**SEVENTH FRAMEWORK PROGRAMME**

**Networked Media**

*Specific Targeted Research Project*

# SMART

(FP7-287583)

# Search engine for MultimediA environment generated contenT

## D2.4 Data collection design

Due date of deliverable: 31-01-2012

Actual submission date: 09-03-2012

Start date of project: 01-11-2011 Duration: 36 months

# Summary of the document

| | |
|---|---|
| **Code:** | **D2.4-v1.0** |
| **Last modification:**<br>**State:** | 08/03/2012<br>Final |
| **Participant Partner(s):**<br>**Author(s):**<br><br>**Fragment:** | AIT, IBM, SDR, ATOS<br>Aristodemos Pnevmatikakis, John Soldatos (AIT)<br>Zvi Kons (IBM), Tomás García Fresno (SDR)<br>Josemi Garrido, Jeldrik Schmuch, Irene Schmidt (ATOS)<br>No |
| **Audience:** | ☒ public<br>☐ restricted<br>☐ internal |
| **Abstract:** | *This is D2.4 Data Collection Design. The document details the sensors, their setup and connection, as well as their arrangement, to facilitate data collection within the SMART project. It also describes the data to be collected for training and evaluating the various audio-visual algorithms in the project.* |
| **Keywords:** | • Audio-visual sensors: Cameras, microphones<br>• Data capturing<br>• Training data<br>• Sensor setup |
| **References:** | DoW |

# Table of Contents

# 1 Executive Summary

## 1.1 Scope

The main objective of the SMART project is to build a multimedia search engine, which could provide scalable search capabilities over environment generated content i.e. content captured by the physical world via sensors. A main part of the project will be allocated to the processing of multimedia content derived from visual and acoustic sensors (notably cameras and microphones). The purpose of this processing is to extract context from the surrounding environment of the sensors based on leading edge audio-visual (A/V) processing components. Such components will be employed in order to allow the SMART systems to perceive the status of the surrounding environment and accordingly to make this context available to the search engine for (later) indexing and retrieval, as required by the SMART applications. The development of robust A/V processing perceptual components is therefore a main part of the project and will be conducted within WP3 of the project.

A key prerequisite for the development (training and testing) of A/V processing components, is the availability of relevant data sets that will facilitate technology developers to build high-performance systems for the target physical environments (notably areas/locations of the city of Santander). The purpose of the present deliverable is to elaborate on the design of the respective data collection processes. This elaboration includes information about the sensors to be used for the data collection, the type of data to be collected, the format(s) of the data, as well as the pertinence of the data collection process to the SMART application scenarios. Furthermore, the present deliverable illustrates the locations where the sensors will be placed, along with the exact processes and timeline that will be followed towards processing the data.

## 1.2 Audience

The primary audience of this document consists of the people that will participate in the actual data collection process, including technology developers, infrastructure providers and people involved in the conduction of the on-site data collection. Primarily, the audience concerns members of the consortium (notably IBM, AIT, ATOS, SDR), who need to advance the data collection process in order to enable the development of A/V algorithms as part of WP3 of the project. We argue however, that this deliverable could also serve as a data collection example) to other (even third-party) technology developers and infrastructure providers. This is indeed the case for such parties who wish to deploy other (possibly) similar A/V processing systems, which could be later interfaced into the SMART engine. As such an example, the document is of wider interest to stakeholders of A/V processing and respective search, outside the SMART consortium.

## 1.3 Structure

The document is structured as follows: Section 2 provides the introduction.

Section 3 elaborates on the visual and acoustic sensors that will be used for the data collection. Apart from a presentation of the sensors, this section illustrates also the main rationale behind their selection (which was based on functionality-related, as well as techno-.economic criteria). Note that the sensors' selection detailed in Section 3, served also as a basis for initiating relevant procurement processes (notably by partner SDR) in order to acquire and later install the sensors.

