



## APIDIS

Autonomous Production of Images based on Distributed and Intelligent Sensing

STREP Project, 1st FP7-216023

### **D3.1 Deployment of the acquisition and storage system, including calibration and annotation issues**

Due date of deliverable: 31-12-2008  
Actual submission date: 09-01-2009

Start date of project: 1<sup>st</sup> January, 2008

Duration: 36 months

Lead contractor for this deliverable: ACIC

[Revision Final v1]

<b>D3.1</b>	<b>Deployment of the acquisition and storage system, including calibration and annotation issues</b>
Project Acronym :	APIDIS
Contract No :	FP7-216023
Due Date :	31-12-2008
Reply To:	Christophe Parisot <a href="mailto:parisot@acic-tech.be">parisot@acic-tech.be</a>
Actual date of delivery	09-01-2009

## 1. Executive Summary

This document describes the deployment of the APIDIS system in a real environment. It defines the specifications and explains how to install the acquisition and storage systems. The document also defines the content annotation syntax, and addresses calibration and sensor positioning issues. Finally, the first results of the generation of planar images from omnivision are included in this deliverable.

**Deliverable Identification Sheet**

<b>Project ref. no.</b>	FP7-216023
<b>Project acronym</b>	APIDIS
<b>Project full title</b>	FP7-216023
<b>Security (distribution level)</b>	Public (PU)
<b>Contractual date of delivery</b>	Month 12, 12, 31, 2008
<b>Actual date of delivery</b>	Month 12
<b>Deliverable number</b>	D3.1
<b>Deliverable name</b>	Deployment of the acquisition and storage system, including calibration and annotation issues
<b>Type</b>	Report
<b>Status &amp; version</b>	
<b>Number of pages</b>	
<b>WP / Task responsible</b>	WP3 / ACIC
<b>Other contributors</b>	QMUL, EPFL, UCL, MP
<b>Main Author(s)</b>	Christophe Parisot, Fahad Daniyal, Yannick Boursier, Christophe De Vleeschouwer, Eric Martrou
<b>EC Project Officer</b>	Albert Gauthier
<b>Abstract</b>	This document describes the deployment of the APIDIS system in a real environment. It defines the specifications and explains how to install the acquisition and storage systems. The document also defines the content annotation syntax, and addresses calibration and sensor positioning issues. Finally, the first results of the generation of planar images from omnivision are included in this deliverable.
<b>Keywords</b>	Acquisition, storage, calibration, annotation, omnivision, specs
<b>Sent to peer reviewer</b>	15-12-2008 to EPFL
<b>Peer review completed</b>	18-12-2008
<b>Circulated to partners</b>	15-12-2008
<b>Read by partners</b>	yes
<b>Mgt. Board approval</b>	Pending

## Table of contents

<b>1. EXECUTIVE SUMMARY</b>	<b>2</b>
<b>2. INTRODUCTION</b>	<b>6</b>
<b>3. DISTRIBUTED AUDIO-VISUAL ACQUISITION SYSTEM SPECIFICATIONS</b>	<b>6</b>
<b>ACQUISITION DEVICES</b>	<b>6</b>
Cameras	6
Audio equipment	9
Omnivision sensors	10
Data acquisition	11
Storage equipment	12
<b>DISTRIBUTED SYSTEM CONFIGURATION</b>	<b>12</b>
Sensor positioning for basket	12
Sensor positioning for surveillance	14
<b>4. CONTENT STORAGE AND ANNOTATION SYNTAX</b>	<b>14</b>
FLEXIBLE RANDOM ACCESS CONTENT STORAGE	14
ANNOTATION SYNTAX AND SEMANTIC FOR SALIENT SEGMENTS IDENTIFICATION	17
<b>5. MULTI-SENSOR CALIBRATION</b>	<b>17</b>
<b>6. ACQUISITION SYSTEM DEPLOYMENT AND ASSESSMENT</b>	<b>18</b>
<b>PRACTICAL DEPLOYMENT OF THE ACQUISITION SYSTEM</b>	<b>18</b>
1 <sup>st</sup> deployment for basketball events in April 2008	18
2 <sup>nd</sup> deployment for video surveillance in September 2008	20
Deployment of the first audio acquisition system	23
<b>ASSESSMENT OF THE FIRST ACQUISITION CAMPAIGNS</b>	<b>25</b>
Spatio-temporal resolution and video quality	25
Ease of deployment	27
Scalability	27
Calibration performance	28
Robustness	29
Cost effectiveness and synchronization issues	29
Audio	30
<b>7. GENERATION OF PLANAR IMAGES FROM OMNIVISION</b>	<b>33</b>
<b>8. CONCLUSIONS</b>	<b>35</b>
<b>9. REFERENCES</b>	<b>37</b>

Table of figures

Figure 1: Arecont AV2100M camera..... 8

Figure 2: Fujinon Fish-Eye FE185C086HA-1 lens..... 9

Figure 3: Polar pattern of the MCE 530 microphones ..... 10

Figure 4: A STAC sensor consisting of a single camera and a pair of microphones in stereo configuration..... 10

Figure 5: Visiobox used for (analogue) data acquisition and synchronization ..... 10

Figure 6: HP Proliant DL380 G5 server..... 11

Figure 7: Portwell DVC-5215 8-Channel Digital Video Capture Board ..... 12

Figure 8: ‘T shaped’ array of omni-directional microphones (filled circles) with a wide lens conventional camera in the array. Proposed values are 200 mm for  $d_1$  and 300 mm for  $d_2$ ..... 13

Figure 9: Basketball court ..... 19

Figure 10: Cameras positions for basketball events..... 19

Figure 11: Sample pictures from each of the distributed video sensors in the basketball environment..... 20

Figure 12: Video surveillance site..... 21

Figure 13: Cameras positions at the video surveillance site ..... 22

Figure 14: Sample pictures from each of the video surveillance cameras ..... 23

Figure 15: Camera position in the proof of concept audio video acquisition ..... 24

Figure 16: Sample images for the acquisition set up of Figure 15..... 24

Figure 17 Sample audio for the acquisition set up of Figure 15. Green and blue represent the audio signal received by the two members of a STAC pair ..... 25

Figure 18: Full resolution extract from camera1 ..... 26

Figure 19: Full resolution extract from camera1 with sharpen from Gimp (sharpen parameter is 90)..... 26

Figure 20: Fast Fourier Transform (FFT) of the full resolution extract from camera1 ..... 27

Figure 21 A panoramic view generated by two overlapping camera views (Camera 1 and 6)..... 28

Figure 22: Track of a single person when viewed from camera 1 and camera 4 of the basketball sequences projected on the (panoramic) overhead view..... 29

Figure 23: Proposed Framework for AV synchronized acquisition..... 32

Figure 24: Proposed Framework for AV synchronized acquisition..... 32

Figure 25: Time Difference of Arrival Measurement between two audio sensors. .... 32

Figure 26: Original frame from camera3 ..... 34

Figure 27: Generated planar image ..... 35

## 2. Introduction

---

This document describes the deployment of the APIDIS system in a real environment. The system and its first in situ deployments are detailed within the following topics:

- Distributed audio-visual acquisition system specifications.
- Content storage and annotation syntax
- Multi-sensor calibration
- Acquisition system deployment and assessment.
- Generation of planar images from omnivision

## 3. Distributed audio-visual acquisition system specifications

---

Involved partners: ACIC, QMUL, EPFL, MP

The selection of the audio-visual acquisition system components is driven by the following three main requirements:

- The system should be cost effective with regards to the budget our end users would be ready to spend for an APIDIS system.
- The acquisition system must fit the technical requirements for efficient autonomous scene understanding algorithms.
- The acquisition system must fit the technical requirements for the production of summaries with good audio-visual quality.

### ***Acquisition devices***

#### **Cameras**

The selection of the cameras depends on the scenario which is considered.

For video surveillance, it is assumed that an APIDIS system is only a sub-part of the video surveillance infrastructure. In this particular case, the cameras have already been selected and positioned in order to cover the most critical areas. The number and technical specifications of cameras are driven by the lightening conditions and the fields of views to cover. The cameras may support colour or not. The resolution may vary from one camera to the other.

For sport events on the contrary, the APIDIS system is totally independent from other needs. In this case, the technical specifications and placement of cameras depend only on the objectives of the APIDIS system.

While generally applicable for the video surveillance scenario, the main criterions for the cameras in the basketball scenario are:

- **Resolution**

Since cropped and zoomed images will be extracted from the different static cameras to simulate PTZ cameras, the resolution should be high enough to provide good rendering when zooming on a particular action.

