Computer Vision Technologies for Repurposing Multi-distribution Systems

Claire-Hélène Demarty, Siwar Baghdadi
Thomson R&D France, 1, avenue Belle Fontaine – CS 17616 35576 Cesson-Sévigné - France
E-mail: claire-helene.demarty@thomson.net, siwar.baghdadi@thomson.net

Keywords: Multi-distribution systems, repurposing, computer vision, HMM, BN, DBN, metadata.

1 INTRODUCTION

Content creators are looking forward to spreading their offer of services, by repurposing their original content, usually first dedicated to broadcast, to other distribution channels. But multiplying the distribution channels does not and cannot mean multiplying the production cost: one really wants to “produce once, and repurpose many”. Moreover, “the faster the content production the better the workflow” remains a key aspect even for new distribution channels. To answer this need while taking into account the above constraints, innovative repurposing solutions are introduced from the initial stages of the broadcast production workflow. These solutions enable to speed up the process of high quality content creation, while generating new revenue. Simultaneous production of content for the Internet, mobile and IPTV is then achieved through a suite of automated tools, grouped together in a single workflow.

Repurposing content implies a necessary adaptation of this content to the constraints of the targeted distribution channels. These new multi-distribution systems should therefore allow to adapt editing, encoding and reformatting of the content to new context such as Internet or mobile devices. And once again, all this has to be done at the lowest cost, meaning that these new functionalities should be provided as automatically as possible, and early in the creation chain, i.e. in the production studio. However, in the proposed systems already on the market, some of these operations are still done manually. This is often the case for the re-edition of the content: one has to understand its semantic meaning to re-create a new content which fulfills the characteristics of a new distribution channel. For example, when mobile devices are targeted, one wants to reduce the duration of the original video by extracting some key events summarizing its semantic meaning. When Internet is targeted, some new access to content is required: one wants to go directly to the interesting parts of the video. Both cases, extracting highlights and structuring the content, require a semantic understanding of the content. One challenge is thus to move from a tedious and expensive manual semantic indexing to a fully automated solution.

In order to help and increase the automated aspect of repurposing, and especially the re-edition step, this paper proposes some new software modules based on computer vision algorithms. These modules either create some metadata from a video analysis of the content, or use these metadata to help, among others, the content re-editing. The addition of these modules into the production workflow will result in a modified workflow. This workflow and the underlying modules will first be introduced. Third section will then describe in more details the advanced computer vision technologies these modules rely on, among which the statistical techniques HMMs (Hidden Markov Models) and BNs (Bayesian Networks). Finally, we will focus on the application of these techniques to structuring and event detection in the context of sports and commercials contents, as these contents generate huge economic interests.

2 DESCRIPTION OF THE MODIFIED WORKFLOW WITH COMPUTER VISION MODULES

The professional content production workflow is no longer concentrated on the only broadcast distribution channel. It has to address other distribution means, such as mobile devices, Internet or IPTV. The associated multi-distribution systems are inserted right after the production studio to help automating the process of content adaptation. The resulting workflow is shown in Figure 1.

To ensure the metadata permanence and consistency, an ideal production workflow should be able to propagate these metadata in the whole production chain. The metadata produced in the workflow should be re-usable anywhere. This is actually not the case, as some key elements of production studios (e.g. the switcher) do not propagate such information, and most of the time the metadata produced are lost and should be generated again.
Fig.1: Production workflow with multi-distribution system.
We propose to add some software functionalities to increase the automated aspect of the repurposing process. These functionalities could have been added in parallel to the multi-distribution system itself, in one single module, with the advantage of lowering the integration cost. As long as the production workflow is not able to propagate the metadata, this solution is without contest the best one. It should be nevertheless underlined that, in that case, a potential re-use of the metadata produced by these functionalities in more classical distribution channels such as broadcast will not be available. We think nevertheless that future production studios will ensure the metadata persistency. Besides, it is interesting to show that computer vision technologies can be applied on raw or edited content with different applications schemes. The available prior knowledge is moreover not the same depending on where you are in the production studio. After the switcher for example, production rules are available and may be used by computer vision, whereas when at the capturing stage, one may benefit from the known location and the different points of view of the cameras.
For these reasons, we propose to separate these new functionalities in three different software modules, even if, in our application schemes detailed in section 4, only repurposing of produced content is targeted.
The modified workflow, shown in Fig.2, presents therefore three new software modules:
- before the switcher for real-time processing of live raw streams;
- at the output of the switcher for real-time processing of the already produced stream;
- in parallel of the multi-distribution system for the re-edition of content.

