Contextual Cropping and Scaling of TV Productions

Joerg Deigmoeller, Itagaki Takebumi, Norbert Just, Gerhard Stoll

In this paper, an application is presented which automatically adapts SDTV (Standard Definition Television) sports productions to smaller displays through intelligent cropping and scaling. It crops regions of interest of sports productions based on a smart combination of production metadata and systematic video analysis methods. This approach allows a context-based composition of cropped images. It provides a differentiation between the original SD version of the production and the processed one adapted to the requirements for mobile TV. The system has been comprehensively evaluated by comparing the outcome of the proposed method with manually and statically cropped versions, as well as with non-cropped versions. Envisaged is the integration of the tool in post-production and live workflows.

Cropping and scaling, computer vision, regions of interest, visual attention, global motion estimation.

Introduction

Nowadays, broadcasters distribute their services over various channels. In addition to the traditional broadcast via antenna, satellite or cable, content is provided to the viewer via internet streams, podcasts or adapted broadcast systems for mobile devices, the latter of which is a growing and quite promising market. Especially in Korea (T-DMB) and Japan (One-Seg, based on ISDB-T) the mobile TV market is gaining a substantial market-share.

Key requirements for the success of mobile TV services are the adaptation of video content for optimal viewing conditions on mobile devices as well as an appropriate video quality. European mobile TV trials have shown that 24 % of users stopped using the service because of quality issues [1]. The study indicates that there is a high demand for made-for-mobile, bite-sized content.

Adaptation of content for mobile devices should be more than just a replication of traditional linear TV content. Mobile TV has to attract an audience with new programming and viewing experiences in order to co-exist with traditional TV on stationary receivers. Watching TV on portable devices should be complementary to the trend towards larger displays at home, such as 42” or even 50” flat screen displays. Unfortunately, all too often, identical TV content is presented on the various distribution channels as the generation of specific content for mobile TV

is very costly and time consuming for content providers. The creation of different video formats already needs to be implemented on program production level with direct implications on artistic design. Alternatively, content adaption can be performed manually during postproduction which is not feasible for live productions.

The proposed work addresses the problem of content adaptation, in particular for mobile TV applications, by means of contextual automatic cropping. The sub-region to be displayed on the mobile device is computed by using metadata information available from the broadcaster's production workflow in combination with video analysis methods. The metadata information feeds the adaptation system with a priori knowledge about the content and is used to guide the feature extraction algorithms. By doing so, the algorithms become aware of the content properties and therefore work more efficiently and deliver more reliable results. If no metadata is available, the system can still be applied on any type of content in a default mode. In this case, salient regions are detected without any background knowledge.

Related Work

The automatic and accurate detection of Regions of Interest (ROIs) is essential for a high-quality cropping method. In the research area of analyzing images for saliency, most work is based on the Feature-Integration Model by Treisman [2]. This model describes the processing of the human eye to recognize saliency. Treisman distinguishes between top-down and bottom-up features. Top-down features lead the search for salient objects by prior knowledge on context and/or object properties. In turn, the bottom-up approach relies on image features attracting the human eye, such as color, orientation, intensity and stereo distance. Itti et al. [3] implemented the bottom-up recognition of salient objects in still images based on the Feature-Integration Model.

For video content, motion becomes available as an important additional parameter to automatic cropping algorithms. To compute motion histograms, [4] applies a block matching algorithm. In addition, they combine the motion map with an attention model similar to Itti et al., thus adding color and intensity information. In a last step, the authors apply a median filter on extracted ROIs for temporal and spatial smoothing.

In [5], a motion map is determined by optical flow and an appearance map is obtained from color information. A saliency map is additionally computed by an alternative attention model called “Spectral Residual” [6].
A prototype solution (“Helios”) was presented by Snell & Wilcox [7]. The system is able to zoom in if clearly defined ROIs are available. Recent S&W publications explain the techniques planned for future professional conversion tools, using different approaches of foreground and background estimation [8].
In 2008, Thomson (now Grass Valley) developed the ViBE Mobile TV encoder which also relies on ROI detection for repurposing video content. The applied visual attention model [9] considers color contrast, visual masking effects and orientation features. Furthermore hierarchical block matching, a 2D affine motion model and M-estimator regression is used to determine temporal saliency.
Most of the presented approaches do not make use of prior knowledge about the content in order to analyze and choose important areas. Therefore, the rating of importance of ROIs can only be done independently of the context and is based on general assumptions.
However, content-specific methods exist as well. In [10], a football player tracking system is presented. Players are extracted by histogram backprojection of the pitch color. In the next step in the process, particle filters are used to track players.

Motivation

Currently, top-down information is rarely exploited for ROI extraction in broadcast applications. Specific applications based on such information work exclusively for a single type of content and sometimes even need a special set-up on site. On the other hand, a classifying or weighting of ROIs can hardly be applied based on bottom-up information only.
In this work, prior knowledge of the type of broadcast content is obtained by metadata information, which has recently become available in the production workflow through the transition to tapeless production.
The proposed system applies computer vision methods to different types of sports productions, and uses metadata to enhance the quality of these methods in extracting ROIs from the images. Sports have been chosen because it is very popular content to be viewed on mobile devices. Compared to movies, sports do

not need as much contextual information and can be watched simultaneously with other activities. Moreover, movies usually have a length of up to several hours, which also makes them less suitable for watching on small displays: a user trial from 2006 reports the average usage of mobile TV of no more than one or two sessions per day with an average duration of 23 minutes each [11].
This paper first describes the overall system followed by an in-depth discussion and justification of each feature extraction method. Afterwards, the processing of extracted ROIs on higher level and the definition of cropping areas are introduced. As proof-of-concept, a comprehensive evaluation is then presented that shows the reliability of the system. The paper concludes with a brief discussion of the overall approach.

System Overview

The system is based on the principle of a plug-in system. Each plug-in is loaded at run time and fulfils a certain task. A plug-in can contain several modules, where each of those define a smallest possible combination of associated operations with an as small as possible external interface.
A distinction is drawn between extraction plug-ins which are a collection of computer vision methods and the Classification Plug-In, respectively Cropping Plug-in. The two latter ones work on a higher level and process ROIs returned by the extraction plug-ins.
Within this work, two extraction plug-ins have been implemented. The first one is the Visual Attention Plug-In. It is composed of a motion map combined with still image saliency detection. The second extraction plug-in is the Backprojection Plug-In. It has been developed for specific types of content to detect objects that are moving on a plain background, for example a soccer game. By this, the Backprojection Plug-In is fed by valuable top-down information.
The Classification Plug-In allows a contextual weighting of extracted ROIs returned by the extraction plug-ins. The Cropping Plug-In represents the final plug-in in the complete processing chain. It not only defines final cropping areas in a video sequence, but filters ROIs that move consistently over time.
On the highest level works the plug-in system which is controlled by the system core (see Figure 1). It loads all required plug-ins and creates the structure of the application. Which plug-ins are loaded is determined by the incoming metadata.

Furthermore, the metadata information allows for loading predefined parameter settings for every plug-in, dependent on the type of sport. If no metadata is available, a default setting is loaded.
The metadata format used here is the Broadcast Metadata Exchange Format (BMF) [12]. It was developed at IRT (Institut fuer Rundfunktechnik) in collaboration with broadcasters and is tailor-made for content description in the scope of broadcast production. The system could however, be adapted to any other type of metadata that carries the required content information.

