EBU **TECHNICAL**

MEDIA TECHNOLOGY & INNOVATION

P/SCAIE First Call for Technologies

Author: EBU P/SCAIE (http://tech.ebu.ch/groups/pscaie)

Target: **open public**

Date: June 5th 2009

Motivations & Rationales

"Content is King, Metadata is Queen". The audiovisual entertainment sector adheres to this old adagio, but it is a fact that without metadata content is unusable.

The creation of metadata was until recently seen as the sole task of the archivist, responsible to create high quality metadata, including technical, descriptive, administrative, and even legal information. However, although archivists create meaningful and high-quality metadata, the information is too limited to allow optimal reuse of content. This is a logical consequence of the limited manpower and time an archivist can spend to describe an increasingly faster growing amount of audiovisual material. As a result, reuse of content is hampered and programme researchers spend too much effort to find and retrieve the desired material from archives.

One solution to increase the amount of metadata is the use of automated tools that extract information directly from the audiovisual asset by analysing the audio and/or visual stream.

The EBU P/SCAIE technical group investigates the availability of automated information extraction tools that can be deployed within broadcasting facilities. As such, the group has collected a variety of relevant broadcast material (audio and video) as a test-set.

This first Call for Technologies (CfT) is targeted at researchers developing automated information or feature extraction tools[1]. Within this call, solutions for one (or more) of the following problems are requested:

1. Speech Recognition
2. Audiovisual Segmentation – Shot, Audio and Scene Segmentation
3. Content Summarisation
4. Copy/Repetition Detection

[1]You can register at p-scaie-rfts@list.ebu.ch to be kept informed about the public Calls. To be registered to the list, please send an e-mail to Mr. Jean-Pierre Evain (evain@ebu.ch).

Call for Speech Recognition Tools and Technologies

For certain kinds of material, e.g. news and documentaries, the spoken content accounts for most of the semantics of the programme. In addition, speech to text results have also a value *per se*, that is to tell exactly what have been said, a kind of annotation that is too expensive if made by humans. Speech Recognition consists in translating spoken utterances in written electronic text, augmented with timing information of words and identification of speakers.

Requirements

1. A technology or a set of technologies is requested, which should be able to automatically recognise
    a. The spoken content of a piece of material
    b. The time reference of each transcribed word, with preferably a declared confidence level
    c. The number and occurrences of speakers along the material timeline, together with their gender, dialect, accent, and bandwidth of the communication (e.g., phone talk).
    d. The number and occurrences of speech patterns (pauses, interposes) and of sound patterns (music, noise) along the material timeline.
    e. The speaker turns and sentences breaking points along the material timeline.
2. The provided text should have a correct multilingual encoding of characters.
3. The delivery format of the results must contain all the required information, and preferably follow a known or documented standard format or practise, like e.g. MPEG-7. When available, the results must be returned in accordance with the metadata model and format defined by P/SCAIE for the specific task.

Evaluation parameters

The following parameters will be evaluated by P/SCAIE in order to assess the requested technology:

1. For material for which ground truth transcription is provided:
    a. Training performance in terms of word error rate using word alignment measurements (e.g., using the sclite [2] package). Missed, mislead and wrongly inserted words will be taken into account.
    b. Test performance in terms of word error rate using word alignment measurements. Missed, mislead and wrongly inserted words will be taken into account.
    c. Time alignments of a sub-sample of the transcribed words.
    d. Speaker clusters alignments.
2. For material for which ground truth transcription is not provided
    a. Qualitative evaluation of the results made by a native speaker.

Evaluation will be done on a reference set of material clips, taken from the P/SCAIE material library. The reference set will be subdivided in a training set and a test set for material for which ground truth transcription is provided.

Contacts

Interested people can contact P/SCAIE Chairman Alberto Messina (a.messina@rai.it) or EBU Project Manager Jean-Pierre Evain (evain@ebu.ch), for further information, clarifications, expression of interest in conducting experiments, comments and feedbacks.

A copy of this document is also available at http://tech.ebu.ch/pscaie_rfts.

[2] available for download at http://www.itl.nist.gov/iad/mig//tools/

Call for Audiovisual Segmentation – Shot, Audio and Scene Segmentation Tools and Technologies

Annotating audiovisual material for archive purposes typically starts by segmenting the audiovisual material, i.e, segmenting the video in shots[3] and scenes[4] and the audio into speech, music, silence (and other) blocks. Archivists typically annotate audiovisual material on scene level – i.e. the logical unit of work.

Identifying the shot, audio part, and scene boundaries is a tedious and time-consuming work. Using technology or a set of technologies which disposes of this work, frees up useful resources. There are many different criteria to perform the segmentation on multimedia content, depending on the use, level of description, purpose of the description. Some examples are shown in Figure 1 below. These different segmentations can be used and/or combined.

