Feature extraction in video signals with the aid of available metadata to crop important image regions and adapt them on displays with lower resolution

Mobile TV (Mobile Television) is a growing and certainly promising market. It allows the reception of Television signals on small portable devices like cell phones, smartphones or PDAs (Personal Digital Assistant). As it can be imagined, the display of those devices does not provide such a detailed image as it is known from the common TV at home (currently SDTV, Standard Definition Television). Despite this essential difference of viewing conditions, the same content is mainly shown on both TV systems. On the other hand, producing a separate programme for mobile TV causes an expenditure of human labour as well as costs a broadcaster hardly can bring up.

Therefore, an adaptation from video content with a high image resolution to smaller displays by cropping parts out is obvious. The presented approach in this patent application deals with the automatic detection of regions of interest (ROI) using feature extraction with common video analysis methods. Once detected regions in a video signal are used to find an adequate crop area and compose a new image containing all relevant information adapted on displays of handheld devices.

Cropping parts of an SDTV image is not something new, but the reliability of such systems is mostly not sufficiently enough to deal with a wide range of content. Mainly, those systems fail because of missing semantically knowledge and thus general defined methods.

The approach presented here makes use of the development on TV production side to increase the reliability of such a system. In this context, it is important to know, that new production workflows are essential, which are only feasible by changing from tape to tapeless, i.e. file-based production formats. This shift allows the introduction of various metadata for post-production, programme exchange and archiving. These metadata contain content related information describing the type of genre as well as specific information of the production procedure.

Comment [JD1]: Maybe, this can be explained more general, i.e. not just for TV productions

Combining the information of these descriptive data and feature extraction methods would provide an approach being much more individually adaptable then methods developed up so far.

The introduced system deals with this approach and parses available metadata to apply content tailored feature extractions. Figure 1 depicts the overall system and Figures 1.1 – 1.3 describes each box of Figure 1 more in detail. The basic prerequisite for applying this process is a container format containing video and metadata. Such a container format allows a multiplex of different data in a synchronised way, either as file or stream.

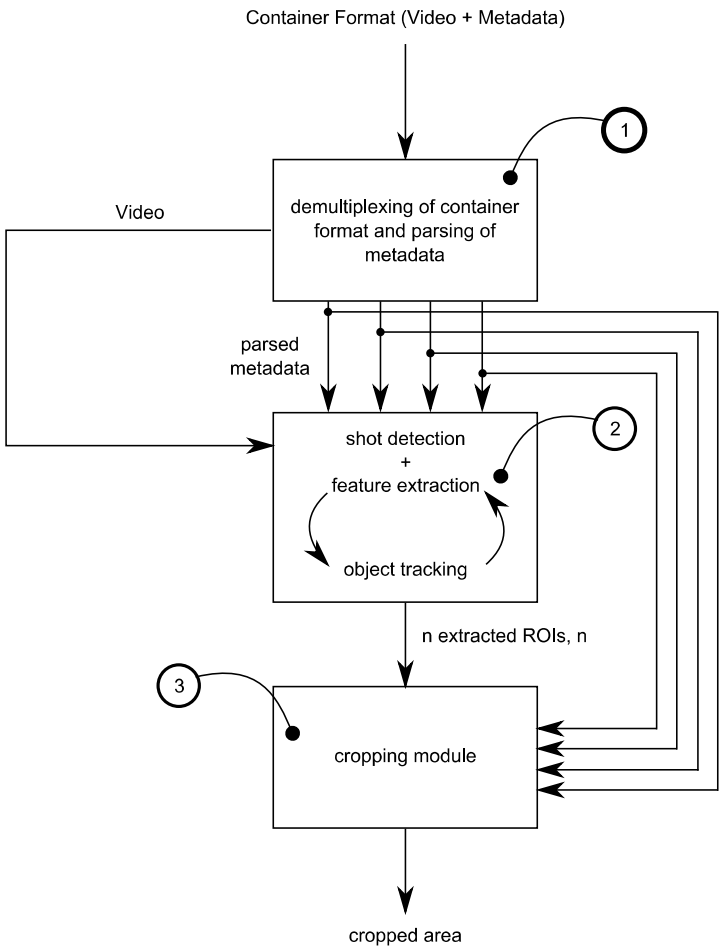

Figure 1: Overall architecture of the process

Box 1 (Demultiplexing and parsing of metadata, Figure 1.1):

To access each type of data, the interleaved data has to be separated by a demultiplexer. The extracted video is passed through to the video analysis (see Figure 1), while the metadata is parsed and important information is categorised in a useful structure. Metadata is a content related description using an easy file structure, e.g. XML (Extensible Markup Language). Here, it is roughly distinguished in descriptive data and technical data. Descriptive data is a content related description. This information can be either static or dynamic. Dynamic means data changing in time is synchronised to the video content, e.g. description of a person appearing in the video. Static data is a description which is valid for the entire video, e.g. type of genre. On the other hand, technical data is related to the format of the essence and can also be static or dynamic. It describes the format of the embedded video. Both, technical and descriptive data is provided to the Feature Extraction Module (Box 2).

