

APPARATUS AND METHOD FOR ANALYSING A VIDEO BROADCAST

The invention relates to an apparatus and method for analysing a video broadcast. In particular, the invention relates to an apparatus and method for determining a likelihood
5 that a candidate commercial boundary in a segmented video broadcast is a commercial boundary. The invention also relates to an apparatus and method for classifying a commercial broadcast in a pre-defined category. The invention also relates to an apparatus and method for identifying a boundary of a commercial broadcast in a video broadcast and classifying the commercial broadcast.

10

TV advertising is ubiquitous, perseverant, and economically vital. Millions of people's living and working habits are affected by TV commercials. Today, TV commercials are, generally, produced for 30 or 60 seconds, costing millions of US dollars to produce and air. One 30-second commercial in prime time can easily cost up to 120,000 US
15 dollars [Reference 1 – see appended list of references]. Millions of people are reached by commercials which modify their living and work habits, if not immediately, at least later.

20

Advertising may be considered an organised method of communicating information about a product or service which a company or individual wants to promote to people. An advertisement is a paid announcement that is conveyed through words, pictures, music, and action in a medium (e.g., newspaper, magazine, broadcast channels, etc.). Although the costs of creating, producing, and airing a TV commercial are staggering, television is one of the most cost-effective media. Its advantages are impact, credibility,
25 selectivity, and flexibility [1]. In the world of satellite and cable television, TV commercials have become indispensable for most clients. Many cable channels fill in 10 to 12 minutes of a 30-minute serial with commercials.

30

In just one day, a TV viewer may be exposed to hundreds of commercials. Over a year, it can be tens of thousands. With the advance of digital video recording and playback systems, many works (for example references [2] – [4]) have focused on automatically locating a commercial disposed within a video stream towards “commercial skip” types

of applications. When a copy of the programme is created for viewing at a later time, many users are not interested in the content of commercials or promotions that are interposed within the television program. Automated commercial detection techniques can replace a user's manual skipping operation. Such work deals with a series of
5 consecutive commercials as a whole block.

Several methods have been disclosed for automatically locating the boundaries of video programs and the boundaries of TV commercials in computerised personal multimedia retrieval systems. Common recording methods of television programmes include the
10 use of a Video Cassette Recorder (VCR), computer magnetic hard disk using an MPEG video compression standard, and digital versatile disk (DVD). However, there is no systematic and generic method that can reliably detect each individual commercial's boundaries. Although we can make the assumption that a black frame and a short silent section of 0.1 to 2.0 seconds appears before and after each TV commercial, any
15 commercial in which post-editing effects are not present cannot secure precise boundaries by detecting black frames and quiet sections.

United States Patent No. 6100941 discloses a method and apparatus for locating a commercial within a video data stream. The average cut frame distance, cut rate,
20 changes in the average cut frame distance, the absence of a logo, a commercial signature detection, brand name detection, a series of black frames proceeding a high cut rate, similar frames located within a specified period of time before a frame being analysed and character detection are combined to provide a commercial isolation apparatus and/or method with an increased detection reliability rate. However, a method for
25 detecting an individual TV commercial's boundaries is not disclosed in this invention.

Reference [2], noted above, discusses a method of extracting a number of audio, visual, and temporal features (such as audio class histogram, commercial pallet histogram, text location indicator, scene change rate, and blank frame rate) within a window around
30 each scene boundary and utilises an SVM classifier to classify each candidate segment into commercial segments or programme segments.

Reference [18], discloses a technique for a commercial video's semantic analysis. However, this work is limited to the mapping between low-level visual features and subjective semiotic categories (i.e., practical, playful, utopic, and critical). It utilises heuristic rules used in the practice of commercial production to associate a set of perceptual features with four major types, namely, practical commercials, playful
5 commercials, utopic commercials, and critical commercials.

Most previous work in the appended reference list on TV commercial video analysis focuses on automatically locating a commercial disposed within a video data stream
10 towards "commercial skip" type of applications. Many audio-visual features about blank frames, scene breaks, action, etc. have been exploited to characterise commercial video segments in general. Heuristic rules or machine learning algorithms (for example, reference [2] above) are employed to generate a commercial discriminator. Shots and sequences are a useful level of granularity, as a few useful features (e.g., scene change
15 rate or shot frequency in [2], etc.) rely on shots directly, and many statistically meaningful features (e.g., blank frame rate and audio class histogram in [2], average and variance of edge change ratio and frame differences) have to undergo the accumulation over a temporal window.

20 In general, features-based commercial detection approach only allows an approximate location of the commercial blocks.

The invention is defined in the independent claims. Some optional features of the invention are defined in the dependent claims.

25 Apparatuses incorporating features defined in the appended independent claims can be used to identify a TV commercial's boundary and TV commercial classification by advertised products or services. A flexible and reliable solution may resort to the representation of intra-commercial characteristics that are of interest to indicate the
30 beginning and ending of a commercial, and to indicate the transition from one commercial to the other.

Thus, apparatuses implementing the features of the independent claims may provide any or all of the following advantages:

Apparatuses implementing the techniques described may provide a generic and reliable system and method for locating each individual TV commercial within a video data stream by utilising machine learning to assess a likelihood a candidate commercial boundary is a commercial boundary (for example, as a boundary or not) on the basis of a set of mid-level features, which are developed to capture useful audio-visual characteristics within a commercial and at the boundary between two consecutive commercials. Some apparatuses implementing the invention utilise a binary classifier to assess simply whether or not the candidate commercial boundary is a commercial boundary.