Furthermore, the higher the resolution the easier it is to automatically detect and recognize objects. When the resolution is high, fewer cameras are needed to cover the basketball court for video analysis purposes.

- **Frame rate**

For evident rendering issues, the frame rate should be higher or equal to 25 frames per second. Furthermore, the higher the frame rate the easier the tracking of the players by video processing algorithms.

- **Color**

The color is required for rendering nice video summaries of the events. The color might also help discriminate between the players of the two teams or for detecting the ball.

- **Sensitivity**

The higher the sensitivity, the better it is. Indeed, high resolution cameras generally suffer from lower sensibility than standard PAL CCTV cameras. For choosing between cameras with the same resolution, the higher the captor size the better it will probably be.

- **Synchronization**

It is important that all camera streams are well synchronized. Synchronization is necessary to be sure to render summaries that simulate real-time changes of cameras (no jumps in time when switching from one stream to another). Furthermore, synchronization is required to extract a global overview of the actions in the same common spatio-temporal referential for all camera streams. A perfect synchronization is ideal for a joint analysis of multiple views.

- **Cost**

The cost for the video equipment should be reasonable to fit a future commercialization of the system. However, since hardware costs will decrease in the next two years and some economical models of the APIDIS system may ask for people to pay for their personalized summaries, the current price of high resolution cameras doesn't necessarily reflect what will be a reasonable cost at the end of the project. The cost was however a limitation in the deployment of our prototype, since it had to fit the APIDIS equipment budget.

The cameras that have been selected for the April'08 acquisitions are Arecont Vision AV2100M IP cameras. Their features are the following:

- Up to 24 fps in 1600x1200 color acquisition mode.
- Power over Ethernet (PoE) network camera.
- ½" captor size with sensitivity of 0.1 lux at aperture 1.4.
- Motion JPEG stream

At the time of choosing the video sensors, these cameras were the only ones providing 2 MPixels with such a high frame rate, given the budget and time

constraints. The recording software provided by Arecont stamps each frame with its acquisition time with an accuracy of one thousandth second.



Figure 1: Arecont AV2100M camera

### Synchronization issue

With these cameras, the video streams are acquired and time stamped independently for each camera.

Other kinds of cameras support the following synchronization mechanisms:

- NTP and tags: here the cameras support internal clock synchronization with Network Time Protocol servers and each frame is tagged with the time the picture has been captured by the camera using the Universal Time Coordinate (UTC) system. With this kind of cameras, the time stamps are coherent whatever the acquisition servers' clocks in case of a distributed storage system.
- Trigger synchronization: here the capture of frames is controlled by an external signal, i.e. one pulse generator for all the cameras. In this case, the frames are captured at the same time for all sensors.

At the time of April'08 hardware selection, we didn't find any 2 MPixels NTP capable camera on the market that could be used for our acquisition campaign (e.g. the latest Axis 211 Network Camera supports NTP but provides only 640x480 pictures).

2 MPixels GigaE cameras with support for an external trigger were available (Pulnix TMC-2030GE) but their cost was almost five times greater than the Arecont solution. 2 MPixels IEEE1394b cameras were also available (Grasshopper) for almost three times the price of the Arecont cameras.

Finally, the selected cameras provide high resolution pictures at 24 fps in color. For synchronization, it is assumed that the recording software is run on one PC or on several NTP synchronized PCs.

### Lenses

While most cameras are equipped with standard C-mount optics, some are equipped with Fish-Eye lenses for very wide fields of views. The selected Fish-Eye lens model is Fujinon FE185C086HA-1. For our cameras, they provide a horizontal angle of view of 136° and a vertical one of 102°.



Figure 2: Fujinon Fish-Eye FE185C086HA-1 lens

### Audio equipment

A first audio acquisition system has been tested in November'08. The equipment that was deployed for those acquisitions consisted in the following configuration:

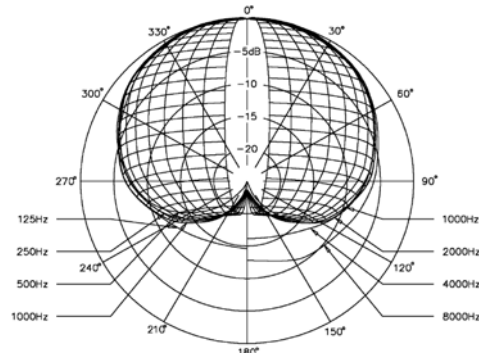
- 3 analogue cameras
- 6 cardioid microphones
- Visioboxes for data acquisition and synchronization

This acquisition campaign was run as a proof of concept for audio analysis.

The cameras used were KOBİ KF-31CD analog CCD surveillance cameras with video acquired at a resolution 360x288 (25 fps).

### Microphones

The microphones are MCE 530 cardioid condenser microphones, with a polar pattern as shown in Figure 3. The AV acquisition hardware is used in a stereo audio and cycloptic vision (STAC) configuration (see Figure 4). This configuration allows us to use the time difference of arrival between two pairs of microphones to estimate the direction of arrival of the sound.



**Figure 3: Polar pattern of the MCE 530 microphones**



**Figure 4: A STAC sensor consisting of a single camera and a pair of microphones in stereo configuration**

### Acquisition and Synchronization

For acquisition and synchronization of video and audio streams we used VisioWave VisioBox (Figure 5). The VisioBox is a professional low port density IP coder, providing PAL or NTSC video for digital video surveillance and CCTV.



**Figure 5: Visiobox used for (analogue) data acquisition and synchronization**

### Omnivision sensors

Two series of omnidirectional images have been acquired.

First, natural lab images have been captured by a paracatadioptric camera with a mirror from the Remote Reality Corporation (<http://www.remotereality.com/>), that has 360 degrees view in the azimuth angle and for which the elevation view ranges from 35 to 92.5 degrees.

The selected camera is an AXIS 2420 PAL Network Camera.

A second set of images has been acquired during the first acquisition campaign using the fish eye lenses described above.

Both cameras and lenses presented here provide images that are perfect for our tasks. Indeed, the generation of planar images partially relies on the mapping between the original geometry and the spherical geometry, and there is a strict equivalence between fisheye images and catadioptric images on the one hand, and spherical images on the other hand.

## Data acquisition

The cameras are connected to a single server through a Hewlett Packard PoE switch with 100 Mbits ports and two Gbits ports. The cameras are connected to 100 Mbits ports while the server uses one of the Gbits port.

Since the cameras support PoE, there is only need to deploy network cables. This simplifies the installation.

The server is a Hewlett Packard quad core server at 2.5 GHz with 2 GB of RAM (see Figure 6).

The synchronization of the video streams is ensured by the fact that all streams are recorded by the same server.

For video surveillance, the server processing capacity and number of disks depend on the type of cameras. Indeed, standard CCTV cameras generate much less bandwidth than high resolution IP cameras. However, while Arecont frames need just being captured and stored in their native (Motion) JPEG format, the acquisition of analog cameras requires computing power for live encoding of the video stream. The same server is sufficient in a video surveillance context for the acquisition, MPEG4 compression and storage of height standard analog cameras. For the acquisition of analog cameras, PCI-Express Digital Video Capture Cards have been selected (see Figure 7).



Figure 6: HP Proliant DL380 G5 server

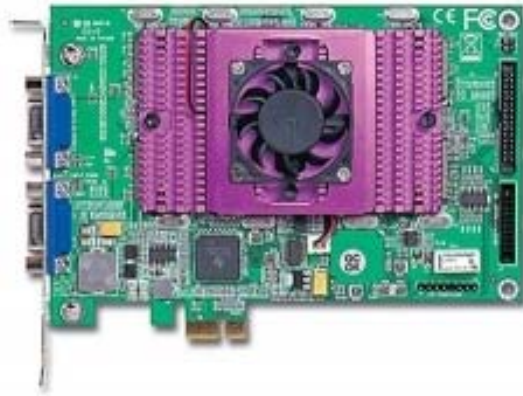


Figure 7: Portwell DVC-5215 8-Channel Digital Video Capture Board

## Storage equipment

The acquisition of seven AV2100M cameras requires distributing the recording bandwidth on several hard drives (one hard drive for three cameras). Therefore, the server is equipped with three standard hard drives. In the video surveillance context, two hard drives are sufficient for the acquisition of eight analogue channels recorded in MPEG4.

### *Distributed system configuration*

This section deals with the APIDIS distributed configuration in the case of basketball events and video surveillance scenarios.