Section 4 is dedicated to the presentation of the actual locations that the sensors will be placed in the scope of the data collection process. The selected locations are from the city of Santander, where real-life sensors will be deployed in order to provide a living showcase for the project's developments. Note that the selection of the locations was also influenced by a number of factors, including the envisaged functionalities to be demonstrated, the requirements of the city, as well as the selection of locations that could provide opportunities for demonstrating added-value capabilities of the SMART search engine. It must be also emphasized that the selected locations have undergone legal and ethical approval, by the respective Spanish authorities (Agencia de Protección de Datos, Data Protection Agency, DPA). Nevertheless, the issue of ethical approvals and related privacy issues is not elaborated in the present deliv-

erable, given that the project includes dedicated deliverables in WP7 (notably D7.7. and D7.8), where ethical/privacy management issues (even for the data collection process) are discussed.

Following the presentation of the selected sensors and their placement, Section 5 illustrates the operational details of the data collection process. Section 5 serves as a guide for the people that will be actually involved in the data collection process.

Finally, section 6 concludes the deliverable.

# 2 Introduction

The main goal of the SMART project is to design, develop and deliver a multimedia search engine for environment generated content (i.e. content stemming from the physical world). The SMART search engine will therefore index, retrieve and consume information derived by sensors (including information stemming from audio and visual sensors). At the heart of this engine will also be a number of A/V (perceptual) processing components, which will allow the SMART search engine to perceive the status of the surrounding environment, in order to make contextual information available for later search and retrieval. The role of these perceptual components for the SMART project is instrumental, given that they will act as the main sources of multimedia content on SMART, while at the same time adding intelligence to the system. Indeed, thank to such A/V processing components the SMART search engine will be able to sense, analyze and perceive the status of the physical world, which could enable the later creation of a wide area of added-value applications.

The development of such A/V processing components is highly dependent on the availability of rich sets of data captured from the locations where the sensors will be deployed and used for the purpose of showcasing the functionalities of the SMART engine. Such data sets will be used in order to train the visual and acoustic technologies to be used in the SMART project (including (but not limited to) visual person tracking, crowd analysis, color density analysis, speaker verification and acoustic event classification). The purpose of the present deliverable is to present the process of data collection design, which will lead to the collection of the required sets of audio and visual data. The design of the data collection process (as presented in the deliverable) presents the sensors to be employed, the formats to be used, the locations where the sensors will be placed, as well as the operational details of the actual data collection process. All this information has been specified following a careful consideration of a number of parameters that affect the data collection process including:

- The selected visual and acoustic sensors, which have been selected on the basis of a range of techno-economic criteria.
- The audio and visual components to be developed, which are at this stage highly driven by the target application scenarios (use cases) of the project. These use cases are currently under development and elaboration as part of the WP2 of the project and they concern the areas of live news and security surveillance. A detailed presentation of these use case is out of the scope of the present deliverable, since they will be elaborated in the scope of another WP2 deliverable (D2.2, dedicated to this task). However, the use cases specifications have been taken into account in the specification of the A/V processing functionalities that have to be supported by the data collection process.
- Requirements of the city of Santander regarding the placement of the sensors, taking also into account the functionalities to be showcased.
- The compliance with legal and ethical requirements, based on relevant approvals that have been requested and granted by relevant Spanish authorities (notably the DPA).
- The need to maximize the effectiveness of the data collection processing (including measures to reduce the required on-site efforts).

The data collection processes prescribed in this document will be carried out by consortium members in order to lead to appropriate datasets that will facilitate the development of robust high-performance A/V processing algorithms. Thus, the document is primarily targeting the involved consortium members. However, the present document could be of interest to other (thirty-party) infrastructure providers and technology providers, which are likely to engage in similar processes towards deploying A/V sensors and A/V processing components for similar search applications. Given SMART's intention and commitment to provide an open architecture for its search engine, the present document will be certainly of interest to other stakeholder wishing to deploy search infrastructures similar to SMART. At the same time, it would be also of interest for organizations wishing to deploy the SMART infrastructure in other physical locations. Overall, the readership of this document could exceed the boundaries of the SMART consortium, especially in the case where the SMART infrastructure will achieve a wide adoption (beyond the consortium's organizations).

# 3 Sensors Selection and Procurement

In this chapter, the sensors to be used for audio-visual data collection in SMART are described. Those sensors are both far-field, for capturing audio-visual scenes (images and sounds), and near-field, for capturing the speech of a user.

## 3.1 Visual Sensors – Cameras

The SMART visual sensors are all far –field, as the project is more focused on crowds than individuals. The only exception is the suspicious behavior of individuals under the security use case.