It may be possible to provide a method for automatically determining key image frames visually marked with relevant information about a product or service such as corporate symbols, brand names, appearance, mild encouragement captions and contact information within each individual TV commercial, which can be fed into OCR or object recognition modules for extracting semantic information. Video shots containing such key image frames, together with some modest encouragement coming from the announcer/voice-over, are often employed to highlight the offer at the end of a commercial. This may be a reliable indicator that the video shot in question is in the vicinity (in the video broadcast stream) of a commercial boundary.

It is also possible to provide a method for modelling audio scene changes, which are used to represent the characteristics of audio signal changes occurring with the transition of different TV commercials. Optionally, an alignment algorithm is carried out to seek the most probable position of audio scene change within a neighbourhood of a video shot transition point.

Boundary classifier modules may comprise a set of mid-level features to capture audio-visual characteristics significant for parsing commercials' video content (e.g. key frames, structure), Black frame inclusive/exclusive multi-modal feature vectors, and a

supervised learning algorithm (e.g. support vector machines (SVMs), decision tree, naïve Bayesian classifier, etc.).

- Apparatuses implementing the techniques described may provide a system and method
- 5 for automatically classifying an individual TV commercial into a predefined category. This may be done according to advertised product and/or service by making use of, for example, ASR (Automatic Speech Recognition), OCR (Optical Character Recognition), object recognition and IR (Information Retrieval) techniques.
- 10 Commercial categoriser modules may comprise ASR and OCR modules for extracting raw textual information followed by spell checking and correction, keyword selection and keyword-based query expansion using external resources (such as google, encyclopaedia and dictionary), SVMs-based classifier trained from external resources such as public document corpus categorised according to different topics, and IR text
- 15 pre-processing module (such as porter stemming, stopper words removal, and vocabulary pruning); visual-based object recognition (e.g. car, computer, etc.) may be useful in the case of weak textual information.

The present invention will now be described, by way of example only, and with

20 reference to the accompanying drawings in which:

- Figure 1 is a block diagram illustrating an application paradigm of TV commercial segmentation, categorisation and identification. Figure 1 is the Figure 1 used in the published paper not that of the specification;
- 25 Figure 2 is a block diagram illustrating an architecture for a boundary classifier and a commercial classifier;
- Figure 3 is a process flow diagram illustrating a first set of techniques for determining a likelihood that a candidate commercial boundary is a commercial boundary;
- Figure 4 is an architecture and flow diagram illustrating a second technique for
- 30 determining a likelihood that a candidate commercial boundary is a commercial boundary;
- Figure 5 illustrates a series of Image Frames Marked with Product Information (FMPI);

- Figure 6 is process diagram illustrating low-level visual FMPI feature extraction;
- Figure 7 is a line graph shows results of system performance for FMPI classification by using different features;
- Figure 8 shows a series of images incorrectly classified as an FMPI frame;
- 5 Figure 9 is a block diagram illustrating an Audio Scene Change Indicator (ASCI), alignment of audio offset and training process flow;
- Figure 10 is a bar graph illustrating statistics of time offsets between an audio scene change and its associated video scene change in news programs and commercials;
- Figure 11 illustrates a Kullback-Leibler distance-based alignment process for audio-
10 video scene changes;
- Figure 12 is a graph illustrating a series of Kullback-Leibler distances calculated from 200 samples of ASC and 200 samples of Non-ASC;
- Figure 13 is a table illustrating the simulation results of ASCI;
- Figure 14 is a graph illustrating statistics of the number of shots and the duration of TV
15 commercials in the simulation video database;
- Figure 15 is a line graph illustrating the simulation results of an individual TV commercial's boundaries detection;
- Figure 16 is a block diagram illustrating the architecture of a commercial classifier;
- Figure 17 is a process flow diagram illustrating a first process for classifying a
20 commercial;
- Figure 18 is an architecture/process flow diagram for a second commercial classification method;
- Figure 19 is a process flow diagram illustrating the method for keyword determination and proxy assignation of Figure 18 in more detail;
- 25 Figure 20 is a process flow diagram illustrating the method for word feature selection of Figure 18 in more detail;
- Figure 21 illustrates an example of actual speech script, ASR generated speech script, and an acquired article from World Wide Web for the purpose of query expansion/proxy assignation;
- 30 Figure 22 shows a group of key image frames containing significant semantic information in TV commercial videos;

Figure 23 is a pie chart illustrating system performance results for TV commercial classification;

Figure 24 is a bar graph illustrating the number of commercials in which the OCR and ASR of Figure 18 recognise brand names successfully;

- 5 Figure 25 is a bar graph illustrating the F1 values of classifications based on three types of input; and

Figure 26 is table illustrating results of classification processes.

- Referring to the illustrative paradigm in Figure 1, four points are summarised to explain the motivations and potential applications of TV commercial video segmentation, categorisation and identification. Firstly, as advertisers spend a great deal of money, it is possible for them to verify their commercials are broadcasted as contracted. A preliminary stage is to determine the boundaries of individual commercials. Accurate boundaries are useful for effective clip-level video matching and subsequent statistics of real duration in TV broadcast. Secondly, research shows that most people do not mind TV advertising in general, although they dislike certain commercials; they do not like to be yelled at or treated rudely; they want to be respected [1]. With the advance of digital TV set-top boxes in terms of powerful processors, large hard disks and internet access, it may be possible to furnish consumers with a TV commercial management system, which detects commercial segments, determine the boundaries of individual commercials, identify & track new commercials, and summarise the commercials within a period by removing repeated instances.