## Sensor positioning for basket

### Positioning

The positioning of cameras in a sport event has to follow several rules.

Among them, all cameras must be distributed on the same side of the court. They must not be placed on both sides of the game main axis because switching from one view to another would perturb the spatial representation the viewer has of the game.

However, this rule applies only for cameras of which streams will be used in the video summaries. The other side of the court can be used if this helps video processing algorithms in analysing and understanding the progress of the game.

A good distribution of video sensors is the one that maximises both the information that can be rendered to the viewer and the information that is available to the video analysis behaviour detection algorithms.

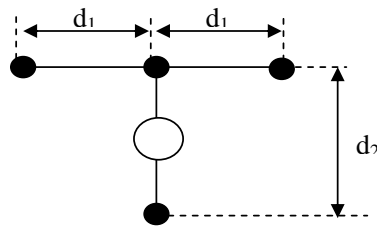
For a basketball game, some cameras should provide an overview of both sides of the court and some cameras should focus on the three-point field goal, the basket, the score panel, etc. Cameras fixed on the roof of the hall can provide

views with minimal overlap between the players and therefore help tracking the different players along the game.

Interested readers can refer to deliverables D2.1 [1] and D2.2 [2] for more details on sensors positioning.

Section 6 details a possible deployment of cameras for basketball events.

The audio setup consists of the 'T shaped' array of microphones with a camera in the centre as shown in Figure 8. QMUL has experience working with STAC configuration of the microphones whereas T-shaped array has been used in the CHIL project [3]. Furthermore, a T-shaped array can be considered as a pair of STAC sensor hence existing algorithms can be used with better accuracy due to the availability of more audio sensors.



**Figure 8: 'T shaped' array of omni-directional microphones (filled circles) with a wide lens conventional camera in the array. Proposed values are 200 mm for  $d_1$  and 300 mm for  $d_2$**

### Calibration issues

The calibration consists in finding the mathematical transformation that relates objects in the 3D world to pixel positions in the video stream. It can be split into getting the intrinsic and extrinsic parameters. Intrinsic parameters depend on the camera, optics, captor, etc. while extrinsic parameters depend on the position and field of view.

In order to facilitate the calibration, it is necessary to get measures and/or 3D positions of objects in the global terrestrial referential and pick them in the video stream pictures to get the projective transformation.

Section 5 shows what kind of inputs can be used for calibrating the system.

### Privacy issues

In a game it is normal to have some shots focusing on the audience. Therefore, the public needs to be warned that video sensors may acquire pictures of them. The public should be informed on the use that is considered for the video streams and it should be informed on its rights to access and rectify any information related to it.

## Sensor positioning for surveillance

### Positioning

In the video surveillance scenario, the APIDIS system has no control on the position and distribution of sensors. The system must be flexible enough to work in a large number of configurations. As an example, some camera may overlap when other cameras may be totally independent each other.

### Calibration issues

The same calibration methods as the ones for the basketball context can be used.

### Privacy issues

When using the APIDIS system for browsing recorded video surveillance streams, the system must follow the rules that apply in each country in terms of storage maximum duration, restricted access, etc. In other words, the limitations that apply to video surveillance data will also apply to APIDIS.

However, the APIDIS system will not change the data that will be observed but only the way they will be browsed. Thus, it is sufficient to restrict the access to the APIDIS tools to people that already have access to video surveillance data to follow privacy rules.

## 4. Content storage and annotation syntax

---

Involved partners: ACIC, UCL, EPFL

### ***Flexible random access content storage***

In order to get flexible and generic access to all files, the following directory structure and files naming conventions are used.

Any information that is stored is considered as a media. Videos, sounds, low level features, high level metadata and all other information follow the same storage and access rules.

A media can be uniquely determined by a name (e.g. camera1) and a type (e.g. objects.xml).

Since some media cannot be saved within single files (e.g. several hours of videos), the filenames include a time stamp following the ISO 8601 date/time syntax. In general, the time stamp of each file refers to the time stamp of the first data it includes (e.g. time stamp of the first frame in a given video file). Time stamps in filenames allow for fast retrieval of media data at a given period.

Then, filenames look like:

MediaName\_ISOTimeStamp.MediaType

Since some media may be split into a large amount of files, files are organized in a directory tree structure that reduces the amount of files per sub-directory and also optimizes the search of a media at a given period of time. Indeed, the tree structure which is implemented is:

MainMediaDirectory / YYYY / MM / DD / hh / mm /... file

where “/” splits a directory into sub-directories, YYYY is the year, MM the month, DD the day in the month, hh the hour (in 24 hours format), mm the minutes, etc. When looking for a particular date, it is very fast to locate the file that is associated with this date.

A library has been developed and distributed to partners for:

- Looking for files containing the data of a given media (name and type) at a given time stamp in a given main directory.
- Providing the directory tree and filename for writing a given media (name and type) given its time stamp and maximum duration.

This way, any data that needs to be exchanged between different APIDIS modules is abstracted in terms of its access. None of the modules needs to know the exact location of data except the common main directory.

### **Random access to video frames**

The format in which video sequences are stored depends highly on the possible capture modes. When the bandwidth of acquired data is very high, e.g. high resolution and/or raw data, encoding/transcoding them in live would require extreme computing resources. Therefore, although transcoding may be performed in a second step, original video data format is first considered.

Depending on the video devices and scenarios, APIDIS will probably have to support the following video formats:

- Raw data (e.g. Bayer color matrix)
- Motion JPEG. This is the most common format for IP cameras
- MPEG 2, MPEG 4, H264 codec families. MPEG 4 will be used when acquiring analog video surveillance streams.

While the first two formats provide random access to any frame in the video stream, the last ones may require decoding several frames to obtain a predicted frame.

In any case, an index file is attached to each video sequence. The index file has the same name as the video file with the “.idx” suffix. This index contains for each frame:

- The offset of the frame, in bytes, in the video file
- The size of the frame in bytes
- The time stamp of the frame
- The type of frame (e.g. I, P, B for MPEG encoded streams)

This index allows for fast retrieval of one frame at a particular date in each video file. When the desired frame is predicted from previous frames, the index file provides an easy access to the first frame (I-frame) from which to decode the stream.

A library has been developed and distributed to partners for browsing recorded video streams. It provides:

- Forward/backward decoding capabilities
- Access to the video at a particular time stamp
- Abstraction of the storage layer (hides the fact that video sequences may be split into several files).

In conclusion, the following are used to get flexible random access to APIDIS data:

- Specific directory tree structure
- Files naming convention
- Library for abstracting read/write access to data
- Index files attached to video sequences
- Library for abstracting read access of recorded video streams

Since the project focuses more on the creation and production of personalized content rather than on its practical efficient manipulation and distribution, we have decided not to convert the JPEG captured content to a more flexible format such as JPEG2000. Conceptually, such kind of format would allow a more efficient access to sub-sampled and cropped versions of the images. However, implementing such a conversion and associated content-access libraries would have a cost in terms of human resources for the project, without bringing any real added value regarding the autonomous production concepts that are central to the project.

## ***Annotation syntax and semantic for salient segments identification***

The syntax and semantic of the annotations that are considered to identify salient segments within captured video content have been presented in details in Section 5.3 of Deliverable 2.1. They have also been published during the NEM summit 2008 in Saint Malo, France (see Annex of D8.1). In short, they have been defined in a way that is expected to support user-driven and personalized summarization of the typical sport events.

A set of tools has then been designed and implemented to manually annotate (part of) the captured content, thereby providing the ground truth and the examples needed to launch the content analysis and production activities of APIDIS. The toolbox consists in a user-friendly interface able to handle multi-view content, so as to manually define events and position of objects of interests over time.

A user manual for the APIDIS tools series is provided in Annex 1 of deliverable 3.2. A running executable is available upon request to fan.chen@uclouvain.be.

### **5. Multi-sensor calibration**

---

Involved partners: QMUL, EPFL

Sensors calibration is highly desirable for consistent and uniform display as well as to survey the monitored scene. Basic calibration techniques require only 2-D point matches in multiple views and work for known cameras parameters or 3-D knowledge of the scene. An iterative statistical model to recover extrinsic camera parameters can be used to perform camera calibration. Given, at least, a single object's information in a network of top-view cameras, this statistical model estimates missing trajectories across the unobservable regions and uses both parametric and non-parametric algorithms.