Fig.2: Production workflow with multi-distribution system and added computer vision-based software modules.
We will detail these modules in the following sub-sections, but we may already separate them in two categories: first and second modules are generating some metadata, which are then re-used in the third one. As already said, the modules are based on computer vision techniques. Computer vision provides some powerful tools to automatically analyze multimedia content. As a result, we have access to a description of the content in terms of features of different
semantic levels. This description and all its associated features correspond to the metadata generated by the first two modules.

As already said, for our current repurposing schemes, content re-editing is performed on the content output by the switcher, i.e. on produced content. The key metadata are thus the ones output by the second module only. We could nevertheless imagine further systems relying also on the first module output. We may want, for example, to select at the same time content from different cameras according to the distribution channels (e.g. a zoom view for mobile distribution and a large view for broadcast).

A part from its potential use for repurposing, the first module may also help automating the production itself. This is what is shown in section 2.1.

2.1 Metadata generation: first and second modules

The first two modules are dedicated to the generation of metadata describing/indexing the content. Their difference relies on the nature of the content to process: raw or already produced content.

Metadata generation from raw content module

Some indexing features may be extracted directly at the live capturing step, in order to help the automation of the production itself. The innovative technology named autoCORRECT™ is an interesting example of how computer vision can help automation. AutoCORRECT (demonstrated by Thomson at NAB this year) allows for unmanned, automatic tracking of anchors and other on-air talent in a studio environment. This technology automatically makes adjustments to pan, tilt, and zoom functions on the Thomson robotic cameras in order to keep the talent properly framed within the shot. The associated requirements are therefore limited to the studio environment. The targeted situation may vary from a seated person presenting the main news, to a standing weather man. No assumption is made on the background appearance. In the present version only one talent is considered. Nevertheless, the system should consider possible short occlusions, slight lighting variations and talent disappearance for a long duration. To address the practical issue of keeping the talent properly framed within the shot, autoCORRECT relies on the use of three main advanced vision-based technologies: face detection in an image, object tracking in a video and robotics to ensure a convergence of the targeted location and the center of gravity of the object. All these three technologies provide an automatic content analysis of the raw stream and result in the generation of metadata, namely the objects (faces) description and their location in the frame.

The computer vision technologies used in autoCORRECT consist in robotics, face detection, object tracking. For the two first items, one can refer to [6, 7] for more information. The tracking technology, patented by Thomson, is named “probabilistic multi cue tracker”. It relies upon a novel combination of probabilistic color and randomized template tracking. The tracker itself was tested against several state-of-the-art trackers using the Performance Evaluation for Tracking and Surveillance (PETS) 2005 VIVID online evaluation program conducted by the Carnegie Mellon Robotics Institute and quantitative comparisons were drawn. The results obtained via this evaluation demonstrate a significant improvement compared to all the existing state-of-the-art trackers in almost all of the tested sequences (see: http://www.vividevaluation.ri.cmu.edu). For further details regarding the tracking algorithm the reader is referred to papers [8, 9].

Metadata generation from edited content module

The second module of Fig. 2 will do video content analysis of already produced content. This implies that information such as production rules may be used as a prior knowledge, which can help the extraction of the indexing features. As a consequence, trying to extract some events of interest or the whole temporal structuring into shots and scenes of the produced stream is relevant in this second module only. An example of events of interest that one may want to extract is the goals or key actions that take place during a soccer match. These events are typically what should be extracted to build a summary of the match to be transmitted to mobile devices. The splitting of a tennis match into its different games, sets and points is also an example of the typical temporal structuring that may be of interest for a repurposing application targeting the Internet. Another example of such a structuring could be the splitting of a TV news program into different news subjects. Section 3 will describe how we may automate the process of extracting such kind of events or structuring, with computer vision technologies. But right now, we may note that these technologies, applied on the stream at this stage, will result in a set of metadata describing the targeted events of interest or the temporal structuring, e.g. by means of timecodes of the events/video parts and labels/features describing them. These metadata will then be re-used in the third module.

2.2 Re-edition step: third module

The third module can be seen as a new component of the multi-distribution system to ease the re-editing of the content, with the help of the metadata coming from modules 1 (if possible) and 2.

For mobile devices, as already said, a known use case is to re-edit a summary of the interesting parts from an original content. In current multi-distribution systems, these interesting parts are selected manually by a user. To do so, the
operator has to visualize the whole content linearly. In the worst case, the user does not know what is in the content; he therefore has to see it all to select the good events. With the use of the third module, the selection is done automatically thanks to the metadata associated with the content. The module proposes only the best events of the video to the user, who will have to validate the selection in a semi-automatic process, but with an evident gain of productivity. A research prototype resulting from our work in computer vision for modules 2+3 is shown in Figure 3, in the context of repurposing soccer matches.