- Given a decent interface, such a system may change a TV viewer's passive position. A user can apply positive actions (e.g., search, browse, etc.) to the commercial video archive. As advertising in the mass media is basically incidental to consumers' use of the media, described techniques may indirectly improve the reachability of TV commercials. Thirdly, all advertisements deal with one of three concepts: ideas, products, and services [1]. TV commercial classification with respect to the advertised products or services (e.g., automobile, finance, etc.) helps to fulfill the commercial filtering towards personalised consumer services. For example, an MMS or email

message (containing key frames or adapted video) on the commercials of interest to a registered user can be sent to her/his mobile device or email account.

- 5 Fourthly, the technology of TV commercials has changed significantly; they are almost always edited on a computer; the appearance all starts with the MTV generation and MTV-type commercials are more visual, more quickly paced, use more camera movement, and often combine multiple looks, such as black and white with colour, or stills with quick cuts [1]. Accordingly, a TV commercial archive system including browse, classification, and search may inspire the creation of a good commercial.
- 10 Marketing companies may even utilise it to observe competitors' behaviours.

- Two challenging tasks are addressed by techniques described below: individual commercials' boundary detection (ComBD) and commercial classification in terms of advertised products/services (ComCL). The first is the problem of video parsing; the
- 15 latter is that of semantic video indexing. In TV streams, a commercial block consists of a series of individual commercials (spots). Each spot may be dealt with as a semantic scene. The process of detecting such scene transitions within a block is referred to as commercial video parsing. In a classified advertisement (often found in most newspapers), one can easily find information useful to determine if the advertised item
- 20 is to be bought. Accordingly, semantic commercial video indexing is meant to accomplish such classified TV advertisement through video content analysis techniques. Since an advertising campaign concerns many topics such as babies, cars, entertainment, fashion, food, money, sports, and so on, one may choose some representative categories of products or services to explore the solution. Some
- 25 described apparatuses deal with this via multimodal analysis.

- Once individuals' boundaries are determined, various video clip matching methods can be used to identify commercials (ComID). One issue lies in a compact and robust signature for representing commercial video content. The other issue is to accelerate the
- 30 clip search in a large database. Compared with ComBD and ComCL, ComID can be easily addressed by existing or modified methods.

A first apparatus for identifying commercial boundaries and classifying commercials in a pre-defined category is discussed with respect to Figure 2. The apparatus 60 comprises TV commercial detector 62 configured to locate boundaries of video programmes and commercial broadcasts in the video broadcast and to derive a segmented video broadcast, video shot (or frame) transition detector 64 configured to identify candidate commercial boundaries in the segmented video broadcast, boundary classifier 66 for assessing a likelihood a candidate commercial boundary is a commercial boundary, and commercial classifier 68. Optionally, boundary classifier 66 is a binary boundary classifier. As shown in Figure 2, boundary classifier 66 comprises FMPI recognition module 70 for determining whether a particular frame comprises an FMPI frame. Boundary classifier 66 also comprises an SVM training module 74 configured to train the classifier model 74 with video frames of the segmented video broadcast which comprise product information (e.g. FMPI frames). Additionally, boundary classifier 74 assesses whether a candidate commercial boundary can be considered to be a commercial boundary. The boundary classifier performs this assessment for an FMPI frame with FMPI recognition module 70. As will be discussed further below, the boundary classifier may, optionally, comprise ASC (audio scene change) recognition module 76, silent frame recognition module 78, black frame recognition module 80 and HMM training module 82 used to train an HMM (Hidden Markov model) utilised in the ASC recognition module 76.

In an example embodiment, (at least) visual features are extracted within a symmetric window of each candidate commercial boundary location from a video data stream as shown in Figure 3. Multi-modal audio-visual features are extracted in apparatuses implementing ASC and/or silence recognition. It will be appreciated that although Figure 3 illustrates a multi-modal technique, it has been found that excellent results are obtainable (again, described below) with an implementation of FMPI techniques only. Boundary classification is carried out to determine whether a candidate commercial boundary is indeed a commercial boundary of each individual TV commercial. The input video data stream can be any combination of video/audio source. It could be, for example, a television signal or an Internet file broadcast.

The disclosed techniques have particular application for digital video broadcasts. Implementation of the techniques described are extendable to analogue video signals. The analogue video signals are converted to digital format prior to application of the techniques.

5

The disclosed techniques may be implemented on, for example, a computer apparatus, and be implemented either in hardware, software or in a combination thereof.

10

Referring now to Figure 3, the process flows for a first set of techniques for assessing the likelihood a candidate commercial boundary is a commercial boundary will now be described. Referring first to Figure 3a, process 100 starts at step 102. At step 104, the input video broadcast signal is partitioned into commercial and programme sections, as is known. At step 106, a candidate commercial boundary is detected by use of, for example, a video shot detector 64.