### 3.1.1 Main Requirements

The use of the visual sensors in SMART is for crowd and traffic analysis (estimate density, colours and incidents), environment analysis and individual person activity. For these, the selected cameras should have the following characteristics:

- Optical/colour
- Large field of view to be able to capture the complete scene from moderate distances
- High resolution to be able to perform analysis at different scales
- Day and night operation for 24 hours operation (see Fig. 3.1)
- IP connectivity
- Outdoor operation (weather-proof)

**Figure 3.1: Example of automatic switch to night view, after removing the IR cutoff filter.**

There is an additional set of desired features:

- Wide dynamic range, to be able to view both bright and dark areas in the same scene (see Fig. 3.2)
- Ability to pan, tilt and zoom to incident (see Fig. 3.3)
- Ability to connect audio sensor(s) for embedding audio into the produced video stream

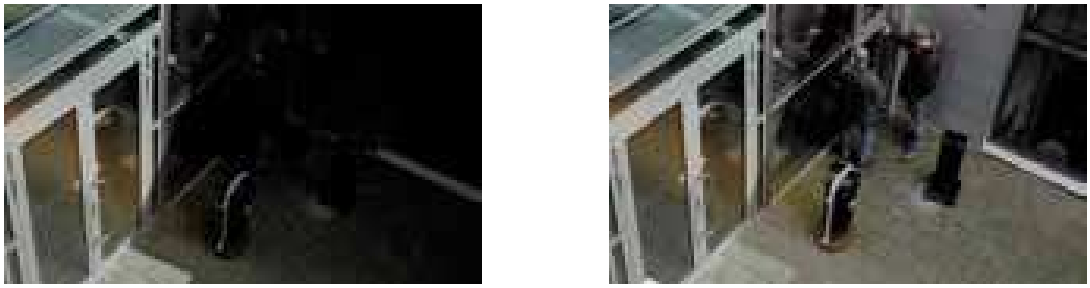

**Figure 3.2: Example of wide dynamic range capability, clearly view both dark and bright parts of a scene.**


**Figure 3.3: Example of pan, tilt and zoom capability, to focus on some insident of interest.**

### 3.1.2 Cameras Reviewed

Table 3.1 lists the cameras under consideration during the SMART kick-off meeting. All these cameras fulfill the set of mandatory features.

**Table 3.1: Cameras considered for initial deployment at SDR**

| Model | Unit price (€) | Type | Res | FPS | FOV (deg) | Min illumination (lux) | | Additional features | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Col | BW | Audio | Image |
| AXIS P1343-E | 863 | Fixed | SVGA | 30 | 61-21 | 0,2 | 0,05 | External microphone input or line input, line output | Compression, color, brightness, sharpness, contrast, white balance, exposure control, exposure zones, backlight compensation, fine tuning of behavior at low light, rotation, mirroring of images, **wide dynamic range - dynamic contrast** |
| AXIS P3343-VE | 898 | PTZ | SVGA | 30 | 72-34 | 0,2 | 0,04 | | |
| AXIS P1344-E | 953 | Fixed | 1MP | 30 | 72-28 | 0,3 | 0,05 | | |
| AXIS P3344-VE | 997 | PTZ | 1MP | 30 | 87-40 | 0,3 | 0,05 | | |
| AXIS P3346-VE | 1075 | PTZ | 3MP | 20 | 84-30 | 0,5 | 0,08 | | |
| AXIS P1346-E | 1235 | Fixed | 3MP | 20 | 72-27 | 0,5 | 0,08 | | |
| AXIS P1347-E | 1514 | Fixed | 5MP | 12 | 89-33 | 0,5 | 0,08 | | |
| AXIS Q6035-E | 3257 | PTZ | 1080p | 25 | 54-2 | 0,8 | 0,04 | | |
| AXIS Q1755-E | 1682 | Zoom/AF | 1080i | 30 | 50-5 | 2 | 0,2 | | Compression, brightness, sharpness, white balance, exposure control, backlight compensation, rotation, mirroring of images |
| AXIS P5532-E | 2186 | PTZ | 720p | 25 | 53-2 | 0,5 | 0,01 | No | **Wide dynamic range**, electronic image stabilization, manual shutter time, compression, colour, brightness, sharpness, white balance, exposure control, exposure zones, backlight compensation, fine tuning of behaviour at low light, rotation, aspect ratio correction, text and image overlay, privacy mask, image freeze on PTZ |

### 3.1.3 Final Selection

Table 3.2 lists the cameras under consideration during the SMART Kick-off meeting. The selection was based on balancing between cost, resolution and coverage of the additional set of desired features.