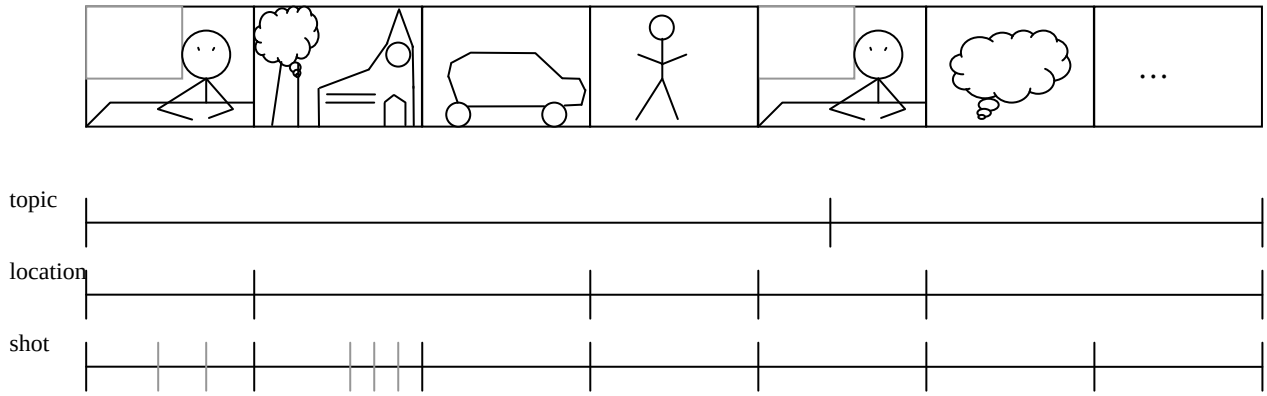

Figure 1: Different segmentation criteria example

Apart from the baseline visual shot transition criterion for segmentation (hard cuts, dissolves, wipes included), in general we foresee five main criteria for semantic segmentation: change in *location*, real or fictitious *time*, *actors*, *action*, and *topic*. In addition, we consider four types of audio features the transitions among which may generate a reason for segmentation: *speech*, *music*, *silence*, and *background noise*.

Therefore, audiovisual segmentation is a procedure that is able, given audiovisual material as input to

- identify the start and end of the different shots within the visual stream of the material
- identify the start and end of the audio blocks containing a particular type of content (e.g., speech, music, silence, background noise) in the audio stream of the material
- identify the start and end of the scenes in the audiovisual material, indicating which is the reason for changing the scene (e.g., change in location, real or fictitious time, actors, action, topic)

Note, each segment (shot, audio, scene) does not need to be aligned with each other. In other words, the start of a scene does not necessarily have to coincide with the start of a shot or audio segment. However, scene segmentation can (and should) use the outcome of low-level feature

[3] A shot (or take) is a continuous segment of motion picture film, created of a series of frames that runs for an uninterrupted period of time.

[4] A scene may be e.g. defined as a section in the audiovisual work such that the action takes place in a single location at the same time with the same actors.

extraction algorithms (such as shot and audio segmentation, and others) to improve the consistency and quality of its results.

A segment is time-boxed section whereby

a) The start of the segment is the first frame or audio sample containing visual and/or aural information about the particular segment.
b) The end of the segment is the last frame or audio sample still containing some visual and/or aural information about the particular segment.

Requirements

A technology or a set of technologies is requested which should be able to automatically identify the shots, audio and/or scene segments within audiovisual material, with the following requirements:

1. The technology must return a list of segments whereby for each segment the type (i.e., shot, audio, scene) and the start and end indication of the time-boxed section is returned. The start and end of a segment is defined as above.
2. An indication on the reason why the end of the segment was at that particular location should be given (e.g., shot ended by hard cut, shot cross-fade ended, scene location changed, ...).
3. If the segment is visual in nature (i.e., shot or scene segments), an indication of the most representative frame(s) of the given segment should be returned.
4. The delivery format of the results must contain all the required information, and preferably follow a known or documented standard format or practise, like e.g. MPEG-7. When available, the results must be returned in accordance with the metadata model and format defined by P/SCAIE for the specific task.

Evaluation parameters

It is not requested that a single tool must be able to perform all the mentioned segmentation types (shot, audio or scene). In any case, the segmentation types will be evaluated separately for each tool.
The following parameters will be evaluated by P/SCAIE in order to assess the requested technology:

1. Objective evaluation of the results by measuring recall and precision rates including the number of correct hits, false positives, and false negatives.
2. Required computational power and execution time.
3. In case that the technology has to be trained, the robustness and generalisation capabilities w.r.t. different conditions.

Evaluation will be done on a reference set of material clips, taken from the P/SCAIE material library. The material will be provided at different encoding quality levels, and assessment will be done at each of the defined levels.

Contacts

Interested people can contact P/SCAIE Chairman Alberto Messina (a.messina@rai.it) or EBU Project Manager Jean-Pierre Evain (evain@ebu.ch), for further information, clarifications, expression of interest in conducting experiments, comments and feedbacks.