15

At step 108, image marked with product information (FMPI) recognition is carried out. From this, and as will be apparent from below, FMPI recognition used in isolation may provide perfectly acceptable results for assessing the candidate commercial boundary is a commercial boundary at step 110. Optionally, the boundary classifier determines a likelihood the candidate commercial boundary is a commercial boundary in dependence of a determination the candidate video frame comprises an audio scene change; that is ASCI recognition may be implemented at step 114 and/or silence and black frames recognition may be implemented at step 116. FMPI recognition is discussed in more detail with reference to Figure 3b and ASCI recognition is discussed in more detail with reference to Figure 3c. The process of Figure 3a ends at step 112.

25

Therefore, as illustrated in Figure 3a, there is provided an apparatus which determines a likelihood a candidate commercial boundary is a commercial boundary. The apparatus comprises a boundary classifier which determines whether a candidate video frame associated with a candidate commercial boundary of a segmented video broadcast comprises product information and determines a likelihood the candidate commercial

30

boundary is a commercial boundary in dependence of the determination the candidate video frame comprises product information.

Optionally, the boundary classifier is determines a likelihood the candidate commercial
5 boundary is a commercial boundary in dependence of a determination the candidate video frame comprises an audio scene change. As a further option, the boundary classifier is configured to make the classification according to a determination the candidate video frame (or frames thereof) comprises audio silence or video black frames.

10

Referring to Figure 3b now, a candidate boundary is detected at step 106 of Figure 3a. At step 120, the MPEG motion vectors of the video signal are queried in order to identify key frames at step 122. The identification of key frames will be described in more detail below. At steps 124, 126, the video frame comprising the candidate
15 commercial boundary is parsed in order to determine local image features at step 124 and global image features at step 126. This is discussed with reference to Figure 6, below. In the present example, the local features derived comprise 128 features (or dimensions) and the global features derived comprise 13 features (or dimensions). At step 128, the local features and global features are merged to form a 141-dimensional
20 feature vector. The 141-dimensional feature vector is examined by a statistical model, in the present example a supervised vector model (SVM) such as the C-SVC (C-support vector classification) model. The SVM model will be trained as will be described below in detail.

25 The SVM model determines at step 132 whether or not the candidate boundary video frame comprises an FMPI frame; that is, it determines whether the candidate video frame which is associated with the candidate commercial boundary of the segmented video broadcast comprises product information. If the query returns a positive result (i.e. the candidate boundary video frame is an FMPI frame), an FMPI
30 confidence/likelihood score for the or each frame in a candidate window (the candidate window comprising a set of video frames associated with the candidate commercial boundary) at step 134. The confidence/likelihood score may be a probability value, as

discussed below. The candidate boundary likelihood assessment is then made at step 110 of Figure 3a; that is, the apparatus determines a likelihood the candidate commercial boundary is a commercial boundary in dependence of the determination the candidate video frame comprises product information.

5

As mentioned above, the assessment of the likelihood the candidate commercial boundary is a commercial boundary at step 110 may be augmented by ASCI (audio scene change indicator) recognition in step 114 of Figure 3a. A process for assessing the audio scene change is illustrated in more detail in Figure 3c. The candidate
10 boundary is detected at step 106 of Figure 3a. At step 140, a symmetric audio window is defined at step 140. This will be described further below. At step 142, the symmetric window is segmented into frames, and a sliding window is derived. Again, this will be described further below. At step 146, audio features are extracted for each sliding window in the segmented window. At step 148, the K-L (Kullback-Leibler) distance
15 metric is applied to the extracted audio features and alignment of the audio window takes place at step 150, looping back to step 148, again as described in detail below. At steps 152 and 154, ASC and non-ASC HMM-trained models analyse the extracted audio features and probability scores for ASC and non-ASC are derived at steps 156 and 158 respectively. The probability scores will be described further below and are applied
20 to the candidate boundary likelihood assessment at step 110 of Figure 3a.

A second system architecture and flow diagram is illustrated in Figure 4.

An input video stream is first partitioned into commercial segments and programme
25 segments. Shot change detection is applied to detect cuts, dissolves, and fade in/fade out, which are considered as candidate commercial boundaries. Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) are employed to construct mid-level features labelled "Audio Scene Change Indicator" (ASCI) and "Image Frame Marked with Product Information" (FMPI) to alleviate the problem of dimensionality and
30 incorporate domain knowledge towards an effective solution. Thresholding is used to detect Silence and Black Frames that constitute an integrated feature set together with ASCI and FMPI. Finally, a supervised learning algorithm is utilised to fuse ASCI,

FMPI, Silence, and Black Frames to distinguish true boundaries of an individual TV commercial. Derivation of these model and features are described below.

An SVM is utilised to accomplish the binary classification problem of an FMPI frame.

- 5 This may be a simple binary (“Yes”/”No” classification). Compared with artificial neural networks, SVMs are faster, more interpretable, and deterministic. Advantages of SVMs over other methods consist of a) providing better prediction on unseen test data, b) providing a unique optimal solution for a training problem, c) containing fewer parameters compared to other methods, and d) working well for data with a large
- 10 number of features. It has been found that the C-Support Vector Classification (C-SVC) works particularly well with the described techniques. The radial basis function (RBF) kernel is used to map training vectors into a high-dimensional feature space for classification.

- 15 We turn now to a discussion of video shot detection as mentioned above,

The term scene transition detection (STD) is used to differentiate commonly known scene change detection that aims to detect shot boundaries by visual primitives.