In order to calibrate (conventional cameras), we use a set of control points whose object space/plane coordinates are already known. The control points are normally fixed to a rigid frame, namely a checkerboard pattern. In addition to this, some parameters that are used as added information include intrinsic parameters such as the focal length, the scale factor associated to the sensor radial lens distortion coefficients and the centre of this radial distortion. The extrinsic parameters include the orientation and the location of the camera in the 3D world coordinates.

When the position of objects must be inferred from one camera view to another one, there should be enough overlap between cameras to establish a common coordinate plane. The homographic projection then enables transfer of the coordinates of one camera to another and also enables us to project the coordinates of the camera on a panoramic view. Joint processing of video

streams also necessitates synchronization across cameras. See Section 6 for further details.

## 6. Acquisition system deployment and assessment

Involved partners: MP, ACIC, QMUL, EPFL

### ***Practical deployment of the acquisition system***

This section presents the first deployments of the APIDIS acquisition system in both sport and video surveillance environments.

#### **1<sup>st</sup> deployment for basketball events in April 2008**

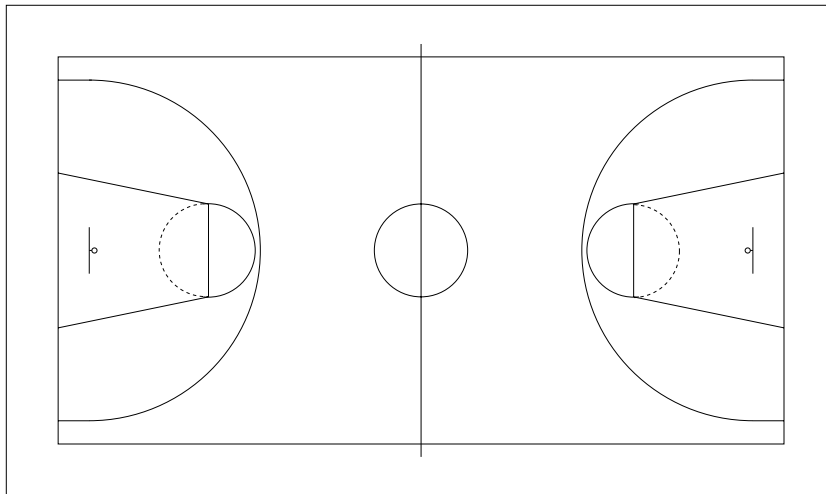
A first basketball dataset has been acquired beginning of April 2008 in Namur, Belgium.

The acquisition system was composed of:

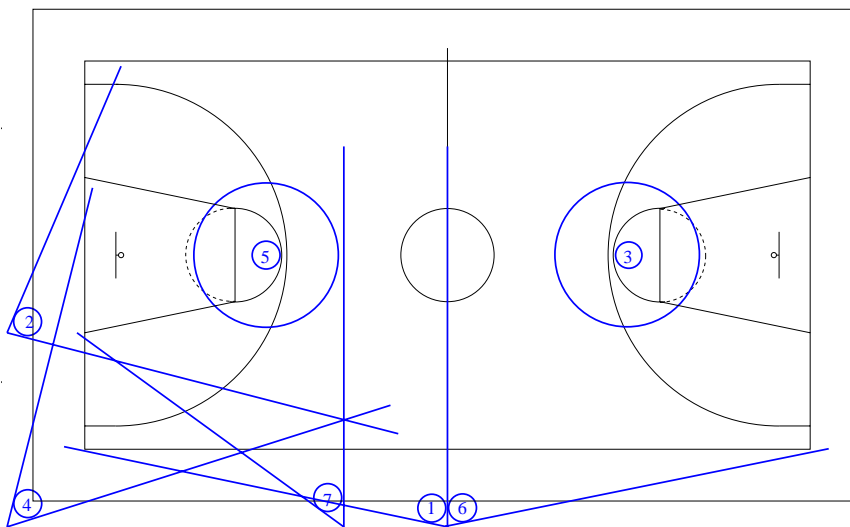
- 7 Arecont AV2100M cameras (1600x1200 at ~ 21.25 fps)
- Fujinon FE185C086HA-1 lenses for wide angles of view
- 1 Hewlett Packard quad core server at 2.5 GHz with 2 GB of RAM.
- 3 hard disk drives (7200 RPM)
- Hewlett Packard PoE switch with 100 Mbits ports and two Gbits ports.
- Arecont IP recording software

Figure 9 presents an official basketball court. Figure 10 shows how the cameras have been distributed all around the basketball court. All the cameras are placed on the same side of the court as required for production (see section 3). It is assumed that for a true installation, the number of video sensors will be increased to get a perfect symmetry in the acquisition of both parts of the court. The current setup, if not complete, is sufficient for demonstrating the key concepts of an APIDIS system.

Since all video streams are acquired by the same server, the time stamps for each frame of each stream are coherent (same time reference).



**Figure 9: Basketball court**



**Figure 10: Cameras positions for basketball events**

Figure 11 shows sample pictures of each camera in the basketball context.

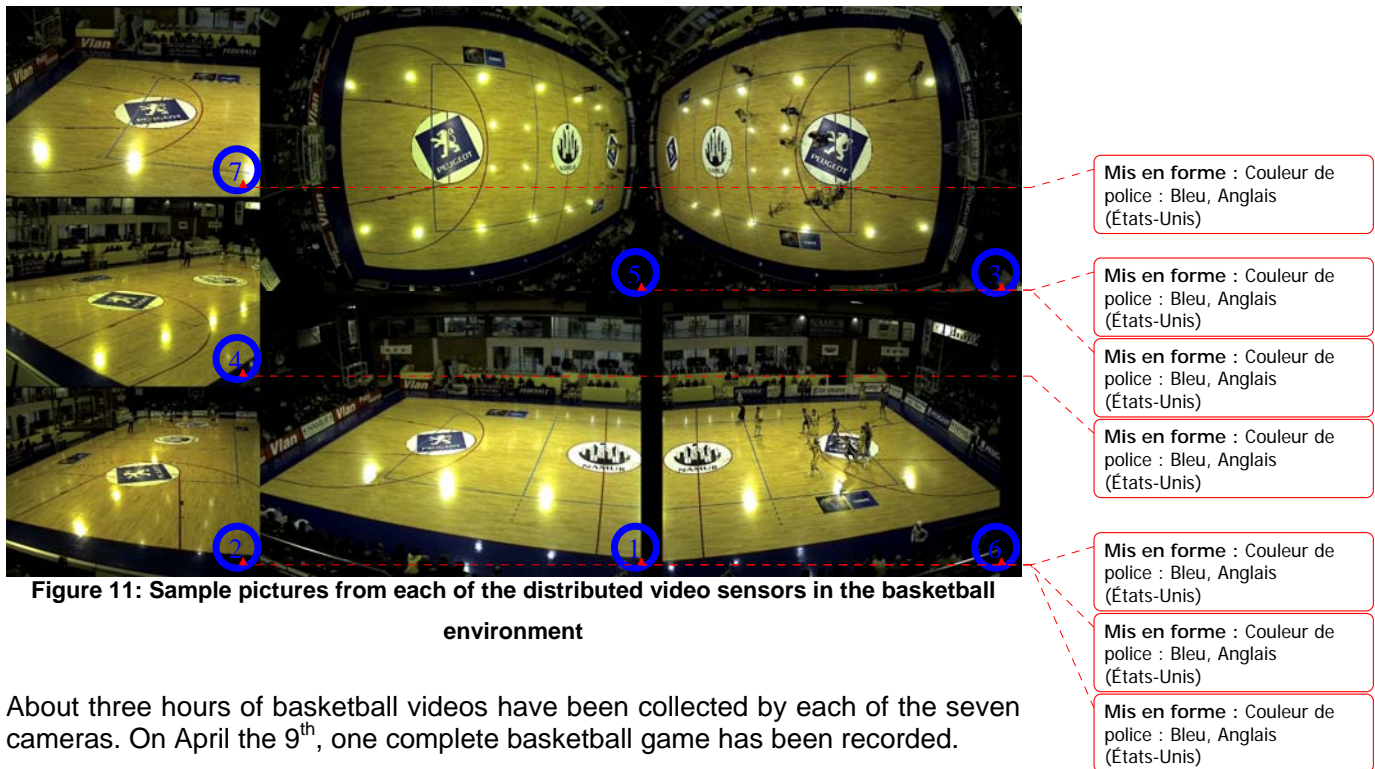


Figure 11: Sample pictures from each of the distributed video sensors in the basketball environment

About three hours of basketball videos have been collected by each of the seven cameras. On April the 9<sup>th</sup>, one complete basketball game has been recorded.