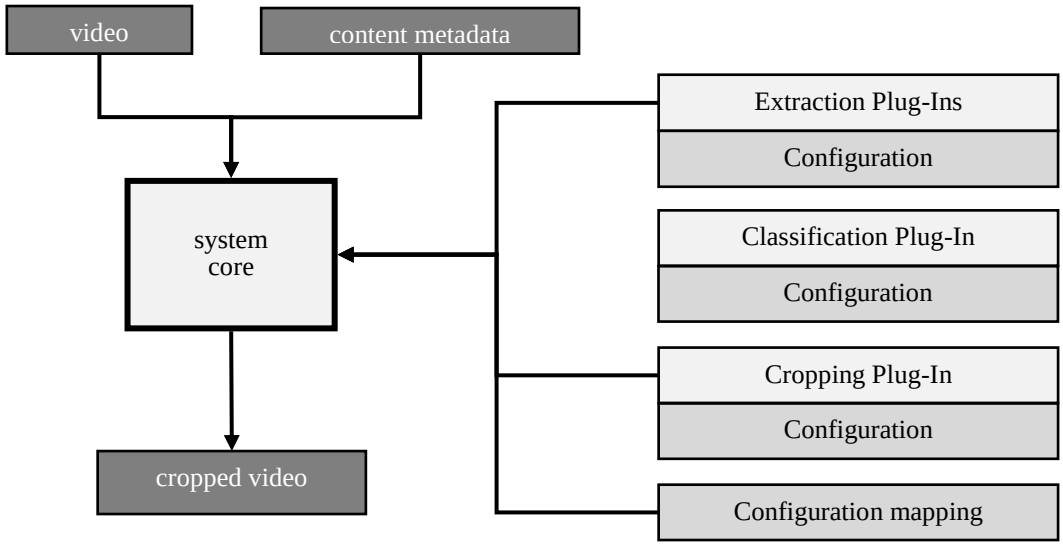

Figure 1. Overall structure of the system. Black boxes represent input and output data whereas gray boxes represent data used for system set-up and configuration. The system core manages incoming data as well as the configuration of the application by loading plug-ins.

Plug-in Configuration

Plug-in parameters can be set by loading their corresponding XML file which contains different parameter settings. Each parameter set is foreseen for certain use cases, i.e. different type of genres. Therefore, each module can be adapted individually to the video content. This allows to feed the system with content related background knowledge which actually is the advantage of this work compared to other approaches in the field of broadcast applications.
The internal processing of such information relies on a description of each video content by its properties. Those properties are not arbitrary but have to be predefined by the plug-in developer. In other words, the developer defines properties of videos that can be extracted by his plug-in. A property (e.g. motion)

has to be further specified by property values (e.g. fast motion). A property value can be assigned to multiple plug-in settings.

Knowing the type of sport from the BMF metadata, video content properties that are extractable by the plug-ins can be linked to each individual content type (see Figure 2). The linkage of extractable properties and present properties in a video signal is a clear and simple assignment. Currently, this only has to be done once, either by the user or by the developer. In future work, a simple graphical user interface could be considered, that allows the assignment.

As the proposed system is a proof-of-concept, the linkage is currently limited to a few type of sports for testing. As already mentioned, the system will run in a default mode, if no metadata is available (see Figure 2).

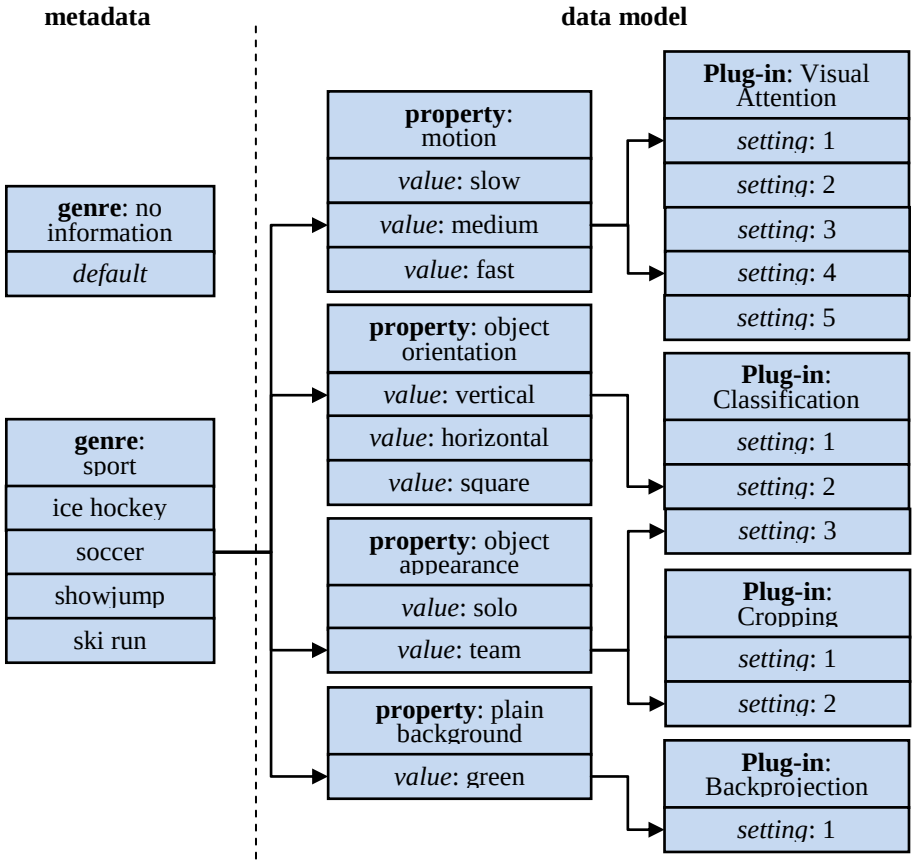

Figure 2. Assignment of properties of a specific sport production (left) and extractible properties (middle) for the example soccer. The listed plug-ins (right) are currently implemented. The system is, however, not limited to these plug-ins and can be extended at a later step.

Metadata Information

Metadata will be the electronic record report of future productions. As already mentioned, the metadata format used here is BMF, which is a comprehensive standard and considers the design and description of a complete production chain as well as the exchange of production components in a container format, e.g. MXF (Material eXchange Format).
Here, only two annotations of BMF are used, the first being the shot boundaries annotation. Shots are treated as tracks being part of a program. They can be linked seamlessly in case of hard cuts or they can overlap, e.g. in case of dissolves. So far, only shot boundary information annotated manually is processed by the system. Such information could be easily received as metadata in future by e.g. recording the editor action or identifying the cameras that are on air. This provides accurate shot boundary positions without any video processing by taking advantage of the production infrastructure. Additionally, detecting shot boundaries based on image processing would be beyond the scope of this work as the main focus is to detect ROIs within shots at first.
The second and most important feature of BMF considered here is the program type annotation. The BMF standard includes only a rough differentiation between genre types. It is up to each broadcaster to decide how those annotation types are to be specified in more detail. To do so, broadcasters have to create thesauri as enumerated data types as well as batched enumerated data types for sub-divisions. Those enumerated types have already been established for German public broadcasters (ARD, ZDF) in [12]. The genre type within this work has been annotated according to these thesauri. This information could be easily annotated during a production workflow either on site or in advance by the editor.

Plug-Ins

In the following sections, each plug-in that has been implemented is introduced. As already mentioned, plug-ins are no inherent part of the application. They are loaded at run-time. First, each extraction plug-in is discussed and finally an introduction to plug-ins working on higher level finishes this section.

Visual Attention Plug-In

The Visual Attention Plug-In consists of two sub-modules: the Still Image Saliency Module and the Motion Saliency Module. The Visual Attention Plug-In is a general approach to combine saliencies of still as well as moving pictures. Both modules analyze the video content independently and hence can run in parallel. Results of both modules are finally fused into one weighted saliency map.