- Generally, a scene or a story unit is composed of a series of “interrelated shots that are
- 20 unified by location or dramatic incident” [9]. STD aims to detect scenes on the basis of computable audio-visual characteristics and production rules. Many prior works deal with STD concentrating on sitcoms, movies [5] – [9], or broadcast news video [10] [11].

- Instead of a single shot, a scene is often treated as an elementary and meaningful unit
- 25 for effective browsing, navigation, and search in video programs where rough scene boundaries suffice for organising video content. Rather than exactly locating scene boundaries, most works deal with STD via the aggregation of consecutive shots.

- Clearly, a scene lies in the marriage of video structure and semantics. By investigating the temporal consistency of audio-visual contents, only an approximation for the actual
- 30 scene should be expected. That is, exact scene boundaries cannot be secured. In particular, commercial videos are featured by dramatic changes in lighting, chromatic composition, and tempo (determined by shot length, motion, zoom, sound, etc.)

amongst shots, and by creative stories. It makes existing STD methods less effective for ComBD, as video shots lack uniform agglomeration within a commercial.

One exemplary approach described herein reduces the problem of commercial STD to that of a classification of True Scene Changes versus False Scene Changes at candidate positions consisting of video shot change points. It is reasonably assumed that a TV commercial scene transition always comes with a shot change (i.e., cuts, fade-in/-out, and dissolves). Features (e.g. multi-modal features) are extracted within a symmetric window at each candidate point. Different or multi-scale window sizes may be optionally applied to different kinds of features. A supervised learning is subsequently applied to fuse multi-modal features. Particularly, the two concepts of ASCI and FMPI characterise computational video contents (structural or semantic) of interest to signify the boundaries of an individual commercial. As noted above, it is infeasible to decipher a commercial video's temporal arrangement via a predefined set of shot classes. The role of mid-level features is to condense high dimensional low-level features by using adequate classifiers to generate as many useful concepts as possible that are supported by commercial video production rules or knowledge. The framework is illustrated in Figure 4.

Accurate cuts and fade-in/-out is significant in the described techniques. In one system performance, videos are in MPEG-1 format, and the compressed domain approach in [12] is employed to determine cuts. In terms of parameter tuning, a higher recall of cuts is preferable. Fade-in/-out is determined by detecting monochrome frames and detecting gradual transitions simply via the twin comparison method (TCM), as the fade-in/-out between two successive spots is in a short duration (often less than 8 frames) and TCM can work well for short gradual transition [17]. In system performance assessments, detected cuts and fade-in/-out have covered about 98% true individuals' boundaries.

FMPI and ASCI are two mid-level features on the basis of video and audio content within an individual TV commercial. Silence and Black Frames are based on the post-editing of a sequence of TV commercials. FMPI – whether or not in combination with ASC – provides a post-editing independent system and method. The combination of

FMPI, ASCI, and as a further option with Silence and Black Frame provides a more reliable system and method if Silence and Black Frame are used in post-editing process. (Silence and Black Frames are created in post-editing processes.) Further, as different countries make use of them differently, it is a significant advantage of the disclosed techniques for FMPI and, optionally, ASCI not to depend on these features.

FMPI is used to describe those images containing visual information explicitly illustrating an advertised product or service. The visual information is expressed in the combination of three ways: text, computer graphics, and frames from a live footage of real things and people. Figure 5 illustrates some examples of FMPI frames. The textual section may consist of the brand name, the store name, the address, the telephone number, and the cost, etc. Alongside the textual section, a drawing or photo of a product might be placed with computer graphics techniques. As graphics create a more or less abstract, symbolic, or “unreal” universe in which incredible things can happen (from a viewer’s perspective), live footage of real things or people is usually combined with computer graphics to solve the problem of impersonality. Each frame of film can be layered with any number of superimposed images.

Let us investigate those examples. Figure 5 (a)–(e) are the simplest yet most prevalent ones. For Figure 5 (f)–(j), the product is projected into the foreground, usually in crisp, clear magnification. For Figure 5 (k)–(o), the FMPI frames are yielded by the superimposed text bars, graphics, and live footage. From the image recognition point of view, Figure 5 (a)–(e) produce a fairly uniform pattern; for Figure 5 (f)–(j), the pattern variability mainly derives from the layout and the appearance of a product; Figure 5 (k)–(o) present more diverse patterns due to unexpected real things.

The spatial relationship between the FMPI frames and an individual commercial’s boundaries is revealed by the production rules as below. For the convenience of description, we define the shot containing at least one FMPI frame as an FMPI frame. Firstly, in most TV commercial videos, one or two FMPI frames are utilised to highlight the offer at the end of a commercial. A good example is a commercial for services, expensive consumer durables, and big companies. These commercials usually work

through context or setting plus the technical sophistication of the photograph or camera work to concentrate on the presentation of luxury and status, or to explore subconscious feelings and subtle associations between product and situation. For these cases, it is sometimes hard to see what precisely is on offer in commercials since the product or service is buried in the combination of commentary and video shots. Accordingly, an FMPI frame is a useful 'prop'. Some modest encouragement coming from the announcer/voice-over, together with one or more consecutive FMPI frames, makes the finishing point. Secondly, an FMPI frame might be irregularly interposed in the course of some TV commercials (say, 30-second or 60-second), as our memories are served by of course endless repetition besides brand names, slogans and catchphrases, and snatches of song. Occasionally an FMPI frame may be present at the beginning of a commercial.