The dataset is available from the APIDIS website.

**2<sup>nd</sup> deployment for video surveillance in September 2008**

ACIC has obtained an access to the streams of the video surveillance cameras installed all around the building hosting its premises. Since September 2008, those streams can be captured whenever desired for APIDIS.

The acquisition system is composed of:

- 8 existing standard CCTV cameras (7 outdoor + 1 indoor)
- 1 Hewlett Packard quad core server at 2.5 GHz with 2 GB of RAM (same as for the first basketball acquisition)
- 2 DVC-5215 8-Channel PCI-Express Digital Video Capture Cards
- 2 hard disk drives (7200 RPM)
- APIDIS acquisition software based on Video For Linux API running under Linux.

Figure 12 presents an aerial view of the building. Figure 13 shows how the cameras are distributed all around the building. Some of the cameras are able to capture colour while others aren't. This test site is representative of already installed analogue video surveillance systems.

Since all video streams are acquired by the same server, the time stamps for each frame of each stream are coherent (same time reference).

Figure 14 shows sample pictures of each camera in the video surveillance context.



**Figure 12: Video surveillance site**



- Mis en forme : Couleur de police : Bleu, Anglais (États-Unis)
- Mis en forme : Couleur de police : Bleu, Anglais (États-Unis)
- Mis en forme : Couleur de police : Bleu, Anglais (États-Unis)
- Mis en forme : Couleur de police : Bleu, Anglais (États-Unis)
- Mis en forme : Couleur de police : Bleu, Anglais (États-Unis)
- Mis en forme : Couleur de police : Bleu, Anglais (États-Unis)
- Mis en forme : Couleur de police : Bleu, Anglais (États-Unis)
- Mis en forme : Couleur de police : Bleu, Anglais (États-Unis)

Figure 13: Cameras positions at the video surveillance site



- Mis en forme : Police :Gras, Couleur de police : Bleu, Anglais (États-Unis)
- Mis en forme : Police :Gras, Couleur de police : Bleu, Anglais (États-Unis)

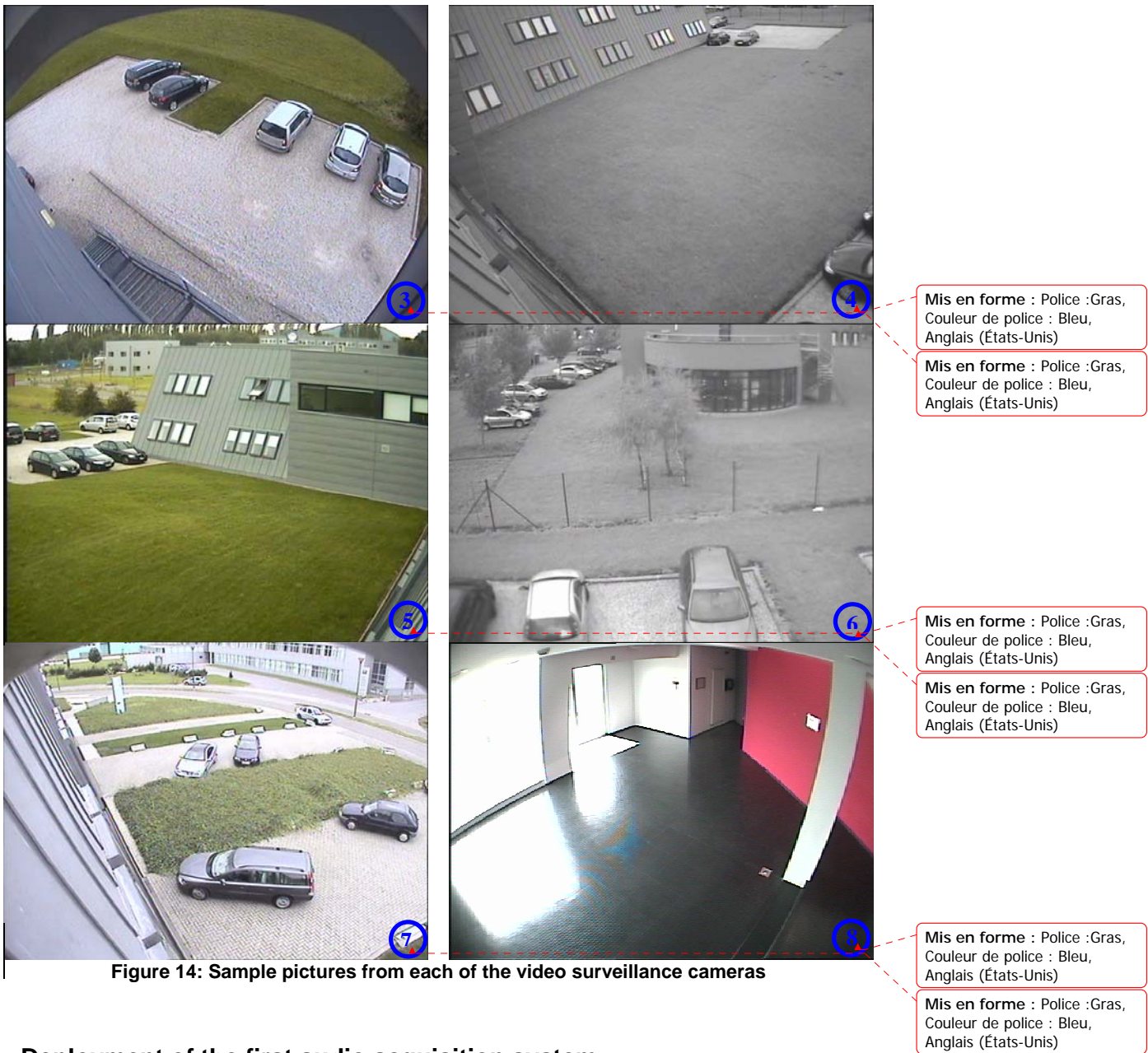


Figure 14: Sample pictures from each of the video surveillance cameras

**Deployment of the first audio acquisition system**

As a proof of concept an audio visual acquisition campaign was set up using the configuration shown in Figure 15. Figure 16 shows the sample images obtained from this setup. Figure 17 shows the amplitude plot of the audio signal captured by a single STAC pair.

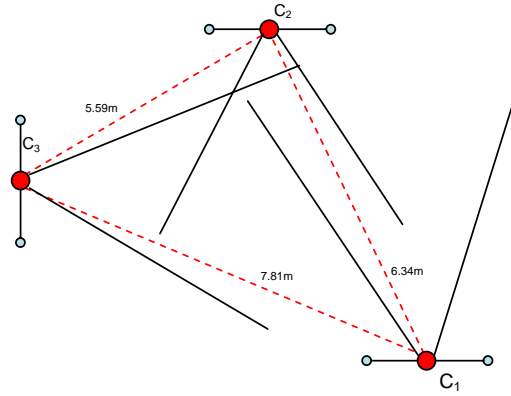
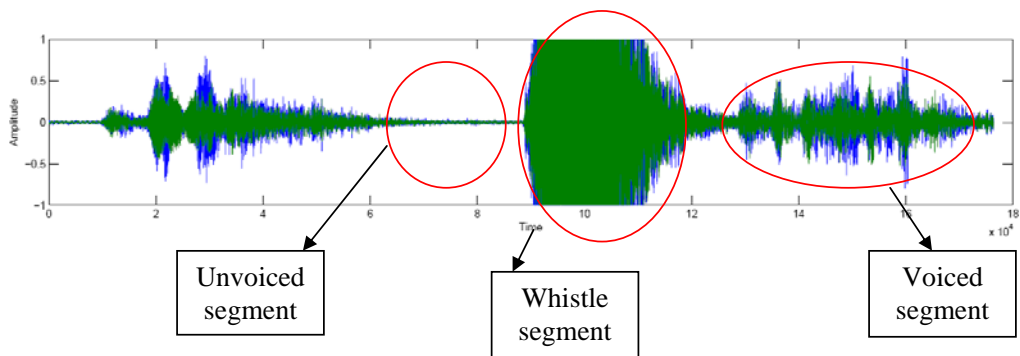


Figure 15: Camera position in the proof of concept audio video acquisition



Figure 16: Sample images for the acquisition set up of Figure 15



**Figure 17 Sample audio for the acquisition set up of Figure 15. Green and blue represent the audio signal received by the two members of a STAC pair**

## ***Assessment of the first acquisition campaigns***

### **Spatio-temporal resolution and video quality**

The Arecont AV2100M cameras theoretically provide 2MPixel coloured frames at 24 fps. In practice, we got an average of 21.25 fps. This frame rate is enough for video processing. Indeed, in our dataset the maximum speed that could be measured (by hand) for a player was close to 4 m/s. This corresponds to a maximum displacement from one frame to the other of 0.19 meter for the players. Hence, the maximum displacement is less than the players' size. The fact that players' positions overlap themselves from one frame to the next one highly facilitates their tracking.