Motion Saliency Module

Obviously, most sports productions contain both, camera motion and object motion. Individual sports are often shot by keeping the object of interest focused. In that case, the most interesting part of the image is kept at a rather static point and does not move much at all. Therefore, the objects of interest are those which move in a different manner than the camera does.

When the amount of camera motion (camera motion is expected here to be the global motion in an image) between two consecutive video frames is known, an inverse warping of one of both images by the computed transformation can be applied. As a result, two images shot at different times then get coincident backgrounds. The objects of interest were also warped, but are not coincident in both images. Taking both images and subtracting one from the other, the background is blanked and the foreground brightens up.

The challenge of this approach is to have robust and accurate camera movement detection, which can only be obtained by a proper motion vector field. For the proposed system, the gradient based motion estimation of Lucas & Kanade [13] has been selected. This method exclusively computes movement at clearly identifiable pixels dependent on the image structure which leads to a sparse vector field. Therefore, the method delivers less, but more reliable motion vectors compared to methods like block matching. The gradient based motion estimation used here, is the OpenCV C/C++ [14] pyramid implementation of Lucas & Kanade [15].

Having created a motion vector field, the challenge is to categorize the motion vectors, respectively to identify whether a vector corresponds to camera motion, such as pan, tilt, zoom, rotation or combinations thereof. Assuming that radial distortions as well as global motion between two frames are small, the

computation of 2D affine homographies is absolutely adequate. In case of homogenous coordinates the 2D affine homographies describe a mapping of points $x_{\mathrm{i}} = \left(x_{\mathrm{i,x}}\ x_{\mathrm{i,y}}\ 1\right)^T$ in frame t and matched points $x'_{\mathrm{i}} = \left(x'_{\mathrm{i,x}}\ x'_{\mathrm{i,y}}\ 1\right)^T$ in frame t+1 by:

$$x'_i \times H \cdot x_i = 0 \tag{1}$$

where

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

Geometrically, the equation above can be interpreted as the matrix H maps two position vectors x'_i and $H \cdot x_i$ in a way that they point in the same direction but with not necessarily the same magnitude and hence their cross product is zero. The parameters of H describe translation (corresponding to pan and tilt), rotation, shearing and scaling (corresponding to zoom).

Obviously, there is no exact solution for H, i.e. for the over-determined system, because the measurement of pixel coordinates is inexact. Additionally, motion vectors that correspond to object motion represent outliers to the camera motion and follow a different and unmodeled noise distribution than the imprecise coordinate measurement does.

The approach chosen here to remove outliers from the homography estimation is called Least Median of Squares (LMedS). At first, it randomly chooses three vectors to compute their transformation matrix based on Equation 1, where the deviation from zero provides information how well the homography fits. This deviation is called residual. After estimating a set of homographies, their residuals are sorted in ascending order and residuals above the median values are removed from the set as they are assumed to represent outliers. Finally, H is re-estimated by minimizing the least squares error of the inliers data set only. This over-determined system is solved with the aid of the Singular Value Decomposition [16].

Additionally, only a set of vectors that are spatially distributed are considered for estimating H, because vectors lying close to each other do not reveal much about global geometry. This is achieved by a bucketing technique, proposed in [17].

The regression method introduced above is implemented in the C/C++ library homest [18] by Manolis Lourakis. For further information, the author refers to [16] and [18].

Still Image Saliency Module

Complex saliency models like the ones by Itti et al. [3] or le Meur [9] are close to the perception of the human visual system. In some cases, however, they may be too general, because they rely on pure bottom-up information. Therefore, two attention models were evaluated with respect to the requirements of the system. One was the Matlab implementation of Itti's model available at [19] and the other one was the Spectral Residual approach by Xiaodi Hou, consisting of five lines Matlab code available at [20].

In sports productions, an attention model has to deal with dazzling colors, which may not be of contextual importance, e.g. advertisement banners or color markings. Comparing both attention models, Itti's model is more sensitive to dazzling colors because it uses color information. Spectral residual relies on gray images only and hence is less susceptible to this color information.

Additionally, some important information already exists in the composition of the image. Especially for individual sports, objects are mostly focused by the cameraman and in case of fast movements, the difference between foreground and background can be recognized due to motion blur. Whereas Spectral Residual strongly responds to these image properties, the attention model by Itti does not consider such depth information.

Based on the outcomes of the evaluations, the Spectral Residual approach was selected. It best meets the requirements of the system and in addition has the highest computational efficiency. It should be mentioned, that this is no general assessment, but rather one specific to the demands of this particular application. Ruderman stated in [21]: "…We can easily distinguish images of the natural world from man-made pictures or those created randomly by a computer. Natural images are distinctive, because they contain particular types of structure. They are far from random ...". In the Spectral Residual approach, such structure in images is defined as redundancy and is removed.

To estimate Spectral Residual, the absolute values of a Fourier transformed gray image $\mathsf{F}(I(x))$ are computed. Next, the natural logarithm of the absolute values is

calculated. To compute the redundancy of an image, a copy of it is approximated by a local average filter. Subtracting this copy from the non-approximated image, results in the Spectral Residual:

$$R(f) = \ln(|\mathcal{F}(I(x))|) - \ln(|\mathcal{F}(I(x))|) * h_n(f)$$

where $h_n(f)$ is a local average filter, e.g. 3x3.

Finally, the saliency map $S(x)$ of the gray image $I(x)$ is obtained by:

$$S(x) = g(x) * \mathcal{F}^{-1}[\exp(R(f) + i \cdot P(f))]^2$$

where $g(x)$ is a Gaussian filter and $\mathcal{F}^{-1}$ depicts the Inverse Fourier Transform. Squaring each pixel and subsequently normalizing $S(x)$ again to a range of 0 to 255 applies a gamma correction, which removes noise by spreading small intensities and clinching higher intensities. $P(f)$ describes the phase spectrum of the transformed image and is denoted by:

$$P(f) = \arg(\mathcal{F}(I(x)))$$

Map Fusion

Both maps – motion and saliency map – present a good complement. The accuracy of the saliency map increases for faster camera motion due to increasing motion blur, whereas camera motion detection may worsen because of less clearly identifiable points.

The saliency and motion maps are weighted and combined into one final map. The weight for each map is mainly determined by the reliability of the estimated camera motion, where the reliability should be high if the following conditions are satisfied: the number of clearly identifiable pixels must be sufficiently high to assure statistical significance and the error between computed affine homography and measured motion vector field should be small.

To evaluate the error of the fitted affine model parameters, the root median square error (*RMedSE*) is computed as follows:

$$\mathrm{RMedSE} = \sqrt{Med\left[\frac{1}{2} \cdot \left(d(x'_i, Hx_i)^2 + d(x_i, H^{-1}x'_i)^2\right)\right]}$$

where the notation $d(x,y)$ describes the Euclidean distance between x and y and Med is the median value of the term computed in square brackets for the entire data set. This error, also known as transfer error, defines the Euclidean distance between projected points from the first image and matched points from the second image by the affine matrix H and vice versa.

To compute the global weighting factor w, $RMedSE$ and the number of features are combined in a way that w only acquires a high value if the number of features is high compared to the image size and $RMedSE$ is close to an accuracy of one pixel. On the other hand, if w decreases, either the optical flow computation delivers sparse motion information because of missing structure in the image or too fast camera motion, or no clearly identifiable global motion exists.

This weighting factor is used to manage the influences of each map on the final saliency map Q:

$$Q = \frac{1}{2} \cdot w \cdot M + \frac{1}{2}(1-w) \cdot S$$

where M is the motion map and S is the saliency map.