In this prototype, the resulting summary is a raw concatenation of all the events of interest. The creative part of building the summary with some transitions between the events for example is still to be done manually. Besides, some work remains on the definition of the events limits. If our technology enables to find markers of the events (some precise shots in the video), we do not have access to the exact beginning and ending of the event, which may somehow be quite subjective anyway: when does an action really begin in a soccer match? And when does it end? When the ball passes the goal line, when the players cheer or at the end of the following replay? There is still some work to do to answer questions such as: how to select among the detected events those which will better match some specific requirements on the summary itself. The summary could, for example, have a limited duration; it could be user-profiled to satisfy a fan of team1, etc.

These points can be seen as some perspectives of our current research in order to go further in the automation process, toward a summary proposal as close as possible to what the user expects.

3 FOCUS ON UNDERLYING TECHNOLOGIES

In this section, two powerful and innovative statistical techniques used in computer vision will be presented, namely Hidden Markov Models and Bayesian Networks. Both are probabilistic model-based techniques; we will therefore present first a common and general description of this kind of techniques, before detailing their own specificities.

3.1 Probabilistic techniques principle

In computer vision, when structuring of video content is targeted, one usually distinguishes two kinds of methods: deterministic and probabilistic. Deterministic methods often consist in a set of decision rules. These rules, which are specific to the application domain, are heuristically defined by an expert. Depending on the application, their definition is not always explicit. Most of the time, it consists in fixing some thresholds that are generally not easy to tune. The deterministic methods can however be implemented rapidly and give a quick access to the resulting structuring of the input content. Besides, for some specific applications with rules explicit enough, the structuring results are good. Their lack of generality and rigidity against applications is nevertheless a serious drawback.

Probabilistic inference techniques, to which HMMs and BNs belong, offer some smoother ways to model data. They allow to take into account more complex rules, without the step of choosing thresholds. Once the relevant model for a given problem is built, the probabilities that some input data match a model are computed. The input video model consists in both a structure and some parameters. In the case of BNs and HMMs, the structure is represented by a directed acyclic graph whose nodes represent random variables and whose edges represent statistical dependence relations between these variables. In such a graph, an edge between two nodes A and B can be informally interpreted as...
A causes B. As for the parameters, a conditional probability distribution (CPD) is associated to each variable, i.e. to each state. These CPDs are the parameters of the model.

Once the structure is defined, such probabilistic techniques work in two phases, described in Fig. 4. In a first phase, the structure is built and the parameters (CPDs) of the model are estimated through learning from some training data. In the inference or decoding phase some new data is put in the model and, according to its answer, a decision is taken whether this data corresponds to the model or not. The parameters’ estimation is obviously dependent on the choice of the learning data which should be representative of the model to describe. These learning data consist in manually annotated data that correspond to the expected content description to be output by the model. Once the parameters of the models are learned, one may use it to decode some new content. If the first learning phase is done off-line and is not real-time (but is done only once during an initialization step), the decoding phase is usually real-time.

Thanks to the parameters’ estimation done during the learning phase, these models allow the structuring of more complex content and with an increased generality, compared to deterministic techniques. For a given application, the diversity of input content (diversity of tennis matches for example) is much better represented by this kind of probabilistic models than by some rigid rules.

3.2 Hidden Markov Models

The above general properties and definitions stand for the HMMs. The reader can refer to [1] for a detailed tutorial.

Fig. 5: Different structuring units of a tennis match represented as HMMs: first serve + rally, rally, replay, break.

Some examples of HMMs’ structures are given in Fig. 5. The application targeted is the structuring of tennis matches, which will be further described in section 4.1. In this example, the random variables correspond to the different states of a shot during a tennis match: global views of the ground (G), dissolves (D), close-up (CU), breaks (B) and other shots (O). The input data are the different shots of the video. The presented structures represent the different structuring units of a tennis match. In our tennis case, a CPD is the probability for one shot to be in state B, knowing that the previous shot was in state A.

Compared to the general definition given in the above section, HMMs nevertheless only allow a restrictive subset of structures. This leads to models that still have difficulty handling the temporal characteristic of some events. HMMs have nevertheless proved to be very efficient in numerous applications, especially when a dense structuring is required. Dense structuring means that for each shot, it is possible to find a subset of the structure to which it belongs. Contents with dense structuring are to be opposed to contents with sparse events in which only some parts – the events – correspond to a well-defined structure. Opposed to the events, is the “rest of the content” without any specific structure. Some examples of content with dense structuring are typically tennis matches, whereas soccer matches correspond to content with sparse events. A tennis match can indeed be split into several sets, each set being split into different games, each game into different points, and finally each point into different shots which belong to one of the sub-structures presented in Fig. 5. In a soccer match, on the contrary, no such dense structure exists. We have long phases of game shots with sparse events corresponding to actions (goals, corners, faults, etc.), which may occur at each instant.