**Table 3.2: Cameras selected for initial deployment at SDR**

| No. | Description | Use in SMART | Quantity | Indicative Price |
|---|---|---|---|---|
| 1 | AXIS P3346-VE<br>http://www.axis.com/products/cam_p3346ve/ | Outdoors video recording | 4 | 1075 € per camera |

Note at this point that the SMART cameras are not the only source of video of SMART. The consortium is currently negotiating with the Santander authorities to gain live access to the traffic cameras in the city centre.

## 3.2 Acoustic Sensors – Microphones

### 3.2.1 Main Requirements

The use of the acoustic sensors within the SMART project is to capture environment sounds and noises such as crowd noises, traffic, music and other types of events as will be defined by the different scenarios. The microphones are expected to operate either outdoor or indoor as will be defined by the scenarios.

General requirements from the microphones:

- High sensitivity sound capture
- Frequency response at least 50Hz-16Khz

Additional requirements for static outdoor microphones:

- Ability to operate in all relevant weather condition relevant to the deployment area (rain, wind and temperature)
- Wind protection to reduce wind sounds
- Rain protection to reduce rain sounds

### 3.2.2 Microphones Reviewed

For outdoor deployment the microphones listed in Table 3.3 were reviewed.

**Table 3.3: Microphones considered for initial deployment at SDR**

| Company | Model | Description |
|---|---|---|
| Brüel & Kjær | BK 4198 | http://www.bksv.com/Products/TransducersConditioning/AcousticTransducers/Microphones/4198.aspx |
| G.R.A.S. | 41CN | http://www.gras.dk/00012/00013/00032/00113/ |
| Larson Davis | 426A12 | http://www.larsondavis.com/PermanentOutdoorMicrophone_Preamplifier.htm |

For indoor deployment, the indoor versions of the reviewed video cameras come with built-in microphones. For those cases the quality of the built-in microphone would be reviewed based on the selected camera models.

### 3.2.3 Final Selection

For outdoor microphones the following model was selected based on its specifications and added features as compared to the others.

**Table 3.4: Microphone selected for initial deployment at SDR**

| No | Description | Use in SMART | Quantity | Indicative Price |
|---|---|---|---|---|
| 3 | Larson Davis 426A12-RI<br>http://www.larsondavis.com/PermanentOutdoorMicrophone_Preamplifier.htm | Outdoors audio recording | 4 | N/A |

In case where directional recording is needed than the model should be 426A12-FF.

## 3.3 Speech collection

### 3.3.1 Main Requirements

Speech recordings will be used within the SMART project for speech transcription and for speaker verification / identification. Both those applications perform better when the speech is clear with minimal amount of environment noises. The requirements from those microphones are:

- Close range microphone
- Directionality and noise isolation preferred
- Frequency response of at least 50Hz-16KHz

### 3.3.2 Selected method

Microphones in regular telephones are usually adequate for this task. Two options are available:

- Phone call over telephony channel to a voice messages box where the messages are recorded.
- Direct recording using a dedicated application running on a smartphone. The recorded sample is sent to the processing server.

The second method is the preferred one.

# 4 Sensor Placement

All the following locations have been pre-selected taking into account two main aspects. On the one hand, all of them are located in areas close to the city centre and therefore with a huge affluence of people and, on the other hand those locations have some access point to the Santander Council's optical fibre network which will facilitate the transmission of image and sound.

The description of the candidate locations is accompanied with aerial photos and plans. On both, the preferred location for installing the A/V sensors is denoted by a red circle.

The six candidate locations are shown in Fig. 4.1.


**Figure 4.1: Overview of the six candidate locations.**

## 4.1 Location 1: Plaza del Ayuntamiento

This square is the geographical and business heart of the city. Therefore there is always a huge amount of people and activities of all kinds. Locating a camera either on top of the Town Hall or on some of the neighbouring buildings will provide an excellent vision of all the activities in the square. In the case of microphones, the lamp post in the square could be the most suitable location for installation.