From the manually annotated metadata, the size of the ball is at least 18 pixels in each direction for the top views. In video processing, it is generally considered that objects must at least have a size of 10x10 pixels to be detected. Therefore, the spatial resolution seems good for the Apidis video processing algorithms.

Figure 18 shows an action captured by camera1. Figure 19 shows the same extract after a sharpen operator. From those figures, it can be seen that there is some blur, especially on the moving players. The sensitivity of the camera should be higher to reduce this problem. A better sensitivity would provide both nicer rendering and probably ease the work of some video processing algorithms. Since the AV2100M cameras already use a ½ inch captor size, it seems unlikely that other cost efficient 2MPixel cameras would have provided better performances. Our opinion is that one will benefit from the next generation of video sensors by the end of the project and acceptable video quality will then be available for nice rendering of sport events.

From Figure 19, it can be seen that the pictures quality has not been reduced by the compression since JPEG blocking artefacts are not excessively highlighted by the sharpen filter in the regions with most gradient (e.g.: players contours, lines on the ground, etc.). Figure 20 shows that the high frequencies have been kept by the compression chain. The blur in the pictures seems to be due to the internal cameras acquisition chain (e.g. quality of the sensor), the optics having a maximum aperture of F1.4.

In the video surveillance scenario, the server can record simultaneously eight cameras with 768x576 resolution at 25 frames per second in MPEG 4 with a target bit rate of 3 Mbits per second.



Figure 18: Full resolution extract from camera1



Figure 19: Full resolution extract from camera1 with sharpen from Gimp (sharpen parameter is 90)

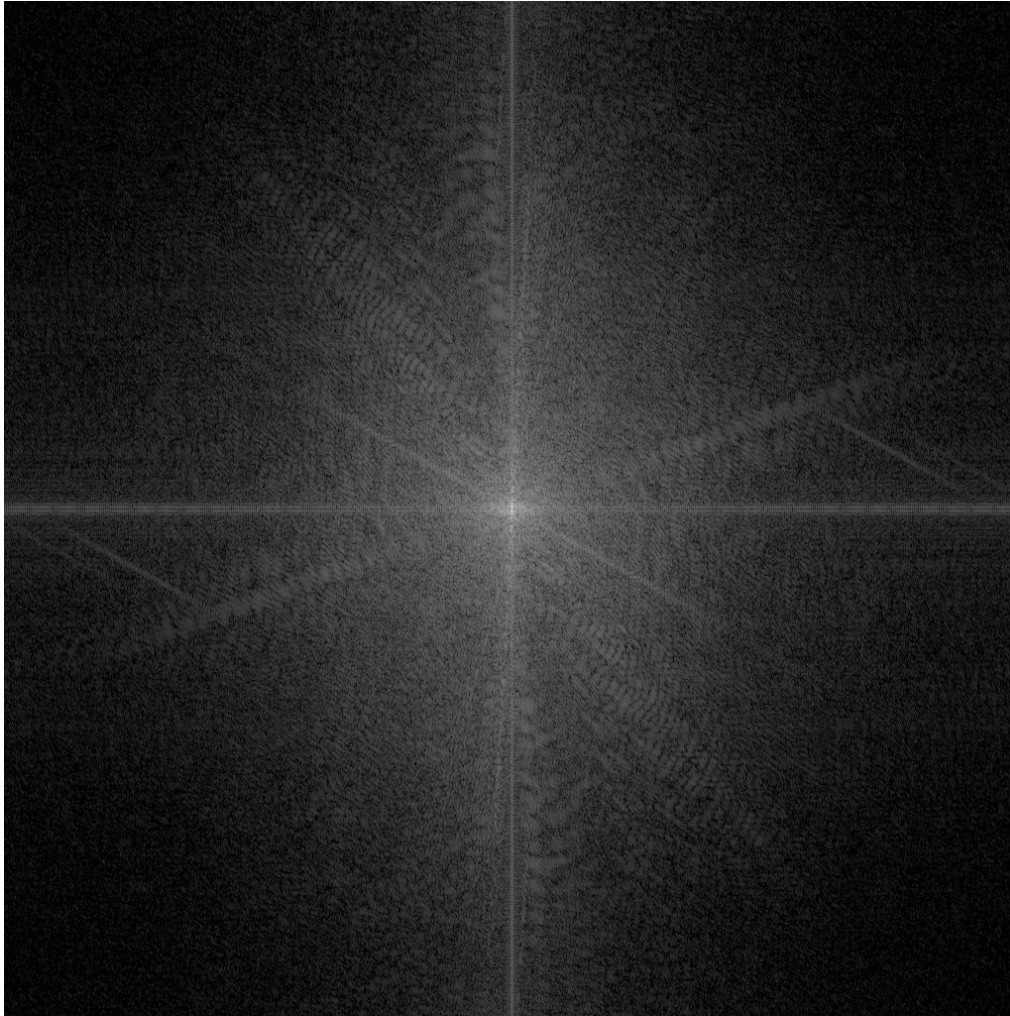


Figure 20: Fast Fourier Transform (FFT) of the full resolution extract from camera1

### **Ease of deployment**

The deployment of the system presents two big advantages:

- There is only one server
- Arecont AV2100M cameras are powered through Power over Ethernet (PoE). This means that only the Ethernet cable is required to plug the cameras into the system. This simplifies the physical installation of the Apidis system and limits sources of defects.

### **Scalability**

The acquisition system that has been chosen for the first acquisition campaigns is scalable in different ways. The server can manage several internal hard disk drives, external storage, and several PCI-Express video acquisition boards. The server also has two Gigabit Ethernet network interfaces.

Hence, the server can be used in a large number of conditions with minor hardware changes (e.g. adding hard disk drives, PCI-Express cards, network cards, etc.).

When the number of video sensors is too high or that it generates too much bandwidth (large number of very high definition video sensors), the acquisition system can either be based on a more capable server or on the use of several networked servers sharing their drives over the network to allow acquisition management and processing of data from the master server.

### Calibration performance

The first step for joint processing is the estimation of a common ground plane between two views of a camera. The overlap between the cameras as shown in Figure 21 (a) and (b) allows us to identify control points across cameras which allow us to establish correspondence among two views. This correspondence can be shown by creating a panoramic view (see Figure 21 (c)).

The second phase of joint processing that is currently employed on the dataset is in terms of fusion of information from Task 4.3 (Visual target tracking). The output for the estimation of the location of a single target in the 3D world coordinates for a pair of Camera 1 and 4 are shown in Figure 22.