Therefore, an FMPI frame can be considered as an indicator, which helps to determine a much smaller set of commercial boundary candidates from large amounts of shot transitions. It is possible to rely on the FMPI frames only to identify commercial boundaries, but performance may feature a higher recall but a lower precision. As illustrated in Figure 4, and particularly by Figure 15 below, by combining FMPI and ASCI techniques, this problem can be alleviated and yet more accurate results may be obtained.

Figure 6 shows an FMPI frame represented by properties of colour, texture, and edge features. As the layout is a significant factor in distinguishing an FMPI frame, it is beneficial to incorporate spatial information about visual features. One common approach is to divide images into subregions and impose positional constraints on the image comparison (image partitioning). This approach is used to train the SVM and also to determine whether the candidate video frame comprises FMPI. In terms of colour, dominant colours are used to construct an approximate representation of colour distribution. These dominant colours can be easily identified from colour histograms. Since Gabor filters exhibit optimal localisation properties in the spatial domain as well as in the frequency domain, they are used to capture rich texture information in the FMPI frame. Edge is a useful complement of textures especially when an FMPI frame

features stand-alone edges as a contour of an object, as texture relies on a collection of similar edges.

As shown in Figure 6, the boundary classifier derives the training data by parsing video frames comprising product information and extracting a video frame feature for one or more portions of the video frame and/or for a complete video frame. A given image is first sub-divided into 4×4 sub-images, and local features of eight dimensions for each of these sub-images are computed. The LUV colour space is used to manipulate colour. A uniform quantisation of the LUV space to 300 bins is employed, each channel being assigned 100 bins. Three maximum bin values are selected as features from L, U, and V channels, respectively, as indicated by solid bars in Figure 6. Edges derived from an image using Canny algorithm provide an accumulation of edge pixels for each sub-image, which finally acts as 16-dimensional edge density features. A set of two-dimensional Gabor filters is employed to extract texture features. The Gabor filter is characterised by a preferred orientation and a preferred spatial frequency. The filter bank comprises 4 Gabor filters that are the results of using one centre frequency (i.e., one scale) and four different equidistant orientations. The application of such a filter bank to an input image results in a 4-dimensional feature vector (consisting of the magnitudes of the transform coefficients) for each point of that image. The mean of the feature vectors is calculated for each sub-image. A 128-dimensional feature vector is then formed to represent local features.

The cues of colour and edge are taken into account for global features. Three maximum bin values are selected from each colour channel, which results in a 9-dimensional colour feature vector for a whole image. Edges are grouped into four categories: horizontal, 45° diagonal, vertical, and 135° diagonal. Edge pixels are accumulated for each category thus yield 4-dimensional edge density features. Finally, a 141-dimensional low-level visual feature vector comprising 128-dimensional local features and 13-dimensional global features is constructed.

Alternatively, let Γ be an $n \times m$ image. *LUV* colour space is used. The colours in Γ are uniformly quantised into $3 \cdot l$ bins, each channel being assigned l bins. To extract local features, Γ is partitioned into $r \cdot c$ sub-images equally. Within each sub-image, the first p maximum bin values are selected as dominant colour features. Note that the bin values are meant to represent the spatial coherency of colour, irrespective of concrete colour values. Based on Canny edges, edge pixels are summed up within each sub-image, thereby yielding edge density features with $r \cdot c$ dimensions. A set of two-dimensional Gabor filters are employed for texture. Within each sub-image, the mean μ_{sk} of the magnitudes of transform coefficients is used. For S scales and K orientations, texture features of $r \cdot c \cdot S \cdot K$ dimensions are finally constructed using μ_{sk} . In terms of global features, colour and edge are taken into account. Similarly the first q maximum bin values are selected from each channel. Edges are broadly grouped into h categories of orientation by using the angle quantiser as:

$$A_i = \left[\left\lfloor \frac{180}{h} \right\rfloor \cdot i, \left\lfloor \frac{180}{h} \right\rfloor \cdot (i+1) \right), i = 0 \dots h-1.$$

By combining local features and global ones, we obtain the feature of $(3 \cdot p \cdot r \cdot c + r \cdot c + S \cdot K \cdot r \cdot c + 3 \cdot q \cdot h)$ dimensions. It has been found that the following parameter settings yield acceptable results: $r = c = 4$, $p = 1$, $q = 3$, $S = 1$, $K = 4$, and $h = 4$. The utilised Gabor filters use one centre frequency (one scale) and four equidistant orientations. Finally, 141-dimensional feature vector is constructed (128 local features and 13 global ones).

Figure 7 shows a set of recall/precision curves yielded by using different visual features and different C-SVC parameters, which have shown the effectiveness of the proposed method in distinguishing the FMPI frames from an extensive set of commercial images.. An accuracy of $F_1 = 89.6\%$ is achieved by using 4632 images comprising 1046 FMPI frames and 2987 Non-FMPI frames selected from the TV commercial video database consisting of 499 clips of individual TV commercial videos covering 390 different commercials. This accuracy is calculated by averaging the results of ten runs, namely, ten different random half-and-half training/testing partitions.

LIBSVM [16] is utilized to accomplish C-SVC learning. Radial basis function (RBF), $\exp(-\gamma \|x_i - x_j\|^2)$, $\gamma > 0$, is used. We are required to tune four parameters, i.e., γ , penalty C , class weight w_i , and tolerance e . w_i is for weighted SVMs to deal with unbalanced data, which set the cost C of class i to $w_i \times C$. e sets the tolerance of

5 termination criterion. Class weights are set as $w_1 = 5$ for FMPI class and $w_0 = 1$ for Non-FMPI class. e is set to 0.0001. γ is tuned between 0.1 and 10 while C is tuned between 0.1 and 1. An optimal pair of $(\gamma, C) = (0.6, 0.7)$ is set.