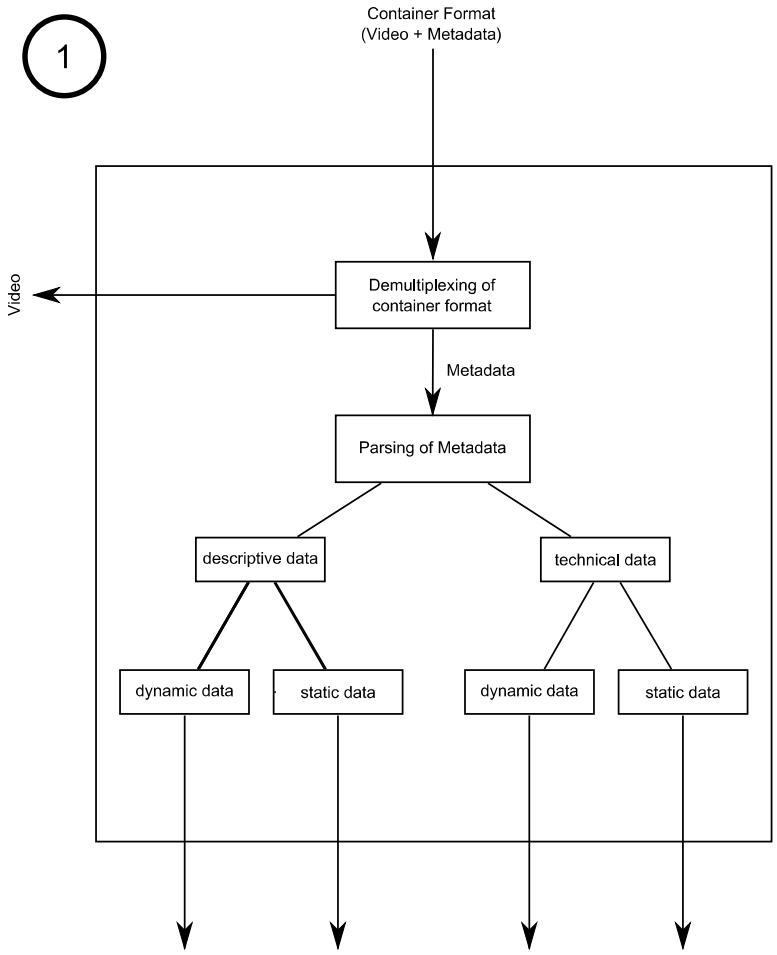

Figure 1.1: Demultiplexing of container format

Box 2 (Feature Extraction Module, Figure 1.2):

The video and metadata delivered by the previous module (Box 1) is combined to extract important features in a context. For this, the categorised metadata are used to initialise a dynamically fitted chain of feature extractions adapted to the delivered video content. Those can be motion detection (e.g. Block Matching), morphology filters (e.g. Erosion), edge detection (e.g. Sobel operator), etc. As additional feature extraction, a visual attention model is implemented and used. Such a visual attention system simulates the visual system of human beings. It detects salient low level features (bottom-up features), like main orientation, colours or intensity and combine them similar to the procedure of the human eye.

Each genre type has a different combination of feature extraction methods and different parameters, which are dynamically controllable by metadata or other information obtained by extracted features. This is depicted in Box 2 by a matrix allocating a genre type with specific feature extraction methods. Following, the detected features are weighted by importance, e.g. by their position or size. Relevant and related features are then combined to a ROI and delivered to the tracking tool. The tracking tool identifies the new position and deformation of each initialised ROI in consecutive frames and returns this information to the feature

extraction. By this, a permanent communication between feature extraction and tracking tool is guaranteed. This can be used to suppress areas for feature extraction which are already tracked. Finally, one or several ROIs are extracted and delivered frame by frame to the cropping module (Box 3).