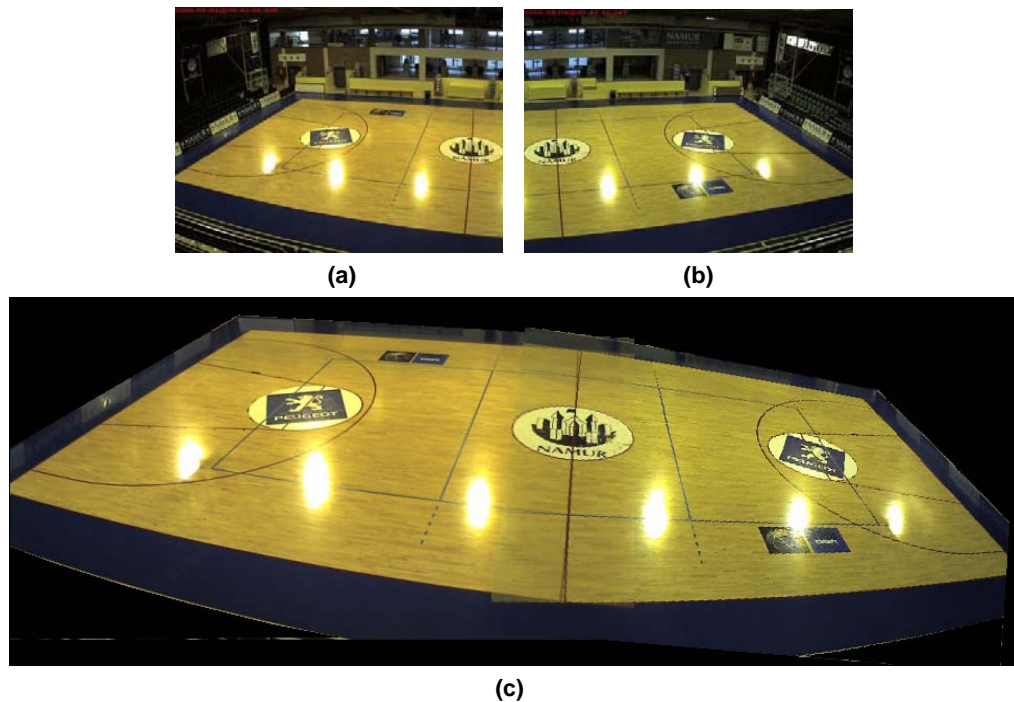
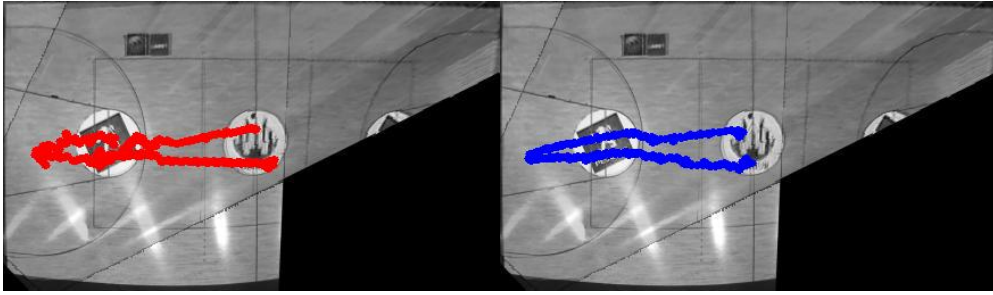


Figure 21 A panoramic view generated by two overlapping camera views (Camera 1 and 6).



**Figure 22: Track of a single person when viewed from camera 1 and camera 4 of the basketball sequences projected on the (panoramic) overhead view**

### **Robustness**

The acquisition system that was used for the first two acquisition campaigns does not depend on the video sensors output frame rate. Therefore, it is able to run with sensors operating at different frame rates and is robust to the fact that one frame might be lost for whatever reason. Since each frame is time stamped by a unique server (i.e. unique time reference), the temporal coherency of data is ensured.

Every part of the acquisition system is professional hardware (server, switch, etc.) and is therefore robust for real deployments in terms of temperature, vibrations, etc.

### **Cost effectiveness and synchronization issues**

Although the economical model of an APIDIS system will be discussed in later deliverables, first figures are discussed here.

Mediapro pays 150 euros per day for a camera man in Spain. But the fee is higher in UK or France where a camera man can easily earn the double per day.

APIDIS aims at providing autonomous production of personalized content. This task cannot be completed by camera men since end-user requests may be very different and therefore difficult to predict at the time of shooting the game. The cost for generating a personalized content from APIDIS recorded video streams can be rounded to 100 euros (half a day of a video editor in Spain).

If the cost effectiveness of the APIDIS system cannot be discussed at this point of the project, the cost of the acquisition system deployed for basketball events can be compared to other solutions.

At the time of writing this document, the system that has been installed in Namur costs around 9,000 euros for seven cameras (11,000 euros for a symmetric installation with ten cameras). The video streams are in colour, close to 25 frames per second, in 1600x1200 frame size. The time stamps of each stream use the same reference server.

Therefore, the video streams are not truly synchronized in the sense that pictures are not captured at the same exact instant by the cameras. However, thanks to the time stamps, it is possible to estimate object trajectories in one video stream and compute the object positions in the frames of the other video streams thanks to multi-camera calibration.

The other possibility is to use a truly synchronized acquisition system in which the frames of all cameras are captured when a signal is sent through an external trigger. Assuming that the same 1600x1200 resolution is targeted, two solutions can be envisaged:

- The first one based on Grasshopper IEEE1394b cameras. The cost of the acquisition system including seven cameras is about 31,000 euros (43,000 euros for ten cameras).
- The second one based on Pulnix Gigabit Ethernet cameras. The cost of the acquisition system including seven cameras reaches about 42,000 euros (60,000 euros for ten cameras).

A truly synchronised acquisition system provides the following advantages/drawbacks:

- All cameras capture a frame at the same moment. This is well suited for joint multi-camera image processing algorithms since no temporal interpolation is needed to predict the position of one object in one view from the positions known in other views.
- Each video stream being captured at a given global frame rate, the time stamps of each frame can be determined accurately from the frame index.
- However, the system requires a very robust acquisition server since the lost of one frame in one video stream would definitively de-synchronized the dataset.

As a conclusion, the solution that has been used for the first basketball acquisitions provides robustness to long term synchronization errors, high resolution, close to 25 frames per seconds at a reasonable cost when compared to a truly synchronized solution (three times more expensive). Furthermore, the cost for a 1600x1200 resolution truly synchronized solution could not be envisaged within the project budget.

The consortium currently plans a second basketball acquisition campaign using truly synchronized cameras that are already in use by some of the partners. Only some streams will have a Megapixel resolution. However, such a second acquisition campaign aims at measuring the advantages and drawbacks of each synchronization solution from the video processing point of view.

## **Audio**

In the first audio acquisition setting, the cardioid microphones limit the audio processing so that the user should be facing the microphone pair for accurate target localization. Another major issue in this setup is the problem of accurate audio-video synchronization.

Therefore, a second audio acquisition setup is considered for future APIDIS deployments. This setup is described here.

Six arrays of microphones (4 microphones per array) should be used for the detection and localisation of the whistle of the referee, as well as the analysis of ambient sounds that will help scene analysis. The captured audio should be in mono .wav format at 16-bit, 44.1 KHz. The necessary setup (Figure 23) is as follows.

- 24 Microphones
- 3 Preamplifiers
- AD converters
- Acquisition PCs

#### **Configuration**

1. 8 microphones are connected to each pre-amplifier
2. 3 pre-amplifiers are connected to the AD converter (sound-card)
3. each sound card is connected to a PC through the firewire port

This framework is shown in Figure 24.

#### **Audio-Visual Synchronization**

The triggering pulse from the NI PCI6601 trigger board acts on the AD converter and AV synchronization module. For audio the maximum synchronization error envisioned is no more than 100 microseconds. Audio source localization is done using Time Delay Of Arrival (TDOA). Audio from speaker, to microphones M1 and M2 (Figure 25) reaches at different time instances. The estimation of the direction of arrival is given by the TDOA between the two microphones. Thus, having a delay due to synchronization can result in inaccurate detections. A delay of one millisecond leads to an estimate which can be off target by 20 degrees at 44.1 kHz. With the proposed synchronization of the microphones we aim to reduce this error to less than 2°. This synchronization can be achieved using a common triggering signal for all the sensors. Table 1 provides an estimate of the price of each item and the total cost associated with such setup.

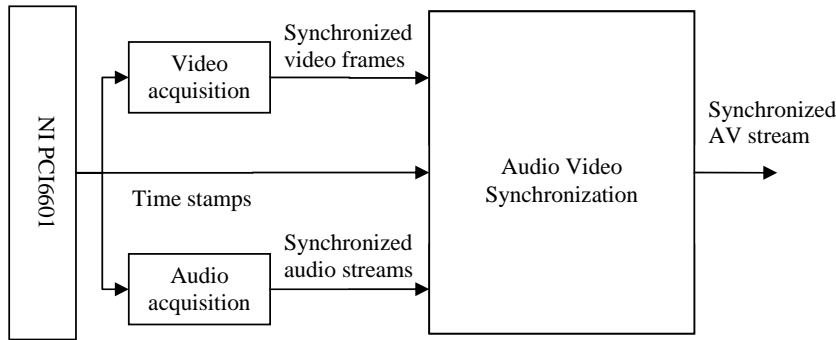


Figure 23: Proposed Framework for AV synchronized acquisition

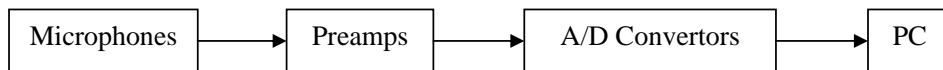


Figure 24: Proposed Framework for AV synchronized acquisition

Equipment	Make/Model	Qty	Unit Price (£)	Total
Microphones	Audio-Technica AT899 Omni Lapel Microphone [4]	24	120	2880
PC interface	Mark of the Unicorn 2408mk3 PC interface [5]	3	525	1605
Pre-amplifier	Behringer ADA8000 preamplifier [6]	3	180	540
I/O Expansions	Mark of the Unicorn 2408mk3 I/O expansion [7]	3	360	1080
Synxhronizer	MIDI Timepiece AV [8]	3	330	990
Acquisition PC with Licences		3	800	2400
Setup cost	Cabling, Mounting etc.			1500

£10995

Table 1: List of audio hardware

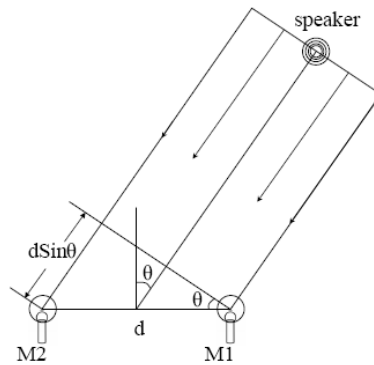


Figure 25: Time Difference of Arrival Measurement between two audio sensors.