In order to evaluate the effects of low-level visual features on the performance, a set of

10 recall/precision curves are yielded by using different visual features and different pairs of (γ, C) as shown in Figure 7.

For each kind of feature combination discussed below, classification is performed with different SVMs kernel parameters (or combinations thereof): colour, texture, edge. Note

15 that different parameters generate different performance figures for recall and precision. Different recall/precision values for each kind of features combination are linked to generate the curves like Figure 7 to reveal some tendency.

In use of LIBSVM as the implementation of the SVM, a set of manually-labelled

20 training feature vectors for FMPI frames and Non-FMPI frames are fed into a LIBSVM to train the SVM classifier in a supervised manner. For an incoming image, the apparatus extracts the feature vector from the image and feeds the feature vector into the trained SVM classifier. The SVM classifier then determines whether the image associated with the feature vector is an FMPI frame or not. Given a set of test images,

25 the SVM correctly classifies some images as FMPI frames and incorrectly classifies some images as FMPI frames. The classification results vary with different SVM kernel parameters. Examples of performance are illustrated in the recall/precision curves of Figure 7.

Two performance curves of “Colour” and “Texture” have demonstrated the individual capability of colour and texture features to distinguish FMPI from Non-FMPI. Texture features play a more important role comparatively. The combination of colour and texture features results in a significant improvement of performance. Edge alone, whilst
5 useful, is less effective than Texture alone. However, the performance can be further improved more or less by fusing Colour, Texture, and Edge together. Edge is a useful complement of textures especially when an FMPI frame exhibits stand-alone edges as a contour of an object, since a collection of similar edges forms texture. In terms of colour a reduced dominant colour descriptor is actually utilised. As shown in figure 6,
10 one maximum bin for each channel within sub-images and three maximum bins for each channel within the whole images are considered. The percentages of selected bins are taken into account to represent spatial coherency of colour. Colour value is less useful for representing an FMPI frame.

15 Aiming to determine an FMPI shot, the FMPI recognition may be applied to those key frames selected from a shot to identify a candidate video frame from a motion measurement of a video frame associated with the candidate commercial boundary. Motion is utilised to identify key frames. The average intensity of motion vectors in the video frame from B- and P- frames in MPEG videos is used to measure the motion in a
20 shot and select key frames at the local minima of motion. Directing recognition at key frames has two advantages: 1) reducing computation, and 2) avoiding distracting frames due to animation effects.

Although perfectly acceptable results may be achieved by FMPI recognition alone, such
25 an implementation may be improved further. Figure 8 illustrates a group of images incorrectly classified as an FMPI frame from FMPI recognition alone. As indicated in Figure 8 (a) (b) (c) (e), strong texture of a large area is one main reason for false alarms. Figure 8 (d) observes clear edges overlapping a large blank area, which exhibits a similar pattern as an FMPI frame. A typical example is given in Figure 8(f) where an
30 object is delineated at the centre of a blank frame. Such a kind of picture often appears in an FMPI frame to highlight the foreground product. To avoid such false alarms, an

algorithm is required to understand what to tell in an image frame. Clearly it is difficult due to the semantic gap between low-level visual features and high-level concepts.

As noted above with respect to Figures 3a and 3c, it is possible to introduce into the
5 assessment a likelihood (of probability) score that the candidate commercial boundary comprises an audio scene change. This is now discussed in detail.

The most common type of TV commercial is a combination of continuous music, sound effects, voice-over narration, and storytelling video. It is easy to imagine different TV
10 commercials exhibit dissimilar audio characteristics. A proper modelling of audio scene changes (ASC) can facilitate the identification of commercial boundaries.

An audio scene is often modelled as a collection of sound sources and the scene is further assumed to be dominated by a few of these sources [5]. ASC is said to occur
15 when the majority of the dominant sources in sound change. It is more or less complicated and sensitive to determine the ASC transition pattern in terms of acoustic classes [13] because of model-based methods' weakness: large amounts of samples required and the subjectivity of classes labelling. An alternative is to examine the distance metric between two windows based on audio features. Metric-based methods
20 are straightforward. A quantitative indicator is produced. Yet human knowledge is not incorporated by labelling training data or others.

The boundary classifier may make the determination the candidate video frame comprises an audio scene change from a distance measurement of audio properties of
25 first and second audio frames of an audio segment of the video broadcast associated with the candidate commercial boundary. Figure 9 shows an audio segment located within a symmetric window at each video shot transition point. The window may be of a pre-defined length. A HMM is utilized to train two models for representing Audio Scene Change (ASC) and Non-audio Scene Change (Non-ASC) on the basis of low-
30 level audio features extracted from the audio segment. Given a candidate commercial boundary, two probability values output from trained HMM models are combined with

FMPI related feature values, Silence and Black Frame related feature values to represent a commercial boundary as illustrated in Figure 4.