**Figure 4.2: Plaza del Ayuntamiento, photo and plan.**

## 4.2 Location 2: Plaza del Príncipe

This is another of the most crowded places in the city with a huge affluence of people due to its commercial activities. This square has an access point to the Council's network close to the point indicated in the pictures.


**Figure 4.3: Plaza del Príncipe, photos and plan.**

## 4.3 Location 3: Mercado de la Esperanza

In this square, just behind the Town Hall, is located one of the most popular markets in the city, where hundreds of people go every day for shopping. The house marked by a red point in the map is a public building where the devices could be installed.


**Figure 4.4: Mercado de la Esperanza, photos and plan.**

## 4.4 Location 4: Alameda de Oviedo / Plaza de Numancia

This other location is also one of the most popular avenues in the city of Santander. Due to the proximity to the Bus Control Centre of the city, it is also an excellent place for the installation of required devices that access the network from City Hall. However due to its location , trees covering the avenue could raise some problems in the image treatment.


**Figure 4.5: Alameda de Oviedo / Plaza de Numancia, photos and plan.**

## 4.5 Location 5: Paseo de Pereda

Pereda's Park is one of the most emblematic points of the city of Santander. Next to this location the construction of a modern art centre that will increase the popularity of the place even more, is planned. The point proposed for the location of cameras and microphones in this place is on top of the local tourist office.


**Figure 4.6: Paseo de Pereda, photo and plan.**

## 4.6 Location 6: Mercado de México

Mexico's Square is next to Mexico's Market and the local bullring. Due to the proximity of these buildings the flow of people and vehicles is more than assured throughout the year. Especially during the summer, the square around the bullring is crowded with thousands of people. Since all mentioned buildings are owned by the town house, access to the corporate network is guaranteed.


**Figure 4.7: Mercado de México, photo and plan.**

## 4.7 Preferred locations

The number of possible collection sites is limited by the number of available video cameras and microphones. Since our budget is limited we need to select two sites in which we'll perform the data collection.

Currently the two preferred locations are 1: Plaza del Ayuntamiento and 6: Mercado de México. Those two locations are preferred since they both hold a lot of activity and both are located near municipal buildings which allow easy connection to the municipality infrastructure.

The final decision on the data collection locations depends on the findings from pre-collection stage (see section 5.4.1).

# 5 Audio-Visual Data Collection Process

The data collection process aims at providing the consortium with adequate material to develop and train algorithms that perform audio-visual analysis.

## 5.1 Data Collection system

### 5.1.1 Microphone to camera connection

The selected microphone already contains the preamplifier and can be connected to the recording equipment with BNC connectors. The microphone also requires a 12V DC power supply through a LEMO EXG.1B.307 7-pin Female connector.

If the microphone isn't connected to the video camera than additional setup is needed for recording the audio. The recording setup should include a PC with a software tool that allows the control of the recording (for example continuous recording or time lapse recording) and storing the audio in a proper format.

### 5.1.2 Configuration of the sensors

Capturing video should happen by a camera looking down to the monitored public space from a height of at least three meters, to avoid occlusions by the crowds. Heights even larger than that are still OK, especially since the camera has zoom capabilities. Pan and tilt capabilities also give us freedom as to the exact placement. The recording frame rate needs to be at least 10 frames per second. Rates smaller than those reduce the correlation between the frames and endanger tracking results.

The microphones should be placed in the vicinity of the events but not too close to any other noisy objects. For example, to record the crowd noises the microphones can be placed on a street lamp.

The microphone amplifier levels should be carefully adjusted to avoid overflow of the signal in case of loud noises while maintaining sufficient recording levels for other noises. This might require some adjustments over time in case the conditions change.

The actual recordings for the data collection will utilize the software for audio-visual recordings of the camera. The microphone will be connected to the camera to ensure synchronization of the audio and visual streams.

### 5.1.3 Networking to the host PC

The network should be carefully designed so as to ensure connectivity, minimizing interference, maximizing performance and Quality of Service. Performance is generally associated to Bandwidth and QoS. An overview of the networking approach that will be followed in SMART is depicted in Fig. 5.1.