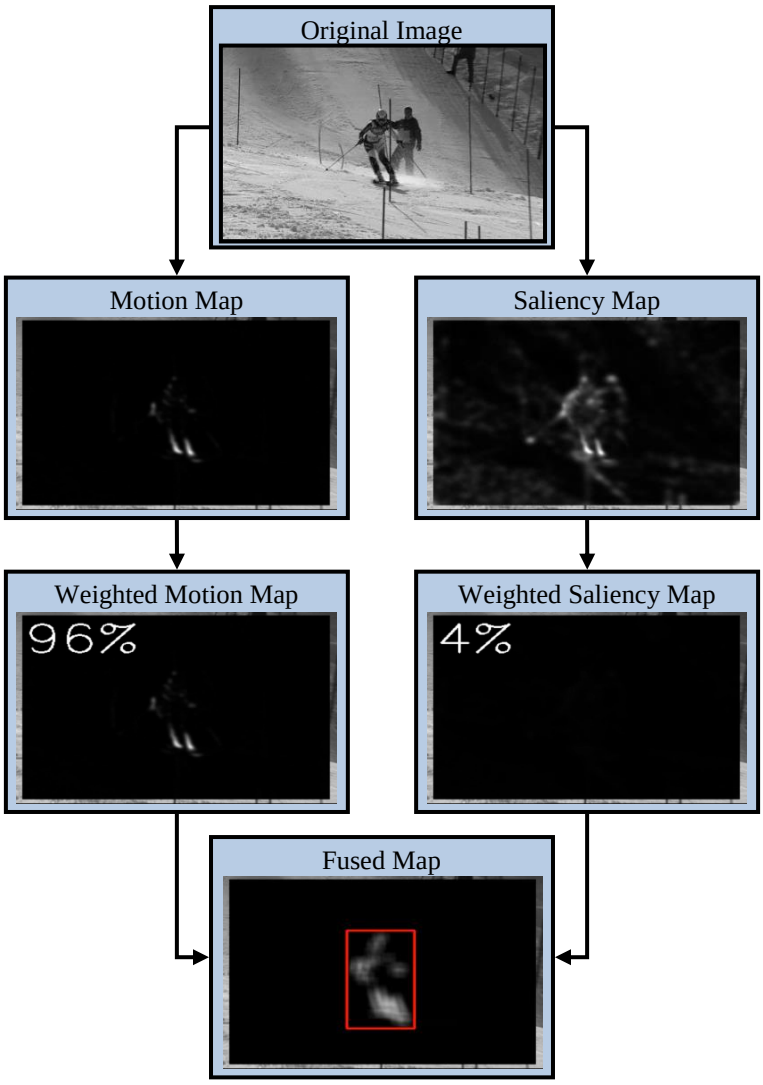

Figure 3. Fusion of weighted motion map (left) and weighted saliency map (right). After fusing the maps, ROIs are segmented by local binarization, filtering and edge linking.

Finally, an edge detection and edge linking is applied on the binarised image to define the Regions of Interest (see Figure 3). For local binarization, edge detection and edge linking methods that are available from OpenCV are used.

Backprojection Plug-In

The Backprojection Plug-In facilitates the detection of plain backgrounds in video images as well as objects which are located within this area. The Plug-In is intended to be optionally applied on sports that take place on more or less plain pitches. Using this component, important top-down information is considered which allows a more accurate extraction of possible objects of interest than simply applying the Visual Attention Plug-In.

The method of histogram backprojection builds the basis for the plug-in. It has originally been proposed by Swain and Ballard in [22]. In a first step, the histogram N of a sample image which contains the desired color pattern is computed. Here, the pattern to be loaded from a data set depends on the incoming type of sport information. Afterwards, the histogram I of the image to be analyzed is determined. For each bin j of both histograms N and I, their ratio is computed which results in a third histogram R:

$$R_j = \frac{N_j}{I_j}$$

Finally, the histogram backprojection is estimated by mapping each three-dimensional color value $c(x,y)$ to a bin by the histogram function $h(c(x,y))$:

$$b_{x,y} = \min\left(R_{h(c(x,y))}, 1\right)$$

where $b(x,y)$ is the backprojected pixel at image position (x,y). For the backprojection computation, the corresponding OpenCV function has been applied, which returns a binary image, where white represents colors that match the histogram bin of the sample image and black represents no matches.
To detect the pitch position and its shape, the backprojected image is scaled down to an eight of its size to remove image details. To further support large areas and suppress details in the image, a median filter with a kernel size of a quarter of the image width is applied. The resulting binary image now represents a rough pitch template. This is subtracted from the backprojected image which removes areas beyond the pitch and highlights elements which are placed within the template. As a result, possible players positioned on the pitch remain. This process is depicted in Figure 4.
Obviously, a drawback is that players which are off the pitch are not or just partly detected. It is assumed, that such effects can be compensated by ROIs estimated from the Visual Attention Plug-In.
To finally separate foreground from background, objects are segmented by another instance of the segmentation method used for the Visual Attention Plug-In (cf. Figure 3).

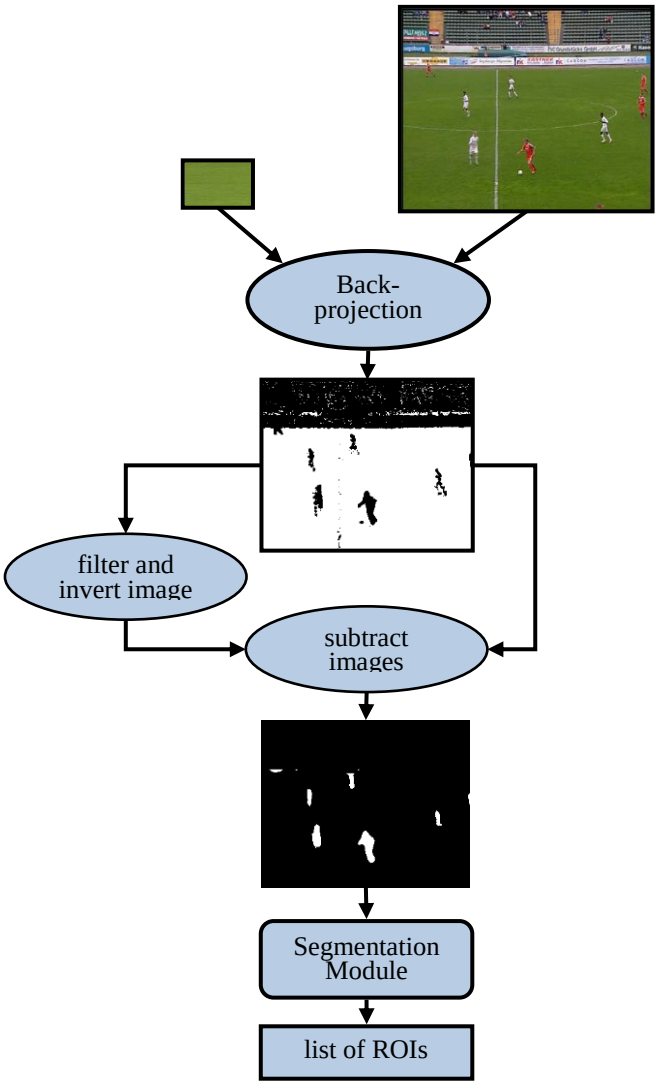

Figure 4. Flow chart of the Backprojection Plug-In. In a final step, objects are segmented by another instance of the Segmentation Module already used for the Visual Attention Plug-In.