---

## 7. Generation of planar images from omnivision

---

Involved partner: EPFL

The software developed in the context of the APIDIS project generates rectangular images taken by a virtual pinhole planar camera and calculated from the images captured by one or an array of omnidirectional sensors.

The inputs of the process are either the conventional parameters of the virtual pan-tilt-zoom camera (line-of-sight, resolution, angle-of-view) or a region of interest defined in terms of spherical coordinates, e.g. in response to a tracking or event detection algorithm.

The theoretical issue addressed here is the problem of mapping images between different vision sensors. Such a mapping could be modelled as a sampling problem that has to encompass the change of geometry between the two sensors and the specific discretization of the real scene observed by the two different imaging systems.

We formulate the problem in a general framework that can be cast as a minimization regularized problem with a linear operator that applies to any image geometry.

Successful experimental results on catadioptric images (taken in our laboratory) and on fisheyed omnidirectional images (provided by the first acquisition campaign) show the superiority of this approach with respect to alternative schemes based on linear interpolation or inpainting approaches : the accuracy of images and the sharpness of edges is clearly better.

An example of a planar image computed from an image taken by a fisheye camera during the first acquisition campaign is displayed below.

Figure 26 is a complete 1200 \* 1600 image acquired by the camera number 3 of the first acquisition campaign.

Figure 27 is a planar image generated from the fisheye image displayed on Figure 26.

Here a basic interpolation technique has been applied, and provide very satisfactory results with very sharp edges.

The parameters chosen by the user are the following: a very wide horizontal angle-of-view of 90 degrees for an image of size 800 \* 1066 pixels, with 1066 pixels along the horizontal axis.

The direction of view that defines the orientation of the image plane of the virtual pan-tilt-zoom camera is determined by a colatitude of 180 degrees, e.g., the virtual image plane is parallel to the ground.

The software that generates rectangular images from an array of omniscams is being developed incrementally, with the goal of having successive versions at T9, T18 and T24.

The next step consists first in improving the quality of images whatever the direction of view and the angle-of-view is.

Moreover, we will consider an array of omniscams to increase the resolution of the reconstructed planar images, and evaluate the gain of using several omniscams instead of one in terms of resolution and quality of produced images. Synchronization is in that case necessarily, so that two cameras describe the same reality at the same time. Images will then be produced when data of the second acquisition campaign will be available.



Figure 26: Original frame from camera3

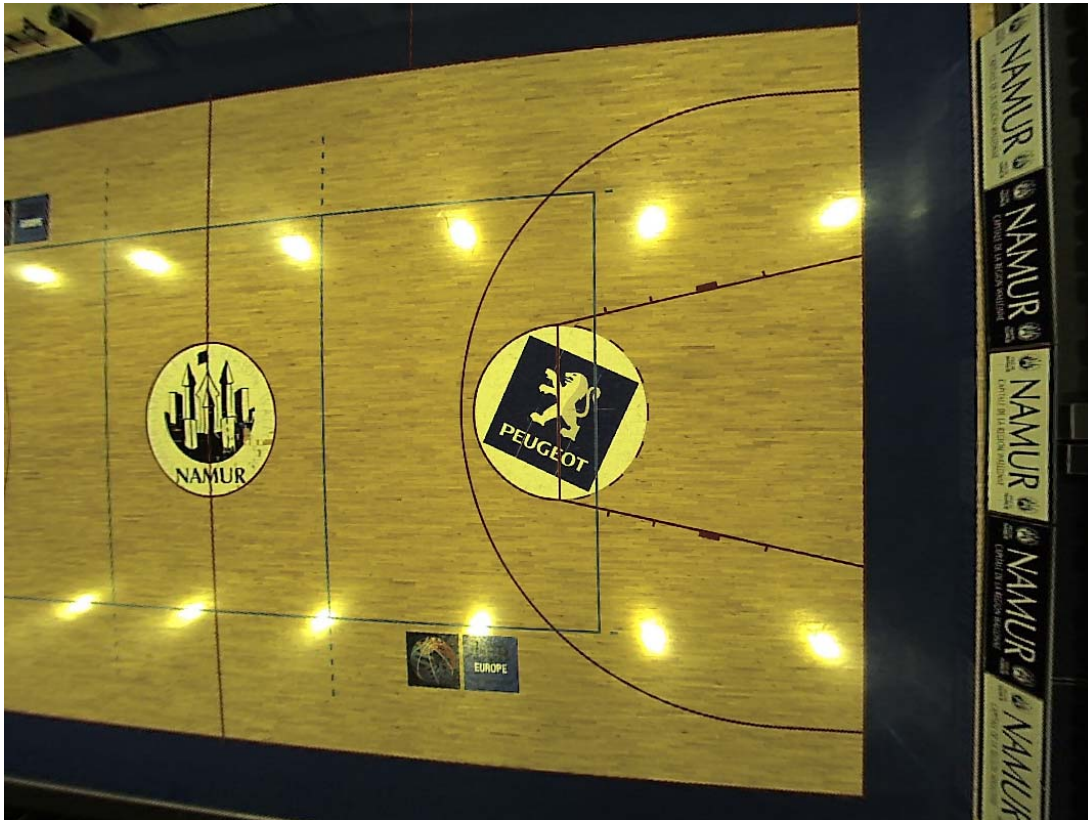


Figure 27: Generated planar image

## 8. Conclusions

---

The first deployments of the APIDIS system have been detailed. We have described the technical choices that have been made for the audio/video sensors and the acquisition servers. We have also discussed the content storage and annotation issues, the calibration of multi-sensors acquisition systems as well as the generation of planar images from omnivision.

The assessments of those acquisition systems show that:

- Although there are some minor deviations in the performances, the results are in line with what was expected from the different sensors
- The pictures quality corresponds to our expectations, even if it could be improved for nicer rendering. Furthermore, higher contrast would probably help the video processing algorithms. It seems reasonable to think that very high quality 2 MPixels cameras will be available in the next two years.
- The deployed systems fit the APIDIS requirements in terms of acquisition bandwidth capabilities and video processing constraints (resolution, etc.)
- A second basketball acquisition system should be deployed to
  - Measure the advantages and drawbacks of sensors synchronized with external triggers when compared to the current time stamping synchronization mechanism.

- Compare the video streams quality of Pulnix cameras with Arecont cameras in the same environment.
- Deploy and take advantage of the proposed new audio acquisition system.

## 9. References

---

- [1] APIDIS deliverable. D2.1 End-user requirements and architecture definition.
- [2] APIDIS deliverable. D2.2 Knowledge associated to content production.
- [3] Alexander Waibel and Rainer Stiefelhagen, Computers in the Human Interaction Loop (CHIL), Universität Karlsruhe (TH) Interactive Systems Labs. Available at: <http://chil.server.de/servlet/is/101/>
- [4] [http://www.audio-technica.com/cms/wired\\_mics/102fa42601dd18dc/index.html](http://www.audio-technica.com/cms/wired_mics/102fa42601dd18dc/index.html).  
Last accessed: 8 Dec, 2008
- [5] <http://www.motu.com/products/pciaudio/2408/>.  
Last accessed: 8 Dec, 2008
- [6] <http://www.behringer.com/ADA8000/?lang=ENG>.  
Last accessed: 8 Dec, 2008
- [7] <http://www.audiomidi.com/2408-MK3-Expander-I-O-P1346.aspx>.  
Last accessed: 8 Dec, 2008
- [8] [http://www.motu.com/products/midi/mtpav\\_usb](http://www.motu.com/products/midi/mtpav_usb).  
Last accessed: 8 Dec, 2008