A copy of this document is also available at http://tech.ebu.ch/pscaie_rfts.

Call for Content Summarisation Tools and Technologies

Nowadays, the explosion of new services based on the Internet and on mobile devices (e.g., Web TV, Mobile TV, IPTV), has a huge impact on how media is consumed. In fact, new delivery models and environments require that content must be adapted to these new conditions in a seamless way for the users and the content creator.

Manual adaptation requires employment of expensive resources, therefore, in this context, automatic mechanisms may offer a facilitation. In particular, content summarisation mechanisms make the typical ad-hoc and brief consumption of media on the Internet, mobile devices and also on archives feasible.

Content summarisation is a procedure that is able to produce a synthetic shortened version of a multimedia item. This can be done in two ways, potentially dependent on each other:

a) Remove the redundant frames so a synthetic version is created which retains the same amount of the information expressed in the input content but in a more condensed form (information redundancy elimination).

b) Producing a derived version retaining part of the original amount of information expressed in the input content (information skimming), w.r.t. to a pre-defined context: location, topic, time, actor, action of consumption (e.g., create a 15 minute version from a 30 minute sports video asset. That shortened version should contain only information about soccer and tennis).

Requirements

A technology or a set of technologies is requested, which should be able to automatically produce the content summarisation of a multimedia item given a pre-defined context of consumption, with one or more of the following requirements:

1. The technology must be able to perform content summarisation by analysing the multimedia content (i.e. video, audio, spoken text, visualised text) itself.
2. The technology should not show statistically different detection performance with the variation of the encoding quality and of the resolution at which the content is analysed.
3. The delivery format of the results must contain all the required information, and preferably follow a known or documented standard format or practise, like e.g. MPEG-7. When available, the results must be returned in accordance with the metadata model and format defined by P/SCAIE for the specific task.

Evaluation parameters

The following parameters will be evaluated by P/SCAIE in order to assess the requested technology:

1. Subjective quality evaluation of the produced items out of the information redundancy elimination process
2. Subjective quality evaluation of the produced items out of the information skimming process
3. Calculation complexity
4. Variation with encoding quality and resolution

Evaluation will be done on a reference set of material clips, taken from the P/SCAIE material library. The material will be provided at different encoding quality levels, and assessment will be done at each of the defined levels.

Contacts

Interested people can contact P/SCAIE Chairman Alberto Messina (a.messina@rai.it) or EBU Project Manager Jean-Pierre Evain (evain@ebu.ch), for further information, clarifications, expression of interest in conducting experiments, comments and feedbacks.

A copy of this document is also available at http://tech.ebu.ch/pscaie_rfts.

Call for Copy / Repetition Detection Tools and Technologies

Broadcasters create archives in order to facilitate the reuse of audiovisual content. A news item, for example, might use archived material to underpin the story. Furthermore, the news item itself might be reused in several other news broadcasts, whether slightly edited (e.g., shortened or completed with new footage) or in its original form. Detecting when material is reused is of particular interest to a broadcaster. For example, it allows the optimization of archive search results whereby multiple copies are displayed as one grouped result. It also makes detection of the original raw footage possible, as well as several other applications.

Copy / repetition detection is a procedure that is able, given a short audiovisual clip as input, to detect all copies in a set of audiovisual material. There are three principal approaches to the problem:

a) Analysing the visual information in the input clip and cross-referencing it to the set of visual footage;
b) Analysing the audio information in the input clip and cross-referencing it to the set of audio footage;
c) A combination of these two approaches.

Requirements

A technology or a set of technologies is requested which should be able to automatically identify all copies of a given clip in a set of audiovisual material, with the following requirements:

1. The technology must return time offsets (start time and duration in each clips) and descriptions of the relationships between all detected copies;
2. The technology should be invariant to any editing operation, such as scaling, cropping, rotation, addition of graphical components (overlay, logos, etc.), slow-motion, fast-motion, video and audio transcoding, etc.;
3. The delivery format of the results must contain all the required information, and preferably follow a known or documented standard format or practice, such as MPEG-7. When available, the results must be returned in accordance with the metadata model and format defined by P/SCAIE for the specific task.

Evaluation parameters

The following parameters will be evaluated by P/SCAIE in order to assess the requested technology:

1. Objective evaluation of the results by measuring recall and precision rates including the number of correct hits, false positives, and false negatives;
2. Required computational power and execution time.
3. In case that the technology has to be trained, the robustness and generalisation capabilities w.r.t. different conditions

Evaluation will be performed on a reference set of material clips, taken from the P/SCAIE material library. When material is provided at different encoding quality levels, an assessment will be performed at each of the defined levels.

Contacts

Interested people can contact P/SCAIE Chairman Alberto Messina (a.messina@rai.it) or EBU Project Manager Jean-Pierre Evain (evain@ebu.ch), for further information, clarifications, expression of interest in conducting experiments, comments and feedback.

A copy of this document is also available at http://tech.ebu.ch/pscaie_rfts.