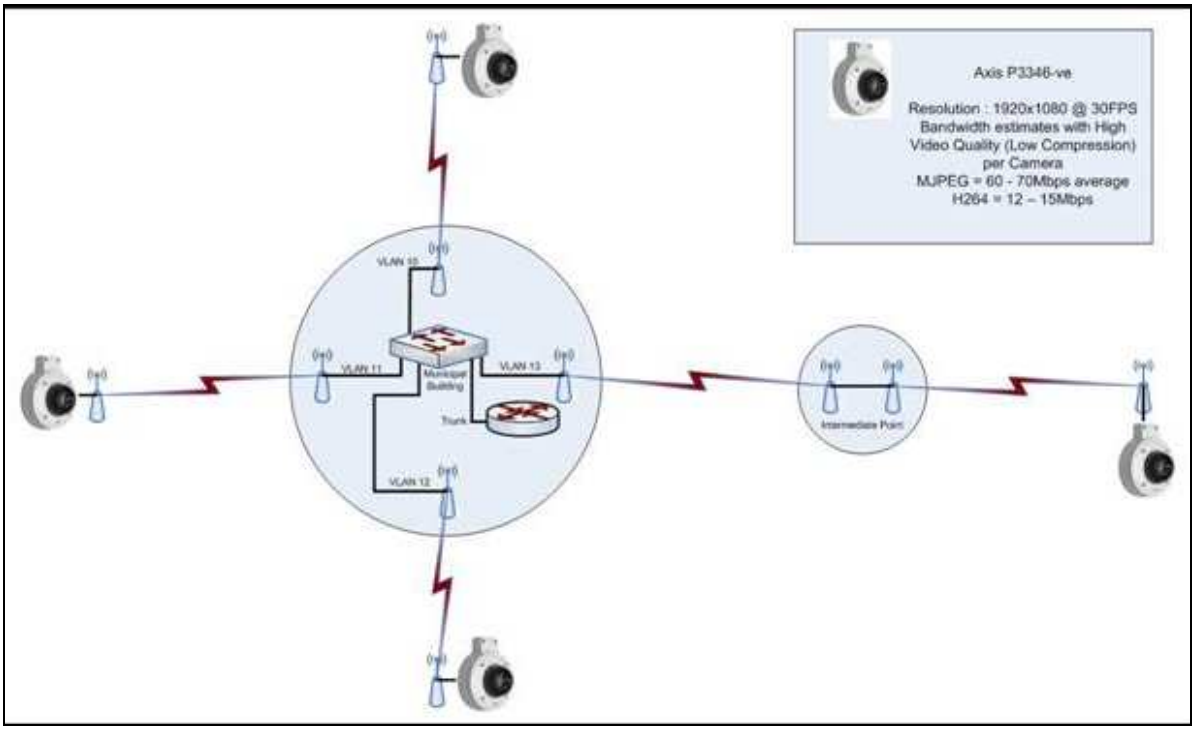


**Figure 5.1: Network interface of multiple IP cameras to the SMART edge node.**

The initial requirements for the various components are discussed next. Regarding the cameras:

- Each Camera can be directly connected to a Wireless Bridge.
- Each Camera requires extra bandwidth for pan-tilt-zoom operations.
- Camera bandwidth requirements decrease with increasing shutter speeds (Night mode).

The involved network has both a wireless and a wired part. For the wireless bridges we need to consider connectivity, interference and performance issues. Regarding connectivity:

- The primary consideration for the design is ensuring Line-Of-Sight (LOS) conditions. If LOS conditions are not feasible, e.g. in the case of trees or worse buildings, OFDM transmission (which is used by carrier-grade wireless equipment) allows the operation in nearLOS and NoLOS conditions.
- In the case of a camera node serving also as an Intermediate Point in the network, one may connect two Access Points (in a back-to-back configuration).
- Preferably a managed network switch should also be included for future administration.

Regarding interference:

- Each wireless bridge should operate at 5.4GHz to avoid interference (in contrary to 2.4GHz where significant interference is expected in a typical city environment).
- The access point antenna must be directional. I.e. to avoid causing interference to adjacent cameras or third party equipment, a selection of directional (narrow-beam) 30º or 60º panel antennas should be made.
- Power budget: Longer distance links require higher power output, however power (or EIRP, i.e. the product of power times the antenna gain) is limited and the limit is 1W for most of Europe).