Classification Plug-In

Due to the fact that different extraction algorithms can be used in parallel, a huge amount of more or less reliable ROIs are detected. Thus some rating has to be applied to reduce wrong detections by making use of known content properties. It has to be pointed out, that it is not needed to make a classification of different types of sports here, because this information has already been received from the metadata. The intention of this classification is to decide whether a ROI is of interest for the present type of content or not. Therefore, it is not aimed to find decision boundaries between multiple classes, but rather how well a ROI fits into a single class, e.g. the class describing players of a soccer game. Each class is formed by three features: shape, size and position. The parameters of these features are currently not learned by the system. They are set manually, because

the number of features and types of sports are limited to a few as proof-of-concept of the proposed approach. How such parameters could be learned, e.g. learning mean and variance for a certain feature, is beyond the scope of this paper, but could be considered in future work.

The features shape and position are expressed as probabilities. The weight of a ROI for the feature size is estimated by a threshold function. To express the probabilities as weights as well, they are normalized by a normalization operator $N(.)$ to a range from 0 to 1, where the maximum value 1 corresponds to the mean value of the probability function. The total weight of a ROI is finally obtained by multiplying all individual weights:

$$w_{ROI} = N(P_{position}) \cdot N(P_{ar}) \cdot w_{size}$$

The feature position is calculated by using a two dimensional Gaussian distribution. The mean value is located at the center of the image, respectively can be shifted if required. By changing the standard deviation it is possible to define the range of the important area. For a single object, a small value might be sufficient as a single most important element is mostly located by the cameraman around the center position. In turn, a large value might be more appropriate for multiple objects.

The feature shape basically represents orientation. Here, the orientation represents the aspect ratio of a ROI. The weighting is applied by a function which describes the desired aspect ratio defined by the class. Aspect ratios a_r are converted to a fixed range from 0 to 2 as follows:

$$a_r = \begin{cases} \frac{height}{width} & for\ width \geq height \\ 1 + \frac{height}{width} & for\ height < width \end{cases}$$

In order to compute the probability of a ROI's aspect ratio P_{ar}, a normal distribution is defined by mean value and standard deviation, where both parameters can be adapted to the analyzed genre.

Finally the size feature is calculated as follows: The ratio between image size and ROI size is computed. A simple threshold is used to decide whether the ROI is below a desired size which results in a Heaviside step function.

After all feature ratings are computed, the final weight of the ROI is calculated as mentioned above. For any further processing of the ROI, the weight value is used

to evaluate the contextual importance of a ROI. As a last step, it is possible to reject ROIs by defining a minimal rating threshold that must be achieved.

Cropping Plug-In

The Cropping Plug-In represents the final plug-in in the complete processing chain. It defines not only the final cropping areas, but filters ROIs that move consistently over time. The filtering is done with the aid of a further sub-module, the Cluster Module. It groups corresponding ROIs across several frames by means of equal time slots, simply called windows in the following. The clustering over time is referred here as *Inter-Frame Clustering* (see Figure 5). In case that several extraction plug-ins run in parallel, the clustering module can be applied as well for grouping ROIs which are returned by multiple components for a single video frame. This additional method is called *Intra-Frame Clustering*.

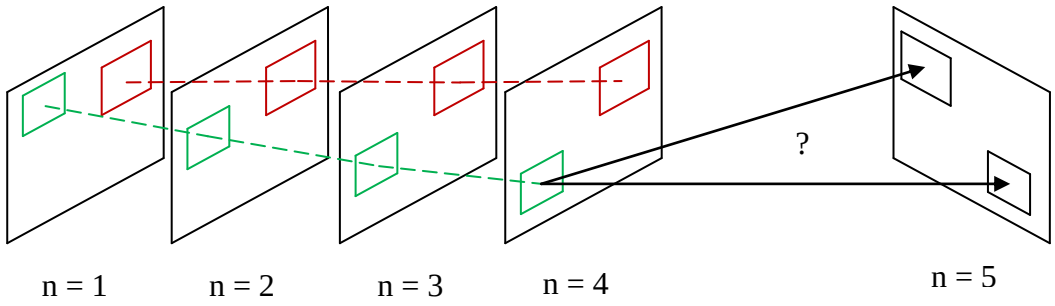

Figure 5. Inter-Frame Clustering applied on a window of 5 consecutive video frames. The ROIs in frame n = 1 serve as initial cluster.

Clustering Module

For Inter-Frame and Intra-Frame Clustering, the same similarity measure is used by means of an agglomerative clustering process. It starts with a single ROI for each cluster and adds further ROIs dependent on relative ROI distance and relative ROI size ratio. Expressing both conditions as probability, the total probability results in:

$$P_{total} = P_{distance} \cdot P_{shape}$$

Whether a ROI is assigned to a cluster or not is determined by a threshold that defines a lowest probability allowed for clustering.
The relative distance between two ROIs is not simply their center distance but their border to border distance. This is motivated by the fact that ROIs which

overlap can have a quite high center-to-center distance whereas their actual distance is zero. Additionally, the ROI size is set into relation with the measured distance which results in the probability for relative distance:

$$P_{distance} \sim \frac{w_1 \cdot w_2}{d^2} \cdot \frac{h_1 \cdot h_2}{d^2}$$

where w_1, w_2, h_1, h_2 are the width, respectively height of two ROIs and d is the border-to-border distance between the ROIs. Values of the similarity measure greater than 1 are set to 1, which reflects the case that ROIs overlap and hence $P_{distance}$ is maximal.

The probability P_{shape} is estimated by means of the relative size ratio, which is measured by the shape similarity of ROIs:

$$P_{shape} \sim \left(\frac{MIN(w_1, w_2)}{MAX(w_1, w_2)} \cdot \frac{MIN(h_1, h_2)}{MAX(h_1, h_2)} \right)$$

where MIN and MAX return the minimum, respectively maximum of two rectangles width and height.

For Intra-Frame Clustering, the ROI with highest probability is kept and others are removed from the cluster.

For Inter-Frame Clustering, ROIs extracted for multiple video frames are buffered at first. Once a time window has been completed and clusters have been created, reliable ROIs are filtered and gaps within trajectories are closed by means of linear interpolation. This requires a sufficient number of ROIs per cluster in order to evaluate their evidence. This can be ensured by a minimal required number of ROIs per cluster and a maximal gap between two consecutive ROIs in a cluster. Clusters with a size below a minimal required size are removed.

Defining the final cropping area

So far, ROIs have been extracted, weighted and filtered over time. The approach to define the size of a cropping area is that the user chooses a desired zooming factor dependent on the source and target image resolution. This is motivated by the fact that a cutter working at an editing desk usually crops broadcast material by a scanning mask of fixed size as well. Once, he decided for a zooming factor, he positions the cropping mask to the optimal position. Additionally, this

approach avoids annoying effects of mixing camera motion – which is already part of the video – and dynamic zooming by the application.

The centre position of the cropping area is determined by searching for high weighted ROIs that fit into the cropping mask. How many ROIs are considered for the search can be influenced by the type of genre information. For example, if it is known that a certain type of sport contains multiple objects of interest, a combination of high weighted ROIs is considered. In turn, for individual sport, only the highest weighted ROI might be of interest.

To find the best combination of ROIs, the search starts with computing all possibilities for a given number of ROIs *m* that form a combination. Afterwards, all combinations are ranked in ascending order according to their average value of ROI weights. In a next step, the same proceeding is done for combinations formed by *m-1* ROIs and so on (see Figure 6). Finally, the last elements in the series of combinations are all single ROIs ranked for their weight value.