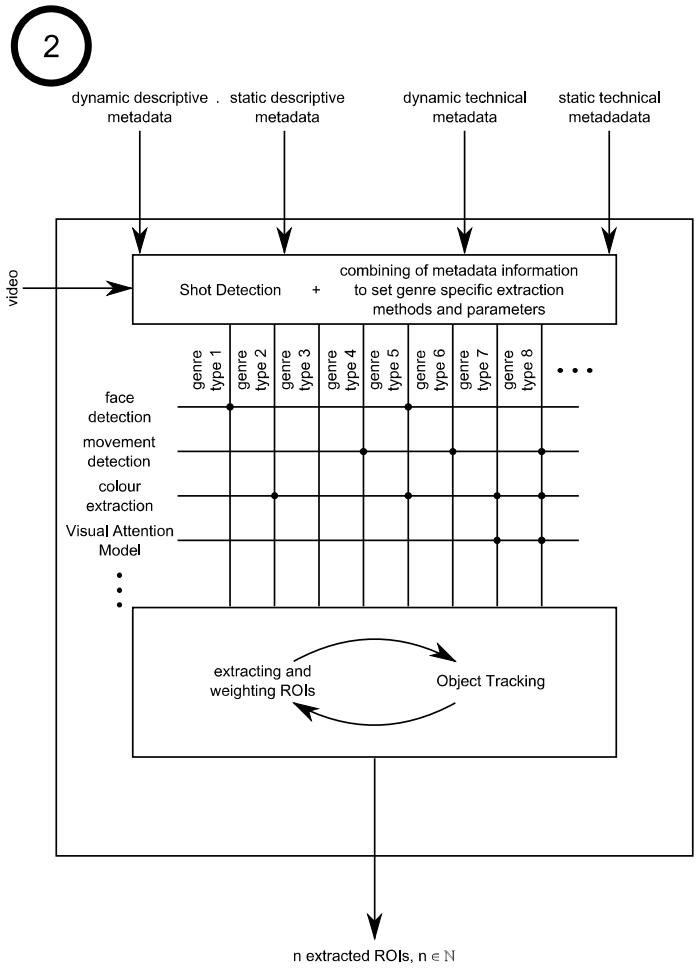

Figure 1.2: Feature Extraction Module

To explain the approach of feature extraction above more in detail, a short example treating a showjumping scene depicts a possible combination of different feature extraction methods. As already mentioned, the used methods are initialised and combined by available metadata. The most important metadata information is which type of genre is present. Here, that information is used to apply special video analysis methods to detect the position of the horse. Figure 2 roughly explains a possible process to get the position and size of the horse and rider. Basic prerequisite in this case is that showjumping is produced with static foreground (horse) and moving background. This leads to an approach to calculate the offset of moving background between two consecutive frames (depicted with f0 and f1 in Figure 2). Knowing the offset, the latter frame can be repositioned by it and subtracted from the previous one. The results are dark areas where background matches and bright areas where pixels differ from the

background. After applying some filters to gain the difference between dark and bright, clearly bring out a rough shape of the horse and rider (s. bottom of Figure 2). Once detected, it would be desirable to keep this ROI as long as it is visible in the following frames. For this, the tracking application is initialised receiving the initialised detected horse and matches it in consecutive frames. Updated tracking positions in subsequent frames are returned from the tracking module to the Feature Extraction Module.