Then, regarding performance/QoS:

- Each Access Point must support MiMO (Multiple-in-Multiple-Out).
- The design should avoid a situation where more than two intermediate points (hops) are used to reach a camera.
- The primary principle to be considered in the design is that the number of Network appliances should be kept to a minimum so as to ensure minimum latency.
- Designing a network with higher than required initial bandwidth requirements will future-proof

the network and guarantees perfect camera operation in any configuration.

Finally, for the wired network part:

- Each Access Point/Camera combo must reside in its own VLAN.
- The switch must forward these VLANs in a trunk.
- The switch must connect to a (VLAN-capable) Router via Gigabit Ethernet.
- The video server must connect to Router via Gigabit Ethernet.
- No QoS requirements as no other traffic will be present.

### 5.1.4 Speech collection system

The speech collection should be done with dedicated software written for mobile devices. The software should be able to record a message and send it along with any additional information (e.g. location, user name etc.) needed to the server.

A collection of a large corpus of messages is required for training the speech software, In order to simplify this process a part of the data can be collected using more common means such as high quality microphone connected to a PC with an audio recording software (such as audacity)..

## 5.2 Format

### 5.2.1 Audio format

The preferred format for the audio recording is uncompressed 16 KHz, 16 bit wav format files. Since the amount of data collected using this format is large, a compressed format would be acceptable too.

The compression should be done using high-quality audio compression methods such as MP3, AAC or OGG Vorbis. The bitrate should be at least 100kbps and with at least 16 KHz bandwidth.

### 5.2.2 Speech format

Speech files should be in uncompressed 16 KHz, 16 bit wav format. If telephony channel is used (i.e. the speech is recorded after passing through a telephone line) then the samples will be in 8 KHz.

### 5.2.3 Video format

The selected camera offers the video stream in two formats:

- H.264 (MPEG-4 Part 10/AVC), Baseline Profile, or
- Motion JPEG

The supported resolutions range from 2048x1536 (3 MP) to 160x90 and the frame rates are 20 FPS for 3 MP and 30 FPS for smaller resolutions.

The preferred video format for data collection is H.264 for size constraints, but on the other hand the quality of the video and the relative bandwidth difference between the two formats need to be checked before any decision is made.

Regarding video streaming to the algorithms, motion JPEG is preferred due to speed of decompression and lack of severe compression artifacts. Again, the final decision is deferred to after experimenting with the camera itself.

For storing the A/V recorded material, the consortium will consider possible offline transcoding from the native camera formats to any high quality compression format deemed necessary. This will depend upon the quality of the native H.264 stream, both in terms of compression artifacts and compression rate. Should any of the two be considered unacceptable, then the recording will be initially in MJPEG and will be transcoded to high quality H.264 using the open source transcoders in FFMPEG [1].

## 5.3 Volume

### 5.3.1 SMART outdoor audiovisual recordings

To be able to develop (train, test and/or tune) the audio-visual algorithms that sense the environment and reason about events of interest, we need an audio-visual corpus rich in these events. If the events are not staged, this can require days of recordings to accomplish. Certainly the corpus can then be shortened, to contain mostly the relevant portions of the original data.

Apart from the events, the environmental conditions should also be captured. Wind and rain affect audio recordings. Rain also affects video recordings. Also, night and day affect video recordings dramatically. All these conditions should be included in the corpus.

For the audio data we can distinguish between two types of events:

- Long term events (e.g. live music). For those we need at least 10 occurrences of the event recorded under different conditions. Each occurrence should contain at least one minute of audio.
- Short term events (e.g. yelling, door opened/closed). We'll need at least 20 occurrences of each events recorded under different conditions.

For the video data we are interested in individuals and crowds. Regarding the former, no data is needed. The partners have enough development data from past projects. Regarding crowds, data is needed at different stages of crowd formation. Also, should the crowd be formed due to an event, then the crowd concentrates in a location. This should be captured. The exact event (accident, street musician performing, etc.) is irrelevant to video.

### 5.3.2 Santander traffic recordings

The collection of visual data from the traffic cameras at Santander is already completed. Clips of five minutes each have been collected from eleven cameras. Sample frames are shown in Fig. 5.2 and 5.3.