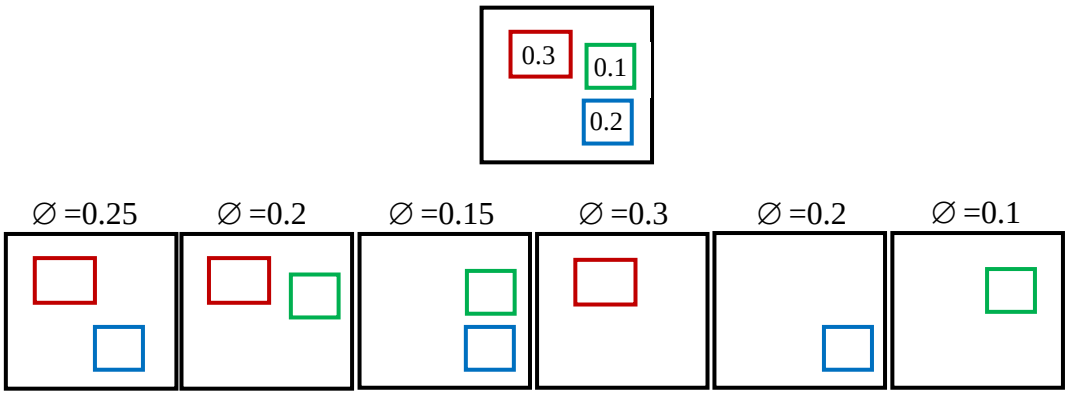

Figure 6: Example for ranked ROI combinations starting with two ROIs per combination. The top box shows the ROIs that have been extracted from the image with their weight values. The bottom row depicts all possible combinations of ROIs with corresponding average weight. The order of the ranked combinations start from the left with the highest average weight for combinations of two ROIs, followed by single ROIs ranked by their weight value.

Starting with the combinations formed by *m* ROIs, the centre position of the first ROI combination that fits into the cropping mask determines the position of the cropping area. In worst cases, either no ROI is found at all or no combinations fit into the scanning mask. In the first case, the image centre position is chosen. In the second case, the highest weighted single ROI is chosen, although it exceeds the cropping mask size.

Even if the ROIs have been filtered in the Cluster Module, scanning masks might rapidly change their centre position from one frame to the next. Therefore, just the median value of the x- and y-position of scanning masks within a time window is kept. By this, it is assumed that the most representative 2D position of a scanning mask does not significantly change within one time window. Rejected scanning mask positions are linearly interpolated with the aid of neighbored windows. This assures a smooth transition of 2D positions at window boundaries.
In Figure 7, two defined cropping sizes applied on an ice hockey example by the Cropping Plug-In are depicted.

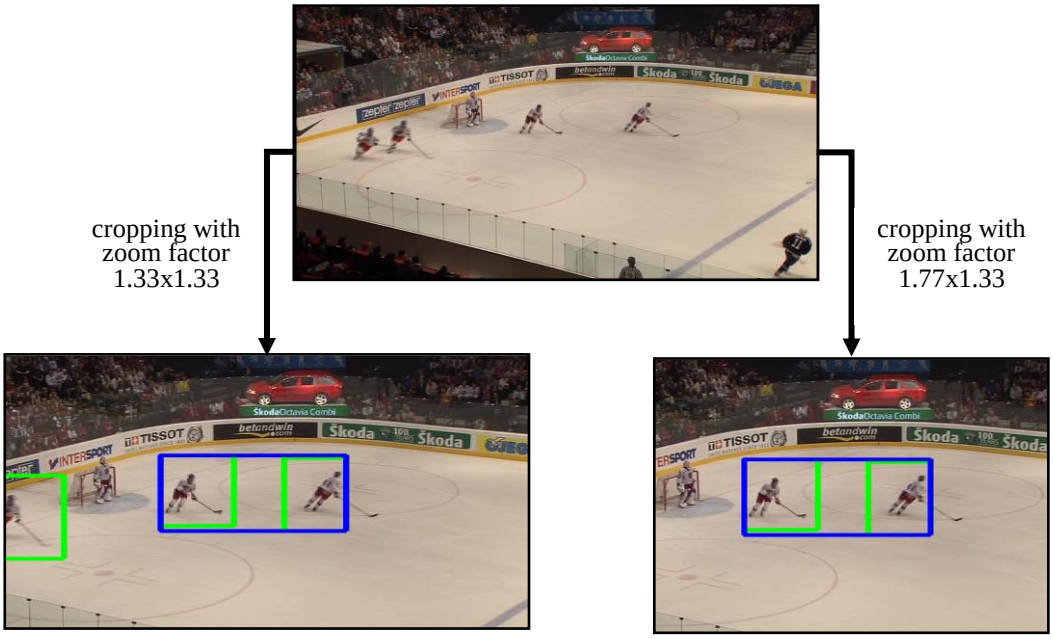

Figure 7. Example for two different cropping sizes. Green rectangles are ROIs that are consistent over time. The blue rectangle describes the best combination of high weighted ROIs that fit into the predefined cropping size.

System evaluation

Within this work, two types of system evaluation have been carried out. One was a subjective evaluation comparing different cropping methods. The other one measured gaze position of subjects watching sports videos and compared those to ROI positions extracted by the system. The combination of both types of evaluation provides a comprehensive statement of the system reliability. Results of the gaze tracking have already been reported in [23]. There, scatter plots have

shown that the system almost always points at the region which could be identified by a viewer watching sports content. Within this paper, the authors concentrate on results of the subjective evaluation.

Subjective evaluation

Fifteen subjects participated in the subjective evaluation (8 experts and 7 non-experts). Non-experts were considered as people who are not directly concerned with picture quality as part of their daily work. The purpose of this test was to compare the output of the introduced application to results from manually cropped, statically cropped (simply cropping the centre area) and non-cropped (simply scaled or scaled with letterbox) videos (cf. Figure 8). The sport sequences which have been used included SDTV sequences (720x576, 25 fps) from ice hockey and soccer matches (team sports) as well as excerpts from show jumping and ski race (individual sports) with a length of approximately 15s each. The material that has been used for this evaluation was exclusively clean feed material from the archive of a German public broadcaster. Other material, for example broadcasted material, is inacceptable as it can contain graphics which can be truncated by cropping. Additionally, the bitrate might be much too low for further processing which can cause heavy artefacts.

The manually cropped content has been prepared by a professional cutter from German public broadcasters. The different cropping levels were 1.33x1.0, 1.33x1.33 and 1.77x1.33 (cf. Figure 8). The cropped material was finally scaled down to common target resolutions according to the DVB-H standard [24] with 320x240 for 4 by 3 content and 400x224 for 16 by 9 content. The viewing distance of the subjects was fixed to 12 height units, which is a good compromise between common viewing distances and theoretically optimal conditions for mobile devices. The video sequences have been presented on PC monitors.