The limitation in the type of allowed structures for the HMMs results nevertheless in very efficient inference algorithms. As the structures are more specific, some optimization can be used to speed the process. Moreover, these techniques are studied for several years now by the research community, resulting to well-known characteristics, advantages and limitations. And this participates also to build efficient algorithms. We will present in section 4.1 an example of application which was used successfully for the structuring of tennis matches.
3.3 Bayesian Networks

BNs are newer in the literature. They are a generalization of HMMs, for which all structures are allowed. Bayesian networks directly correspond to the directed acyclic graph model described in section 3.1. Dynamic Bayesian networks (DBNs) are the dynamic version of Bayesian networks, which allow encoding time evolution in a much more better way than what is done in HMMs. In DBNs, the time flow is discretized, and a static Bayesian network is assigned to each temporal slice. Variables of different time-slices are connected through directed edges, which explicitly represent the time flow in terms of conditional dependence. DBNs handle the temporal correlations that exist in audio-visual signals through the connections that exist between nodes of different time slices. The modelling of event duration is therefore much more precise and efficient in DBNs than in HMMs.

Whereas HMMs allow a good modelling of dense structuring, BNs are well-adapted to sparse events extraction. All these properties work towards an increased generality of the modelling with BNs: more complex content and events can be represented through BNs.

In [5], we have reminded that, in all the video indexing frameworks based on Bayesian networks that can be found in the literature, the structure of the model is usually designed manually. However, Bayesian networks provide interesting opportunities for automatically learning the model's structure from training data. Structure learning is a key property of BNs, as it goes along the line of even more generality. Building the structure manually needs using expert knowledge on the different interactions between the variables used. Unfortunately, this knowledge is not always available. Moreover it is expensive and tedious to get, in particular for complex problems of semantic video analysis. Structure learning offers a first advantage of automatically building a model from some training data. It also allows the system designer to get a better understanding of the system under study, as it may point out some causal relations between features. Moreover, structure learning provides a help to choose between all available features those which may be relevant or not for the targeted application. Non relevant features should indeed be disconnected from the learned network. This was illustrated in [5], in the case of commercials detection. Through two learning steps, one for the structure and one for the parameters, BNs offer therefore more generality against applications than HMMs do, as less expert knowledge is required. It should nevertheless be noticed that the available inference algorithms in the literature are computationally less efficient in the case of BNs. This is due to the fact that they have to handle much more general networks.

As a conclusion, we may say that this property of automatic structure learning is of great interest in a video indexing framework such as what we propose for repurposing multi-distribution systems. Once a change of application occurs, one will indeed only have to provide some different learning data, in order to allow the system to learn another structure, but the whole detection module will not have to be manually re-designed.

4 APPLICATION TO THE REPURPOSING OF SPORTS CONTENT

In this section, we will illustrate how HMMs and BNs may be used in some application cases first for the Internet, and then for mobile devices.

4.1 Temporal modelling for the Internet

In some previous work [2, 3, 4], we successfully used HMMs to automatically structure tennis matches. As already said in section 3.2, tennis matches have a natural dense hierarchical splitting into sets, games, points and shots. This hierarchical structure can be retrieved using an extension of the HMMs, the hierarchical HMMs. In the HHMMs, each state itself is a small HHMM. We distinguish in this case nodes which are really related to observations (leaves of the graph) and intermediate nodes. In our tennis application, the global resulting HHMMs are described in Fig.5 and 6. Fig.6 presents the global hierarchical structure, whereas Fig.5 presents the small sub-structuring units directly related to the observations.