**Figure 5.2: Sample frames from the 11 traffic cameras.**


**Figure 5.3: (cont.) Sample frames from the 11 traffic cameras.**

### 5.3.3 Speech

For the voice transcription we need to build a Spanish voice & language model which gives good representation of the channels (e.g. microphones and telephone channels), the people and the language domain. For that we need at least 40 hours of speech from many different users. This speech can also be used for the speaker identification task. We suggest the following collection protocol:

- Data will be collected from 200 speakers
- The speakers should be native Spanish speakers
- Each user will hold two recording session: one indoor and one outdoor

In each session the user will record about 50 sentences (3-5 min). Each recording session should consist of two parts. In the first part the user should speak freely within the relevant domain. This can be achieved by asking leading questions. For example:

- Describe the weather today
- Tell us about a recent event you saw
- Where do you live

The second part should be reading random sentences from a script. The script should include sentences giving good coverage of the relevant domain. For example the script should include:

- Eskup and news messages
- Street names, location names and people names
- Description of different events
- General information such as numbers, days of the week, months, hours currency, etc.

For the speaker identification task additional data should be collected for a subset of about 50 speakers:

- 4 recording session (indoor/outdoor)
- In each session:
  - Name (or username) repeated 3 times
  - Count from 0 to 9 (or other phrase which is similar to all users) repeated 3 times
  - Free text speech

Additional large text corpus of the relevant domain (e.g. eskup archive) is needed for building the language model.

For the speaker identification we need at least 3 training samples from each speaker and additional test samples.

## 5.4 Process Description

The process of data collection is split into three phases: pre-collection, sensor and recording setup and data collection.

### 5.4.1 Pre-collection

In order to finalize the selection of the recording sites, and get a crude understanding of the sensor placement, a pre-collection phase is planned. In this phase we will be using an amateur video camera with its microphone and a tripod, in order to capture some video clips at various locations and angles. Audio data can be similarly collected using the video camera or with a microphone connected to a recording device (e.g. laptop).

Based on these, we will be ready for placing the actual equipment once procured. On the other hand, early algorithmic development can use these data.

### 5.4.2 Sensor & recording setup

At this stage the actual sensors are procured, and some dummy recordings are carried out to fine-tune settings, sensor positioning and targeting and finalize recording formats based on severity of compression artifacts.

This phase is necessary and cannot be carried out prior to equipment procurement.

### 5.4.3 Data collection

When the equipment described in sections 3.1.3 and 3.2.3 is acquired and installed, we can proceed to the actual data collection,

Given the considerations in the previous sections, the data collection process will comprise of several recordings, in order to try to capture as many conditions (day/night, weather) and events (crowds, activities, short and long term acoustic events) as possible. The following list gives the possible combinations of what and when to record:

- What (Events): Passing by at various crowd densities, gathering, yelling, banging noises, loitering and isolated crossing.
- What (Weather): Sunshine, cloudy, rainy, windy.
- When: Rush and off-peak hours of working days, holidays/weekends, day and night.
- How much: One-hour segments for long term events, 10-minute segments for short term events (yelling, banging, loitering, isolated crossing).

For the live news scenario, data collection can be un-staged. But for the security scenario some staging might be needed for events like loitering and crossing a restricted area. In any case, pending an answer from the DPA, the data collection might all be staged, or at least the public may have to be notified about the recording and somehow give their consent.

# 6 Conclusions

Data collection plays a paramount role in the development of perceptual algorithms. Through this process datasets of relevant data are collected, which can then be used for training and testing the algorithms.

This document has described the hardware and software options for the data collection, covering:

- Audiovisual sensor selection
- Recording sites and sensor placement
- Type of material to collect (recording scenarios)
- Formats for recordings
- Collection process

Due to open questions regarding the data collection process stemming from Spanish law, the consortium is currently waiting for the advice of the Data Protection Agency on the matter. Based on their recommendation, the placement and process will have to be fine-tuned.

Also, tests need to be run (described in the collection process) in order to finalize the format of the data.

For the above two reasons it is possible that some information in this document will change. The updated information will be provided as part of D3.5, Audio and Video Data.

# 7 BIBLIOGRAPHY AND REFERENCES

[1] FFMPEG: A complete, cross-platform solution to record, convert and stream audio and video, version 0.9, Dec. 2011. http://ffmpeg.org/