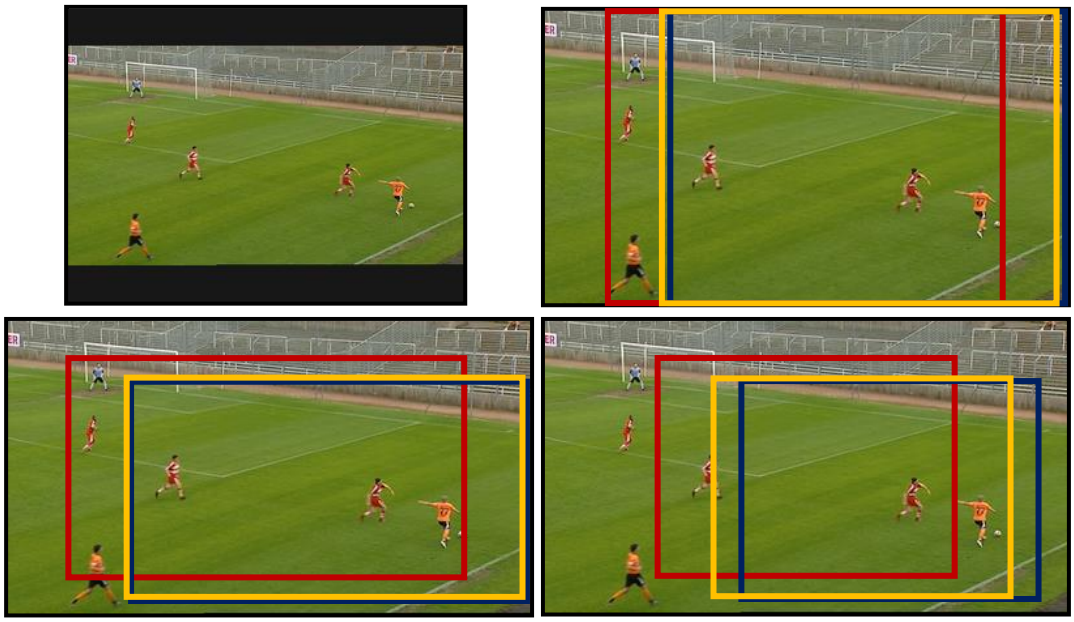

Figure 8: Example for different types of cropping used for the evaluation. The upper left image shows the non-cropped version with letterbox. The other images illustrate the positions of the cropping areas for the statically, i.e. centered cropped version (red bounding box), manually cropped version (blue bounding box) and automatically cropped version (yellow bounding box).The upper right image show results for a cropping level of 1.33x1.0. The bottom left image depicts positions of cropping areas for a cropping level of 1.33x1.33. The bottom right image show results for a cropping level of 1.33x1.33.

The evaluation method that was chosen is SAMVIQ which is specified in [25]. Other methods specified in [25] base on double-stimulus or single-stimulus evaluations. The former allows the viewer to assess one test version in comparison to a reference version of a sequence. The latter defines that the viewer has to grade a single video without reference. SAMVIQ differs from these approaches as it allows the viewer directly to compare more than one version of a sequence to a defined reference version (multi-stimulus). Its advantage over double- or single stimulus methods is that a subject can directly compare all processed versions of a sequence to a reference as often as he likes. Besides this, he is able to loop the complete sequence or even parts of the sequence. Therefore SAMVIQ offers a high flexibility as the assessor is not forced to make a decision within a defined period. This is important for the evaluation, because it was asked for the subjective impression of the cropped videos. Compared to e.g. video codec evaluations, there exists no right or wrong in this evaluation.
Obviously, none of the cropped or non-cropped videos represent a clear reference. Therefore, the reference version has to be well-considered, because the question

of the evaluation significantly differs by changing it. As the main intention of this evaluation is to assess the subjective impression of the position of the cropping area, the statically cropped version was chosen as reference. Results estimated by this set-up directly indicate the necessity of adapting the cropping area intelligently instead of simply cropping the centre position. The statically cropped version does not present a reference of highest quality. Therefore, the recommended quality scale of SAMVIQ (continuous scale from "bad" to "excellent") has been replaced by a comparison scale (from "much worse" to "much better") according to [25].

Subjects were asked to check attributes in addition to their assessment which should reflect the intention of a subjects rating. To do so, a subject had to specify one attribute which had mainly influenced his decision, where the available attributes were: *motion*, *sharpness*, *proportions* and *position of the cropping area*. In case that the subject was not able to justify his decision or the corresponding attribute was not listed, he had the additional possibility to choose *do not know*.

The scale used in this evaluation was a measure for qualitative variables and hence is a non-metric scale. From statistical literature [26], it is highly recommended to use non-parametric statistics in such a case. Due to this fact, median and quartile had been used as statistical method instead of mean and standard deviation. Figure 6 shows the overall results of the subjective evaluation. In the lower part, the frequency distribution of attributes for each assessment is depicted. For illustrative purposes, it is distinguished between attributes according to positive (*better*) and negative (*worse*) ratings by splitting the frequency distribution as well.

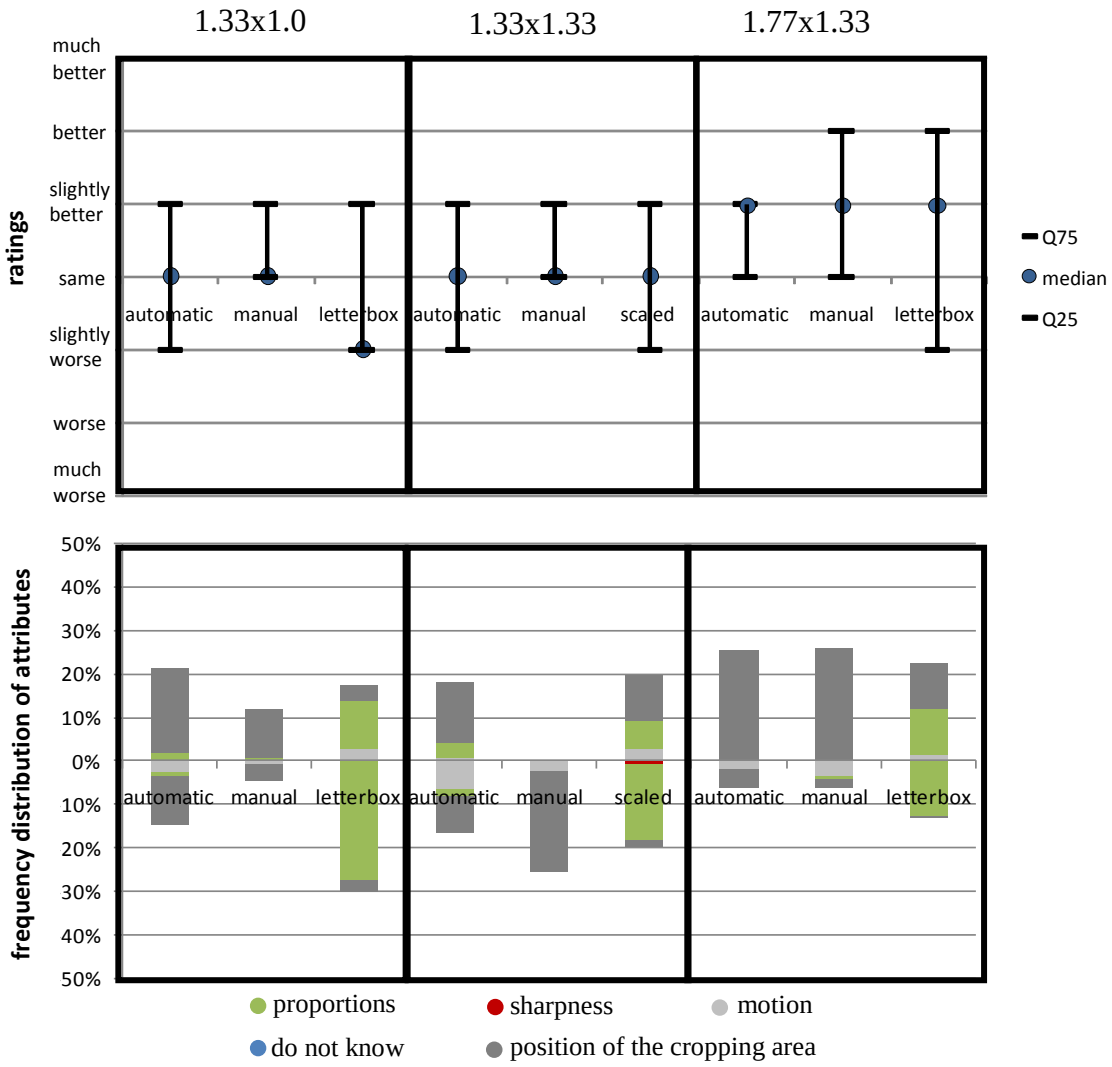

Figure 9. Results from the subjective evaluation over all sports sequences. The original materials were 16:9 SDTV videos cropped and scaled down to 320x240 (4:3 aspect ratio) and 400x224 (16:9 aspect ratio).

