with Ground-truth, The Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, In conjunction with ICCV 2005, 15-16th October 2005, Beijing, China.

[4] J. Black, T. Ellis and P. Rosin, A Novel Method for Video Tracking Performance Evaluation, Proceedings of the IEEE International Workshop on VS-PETS 2003, Nice.