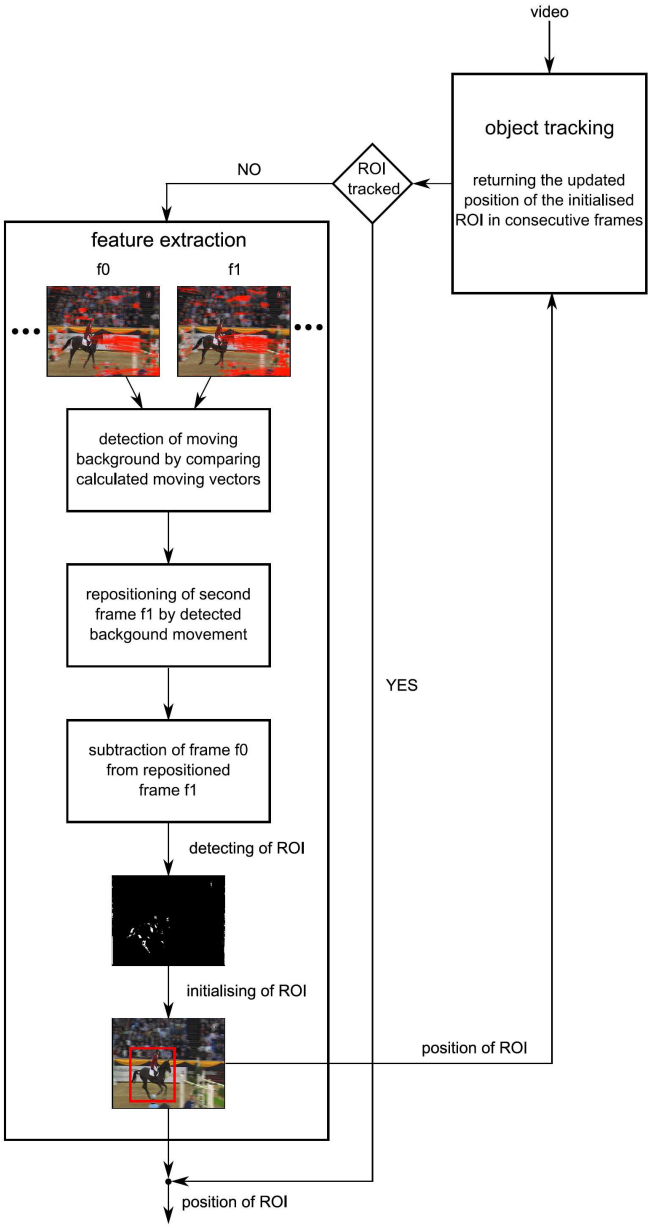

Figure 2: Example of initialised feature extraction methods to detect a ROI

Box 3 (Cropping Module):

The cropping module mainly has the function to crop a well composed image part. For this, all received ROIs are classified by importance and available metadata are used again to aid the decision of positioning the cropped area. Besides simply choosing an area for cropping, it has to be considered whether an anamorphic video is present (16:9 aspect ratio horizontally clinched to 4:3) and square or non-square pixels composes the image. Dependent of the image format of the target display, these possibilities are considered and adapted to avoid an image distortion. Additionally, viewing conditions for the different displays have to be considered. By this, a benchmark defines which size the cropped area should have compared to the original image. Such a benchmark can be determined by a comparison of viewing distances for both display resolution. Those considerations may change the size and shape of the cropped area again and has to be adapted once more. After coming to a decision of a properly cropped area considering all content-related and technical issues, the image has to be scaled to the size of the target display.

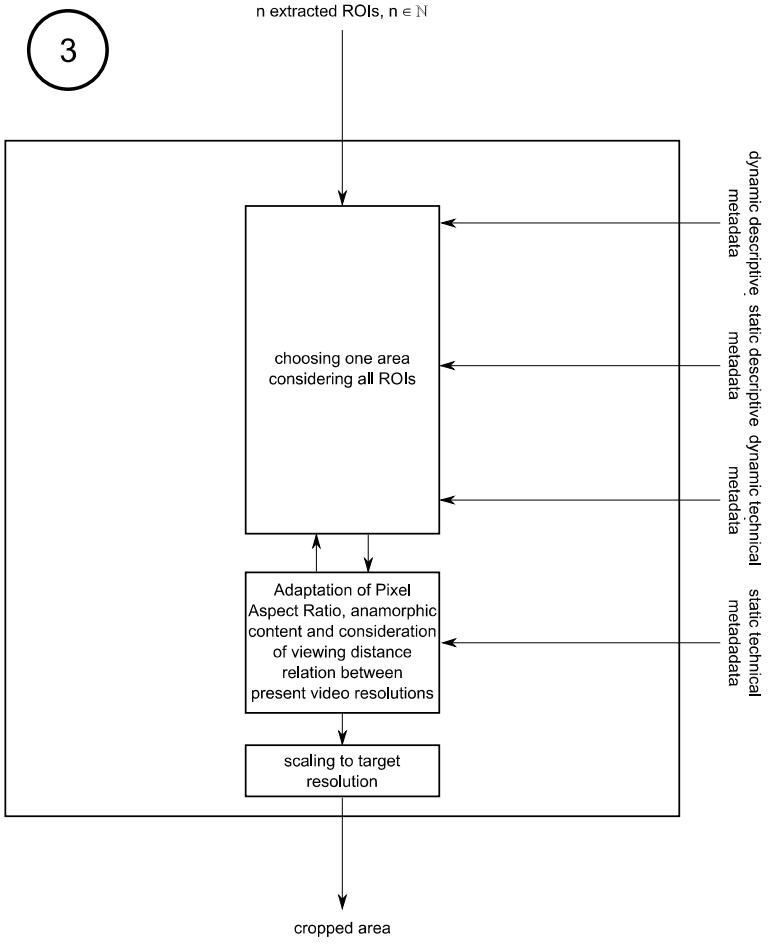

Figure 1.3: Copping Module

As shown above, the example of extracting features for showjumping is a specially-tailored method and would not work properly for other types of content, e.g. soccer. Therefore, the presented approach requires metadata to choose the right extraction method for the present type of genre. In the end, it is desirable to adapt video content like depicted in Figure 3.

Figure 3: Comparison of original and cropped image

The proposed methodology describes a workflow controlled by metadata. By this, a specially-tailored feature extraction and cropping method can be applied to increase the reliability of video analysis and aesthetic of the composed image.

The video analysis and cropping example of showjumping explained above is just for demonstration purposes of one possible workflow more in detail. They are not part of the patent application.