From Figure 9, it can be seen that for cropping level 1.33x1.0, the non-cropped version (letterbox version) tends to be slightly worse than the statically cropped version, which is mainly due to the attribute *proportions*. For the highest cropping level, the letterbox version is slightly better than the statically cropped version, which is mainly due to the *position of the cropping area* and *proportions*. This indicates that the statically cropped area seem to no longer enclose the most relevant areas of the sequences.

In turn, the automatically and manually cropped versions show very similar trends for all cropping levels. For the highest cropping level, both have been clearly preferred to the statically cropped version. This shows that the subjects were more

satisfied with the adapted cropping versions because of the *position of the cropping area*. This emphasizes the previous assumption that with higher cropping levels a statically cropped version is no longer sufficient most of the time.
It has to be mentioned, that this evaluation does not give an answer to the question whether there is a demand for cropping or not. As the subjects were asked to compare in relation to the statically cropped version, no answer is directly given to the question whether cropping or no cropping has been preferred to the scaled, respectively letterbox version. Even if it would be interesting to answer this question, including this statement would be beyond the scope of this work. It can just be stated that all versions have been favored in relation to the statically cropped version due to the *position of the cropping area*, respectively *proportions*.

Conclusion

The presented system follows a new approach to combine top-down information in form of production metadata with computer vision methods for broadcast applications. Such a fusion allows an optimized extraction of regions of interest for specific types of content. These ROIs are filtered by several processes on different levels to estimate a reliable number of contextual important regions. Based on computed weights for each ROI, the system is able to finally define a cropping area that encloses as much important image information as possible.
Results of the subjective evaluation have shown that with higher copping levels statically cropped versions are less satisfying than those with adapted cropping masks (manually and automatically cropped versions). In rare cases, the automatically cropped version was worse than the manually cropped version.
In addition to the subjective evaluation, a gaze tracking analysis has been carried out in [23] to compare a subject's line of vision with extracted ROIs of the proposed system. Results of this test demonstrate the reliability of the system independent of cropping. Scatter plots show that the system almost always points at the region which was also identified by a viewer.
In [23], it has also been shown that the system is able to run nearly in real time on a standard PC analyzing SDTV content. As the intention was not to implement a

real time system, but rather a proof-of-concept prototype, the authors see great potential for a real time application.

With HDTV penetrating the market quickly, a wider range of display resolutions needs to be considered for broadcast productions. This means that the required cropping ratio increases, which also has effect on the amount of work for a cutter compared to SDTV productions. This indicates the necessity of a system which identifies possible regions of interest automatically. In turn, the system does not have the complete contextual knowledge and hence the selection of the specific cropping area should still be in the hands of the cutter. Therefore, the proposed solution tackling the problem of automatic cropping and scaling should be seen as a supporting tool for cutters, for example by suggesting possible cropping areas. Such a solution can provide an improvement of the production work flow.

References

[1] C. Arthur (2007) Television is a turnoff for mobile users. The Guardian. http://www.guardian.co.uk/technology/2007/aug/02/guardianweeklytechnologysection.mobilephones. Accessed 26 June 2010

[2] A. Treisman (1986) Features and Objects in Visual Processing. Scientific American, vol. 255, pp. 106 – 115

[3] L. Itti, C. Koch, E. Niebur (1998) A Model of Saliency-based Visual Attention for Rapid Scene Analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20, pp. 1254–1259

[4] W.-H. Cheng, W.-T. Chu, and J.-L. Wu (2005) A Visual Attention Based Region-of-interest Determination Framework for Video Sequences. IEICE Transactions on Information and Systems Journal, vol. E-88D, no. 7:1578 – 1586

[5] T. Deselaers, P. Dreuw, H. Ney (2998) Pan, Zoom, Scan – Time-coherent, Trained Automatic Video Cropping. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage

[6] X. Hou. L. Zhang (2007) Saliency Detection: A Spectral Residual Approach. Conference on Computer Vision and Pattern Recognition, Minneapolis

[7] J. Zaller (2007) Snell & Wilcox's Helios. http://broadcastengineering.com/RF/broadcasting_snell_wilcoxs_helios/index.html. Accessed 26 June 2010

[8] M. Knee, R. Piroddi (2008) Aspect Processing: The Shape of Things to Come. International Broadcast Conference 2008, Amsterdam

[9] O. Le Meur, P. Le Callet, D. Barba (2007) Predicting visual fixations on video based on low-level visual features. Vision Research, vol. 47, pp. 2483-2498

[10] A. Dearden, Y. Demiris, O. Grau (2006) Tracking football player movement from a single moving camera using particle filters. Proceedings of the 3rd European Conference on Visual Media Production (CVMP), London. pp.29-37

[11] S. Mason (2006) Mobile TV – results from the DVB-H trial in Oxford. EBU Technical Review. http://www.ebu.ch/en/technical/trev/trev_306-mason.pdf. Accessed 26 June 2010

[12] BMF Documentation (2007) BMF – Broadcast Metadata exchange Format. Institut fuer Rundfunktechnik, Version 01.00.00, Munich

[13] B. D. Lucas, T. Kanade (1981). An iterative image registration technique with an application to stereo vision. Proceedings of Imaging understanding workshop, pp. 121--130

[14] (2009) OpenCV library Documentation. http://opencv.willowgarage.com/wiki/. Accessed 26 June 2010

[15] J.Y. Bouguet (1999) Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm. Intel Corporation Microprocessor Research Labs

[16] R. Hartley, A. Zisserman (2003) Multiple View Geometry in Computer Vision – Second Edition. Cambridge University Press

[17] Z. Zhang, R. Deriche, O. Faugeras, Q.T. Luong (1995) A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry. Artificial Intelligence, vol. 78, pp. 87-119

[18] M. Lourakis (2009). homest: A C/C++ Library for Robust, Non-linear Homography Estimation, http://www.ics.forth.gr/~lourakis/homest/. Accessed 26 June 2010

[19] D. B. Walther (2010). Saliency Toolbox. http://www.saliencytoolbox.net/index.html. Accessed 26 June 2010

[20] X. Hou (2009) Spectral Residual, http://www.its.caltech.edu/~xhou/. Accessed 26 June 2010

[21] D.L. Ruderman (1994) The statistics of natural images. Computation in Neural Systems, vol. 5, pp. 517-548, 1994

[22] D. A. Forsyth, J. Ponce (2003) Computer Vision - A Modern Approach. Prentice Hall. New Jersey

[23] J. Deigmoeller, N. Just, T. Itagaki, G. Stoll (2010) An Approach to Intelligently Crop and Scale Video for Broadcast Applications. Proceedings of the 2010 ACM Symposium on Applied Computing

[24] European Telecommunications Standards Institute (2006) Specification for the use of Video and Audio Coding in DVB services delivered directly over IP protocols. European Telecommunications Standards Institute

[25] International Telecommunication Union (2007) Methodology for the subjective assessment of video quality. Recommendation ITU-R BT.1788.

[26] L. Sachs, Z. Reynarowych (1984) Applied Statistics: A Handbook of Techniques, Springer Verlag, New York