The input data for the HHMM are the different shots of the video together with some low-level multimedia features, describing these shots. The different shots and features are extracted through computer-vision techniques from the audio and video channels. As for the features, we used the player’s position on the field, visual similarity through dominant color, activity measure, presence of speech, music, silence, applauses and ball noise. All these features are the extracted observations for each shot. From these observations and with the structure designed in Fig.5 and 6, a learning step is launched on a learning database of tennis matches to estimate the model parameters. These parameters and the previous structure are then used to decode some new tennis matches. The percentage of correct shot classification is 60% for the games, 93% for the points and 94% for the logical structuring units of Fig.4.
From this HHMM framework, we therefore end with a temporal and highly semantic splitting of tennis matches, as illustrated in Fig.7. The resulting structuring allows to have a non-linear access to any moment of the document. This type of non-linear access is particularly relevant for web applications. Through one click only, a user may for example jump from the first game of first set to the second game of third set of the match. Some additional computer vision features were also used to compute a level of interest for each part of the tennis match, leading to a ranking of the different shots according to this level of interest. When representing this ranking together with the match structuring, one also allows the internet user to access directly to the most interesting parts of the match. This is what is illustrated in Fig. 7, where the yellow stars stand for the estimated level of interest. The metadata resulting from the structuring are: all low-level indexing features that were extracted, plus the belonging class for each shot, together with its timecodes. All this is provided in XML format.

4.2 Event extraction and summary building for mobile devices

In this section, the targeted application is repurposing of content for mobile devices. In such an application, one usually wants to reduce the duration of the original video by extracting some key events summarizing its semantic meaning. In current multi-distribution systems, such summaries are manually built thanks to an operator who visualizes the entire document, selects the events of interest and builds the desired summary from these events. We propose to use the BNs to automatically extract these events. The role of the operator is then reduced to viewing and validating the event selection and editing the summary. We propose to illustrate the event selection first in the case of commercials detection and then on soccer matches. In the first case, the events to detect are the commercials parts of the content. In the second case, the events are all key actions, goals, faults, etc. that may occur during a soccer match. It should be noted that both types of events are sparse events in the video.

The first application – commercial detection – was already detailed in [5]. Structure learning was used to build the static part of the Bayesian network. The nodes of the Bayesian network were the following audiovisual features for a given shot: shot length, motion intensity, color coherence, surface and number of ‘text caption’, and one audio feature representing the variation of short-time energy (LSTER). One additional node stands for the event to detect, namely the presence of commercials. Once these nodes were defined, the system was able, through the used of training data, to
build the structure of the video proposed in Fig. 8. This automatically built structure underlines the correlations that exist between features. It also proposes a semantic and explicit model of the video content under study, which does correspond to our knowledge of the domain. In [5], the capability of such a system to select only relevant features was also shown: some non-pertinent features artificially introduced in the system were discarded. After another learning step dedicated to the parameters’ estimation, the classification results on new content reached 93% of recall and 80% of precision.

Some current work is on-going to apply structure learning to event detection in soccer matches. This use case is much more complex than what was done for the commercials detection application. We have to handle roughly 50 nodes in the network; 10 audiovisual features were extracted and temporally duplicated around the event instant to five distinct time slices. This temporal duplication of features is due to the fact that, on the contrary to commercial detection where the event and the extracted features occur at the same time, for soccer matches the event may occur before or after the presence of some features. For example, the replay feature has a positive answer after the shot corresponding to a key action. In order to model this delay, we choose to duplicate the features at different time slices in a first step. A future evolution of the system will be to use real Dynamic Bayesian networks to handle some duration models of the event. Our first tests concentrate on a static Bayesian network, but with temporal features. First results have shown that a compromise have to be chosen between the complexity of the network and the ability of the model to truly represent the data. With this in mind, very first results of structure learning tend nevertheless to prove that the system has also the capability to automatically model the correlations that exist between the features and the event, in such a complex application. These are encouraging results, although we still need to go on with the study of the property of structure learning. Thus, we have to try to apply structure learning to the dynamic part of Bayesian networks.

5 Conclusion and Perspectives

In this paper, some new modules based on computer-vision technologies were proposed in order to help automating the process of repurposing content. These software modules may be added directly in the production workflow. They provide an automatic extraction of metadata, which are then used to automatically re-edit the content. We therefore add two new automatic functionalities to the repurposing process, first by automatically structuring the input content and then by extracting some events of interest. The two key computer vision technologies used in the modules, namely HMMs and BNs, were described and some applications for repurposing content for the Internet and for mobile devices were detailed. It was shown that these two technologies go towards an increased generality of the process, as they provided a more general model of video content which adapts to complex content. An additional property of the BNs was detailed, namely structure learning. First results on the use of this property have shown that it participates to the increase of generality of the system. The model structure is indeed no more provided manually by an external expert, but is learnt by the system. Besides, the resulting structure provides some interpretation of the correlations that exist between the features, and helps in the selection of the relevant features. Some work still remains however to study this property in the context of more complex applications and with complete dynamic Bayesian networks. In the application context of multi-distribution systems itself, some additional study may also be conducted to find some means to go ahead in the building of summaries from the extracted events, and to help the events selection in order to match some specific summary requirements.

References