In previous work, an audio scene is usually modelled as a collection of sound sources and the scene is further assumed to be dominated by a few of these sources. ASC is said to occur when the majority of the dominant sources in the sound change. In terms of acoustic sources, previous work has classified the audio track into pure speech, pure music, song, silence, speech with music background, environmental sound with music background, etc. ASC is accordingly associated with the transition among major sound categories or different kinds of sounds in the same major category (e.g. speaker change). Although good results have been achieved on audio classification, it is complicated and sensitive to determine the ASC's transition pattern in terms of acoustic classes in audio streams. It is due to the model-based method's weakness: large amounts of samples required and the subjectivity of their audio classes labelling. An alternative is to use a metric-based approach that examines the audio features' distance measure between neighbouring windows. The metric-based method is straightforward. A quantitative indicator is produced. Yet human knowledge cannot be incorporated by labelling training data.

Given an audio segment of a predefined length (say 4 seconds) around a candidate boundary, the proposed ASCI is meant to provide a probabilistic representation of ASC. As shown in Figure 9, HMM is utilized to train two models for "explaining" the audio dynamic patterns, namely, ASC and Non-ASC. An unknown audio segment is classified by either of the models which returns the highest posterior probability the segment is ASC or non-ASC. This model-based method is different from that based on acoustic classes. Firstly, the labelling of ASC/Non-ASC is simpler and can more or less capture the sense of hearing when one is viewing TV commercial videos. Secondly, according to the framework in Figure 1, two probability values yielded by the ASCI, as intermediate features, can be easily fused with others; in other words, the subjectivity at this stage would not seriously affect the final target. Moreover, a metric-based method is introduced to accomplish the alignment of audio feature windows.

Our simulation results have shown that the combination of model-based and metric-based algorithms yields better performance.

A mixture Gaussian HMM (left-to-right) is utilised to train ASC/Non-ASC recognisers.

5 A diagonal covariance matrix is used to estimate the mixture Gaussian distribution.

Suppose we have two HMM models for representing ASC and Non-ASC, two likelihood values of an observation sequence are generated by the forward-backward algorithm. An HTK toolkit [14] is utilised.

10 The ASCI considers 43-dimensional audio features comprising Mel-frequency cepstral coefficients (MFCCs) and its first and second derivatives (36 features), mean and variance of short time energy log measure (STE) (2 features), mean and variance of short-time zero-crossing rate (ZCR) (2 features), short-time fundamental frequency (or Pitch) (1 feature), mean of the spectrum flux (SF) (1 feature), and harmonic degree (HD) (1
15 feature). An audio signal is segmented into a series of successive 20 ms analysis frames by shifting the sliding window of 20 ms with an interval of 10 ms. Features are computed for each analysis frame. Within each frame STE, ZCR, SF, and Harmonic peaks are computed once every 50 samples at an input sampling rate of 22, 050 samples per sec where the duration of sliding window is set to 100 samples. Means and
20 variances of STE and ZCR are calculated for 7 values from 7 overlapping frames while the mean of SF is calculated for 6 values from 7 neighbour frames. HD is the ratio of the number of frames having harmonic peaks to the frame number 7. Pitch and MFCCs are computed directly from each frame. The non-ASC may also consider the same parameters.

25

Major reasons for using these features are as follows. MFCCs furnish a more efficient representation of speech spectra, which is widely used in speech recognition. STE provides a basis for discriminating between voiced speech components and unvoiced speech components, speech and music, audible sounds and silence. In terms of ZCR,
30 music produces much lower variances and amplitudes than speech does. ZCR is also useful for distinguishing environmental sounds. Pitch determines the harmonic property of audio signals. Voiced speech components are harmonic while unvoiced speech

components are non-harmonic. Sounds from most musical instruments are harmonic while most environmental sounds are non-harmonic. In general, the SF values of speech are higher than those of music but less than those of environmental sounds.

- 5 As illustrated in Figure 9, the alignment of audio feature window is incorporated. The alignment problem is addressed principally for two reasons. Firstly, at most TV commercial boundaries there is an offset of ± 0.25 sec $\sim \pm 1.0$ sec between an audio scene change and its associated video scene change. Secondly, due to video production, a mixed soundtrack made up of music, commentary, and sound effects does not
- 10 necessarily synchronise a video track; therefore a symmetric window at shot transitions cannot secure the extraction of effective features well matching the sense about ASC nearby. This is supported by the statistics of time offsets in news programme and commercials as shown in Figure 10. Based on experimental observations, around 95% offsets lie in the range of ± 0.25 sec $\sim \pm 1.0$ sec wherein the offsets of ± 0.25 sec occupy
- 15 around 85%. In the news video production, an anchor person or live reporter can hardly remain synchronised when camera shots are switched to weave news stories. In order to capture audience attention, the production of commercial video tends to use more editing effects. The time offsets are mainly attributed to post-editing effects. For example, for fade -in/-out, the visual change is located at the middle of the shot
- 20 transition whereas the audio change point is often delayed till the end of the shot transition.

- Figure 11 shows the Kullback-Leibler distance metric in use to evaluate the changes between successive audio analysis windows and to align the audio window. Window
- 25 size is important to good modelling. The difference curves indicate the different locations of peak change for different window sizes. A multiscale difference computing is used since it is unknown what sounds are being analysed. Essentially, the boundary classifier determines the candidate video frame comprises an audio scene change by partitioning the audio segment into a plurality of sets of audio frames, each set of audio
- 30 frames having frames of equal length, the length of one set of audio frames being different from a length of another set of audio frames to determine a set of difference sequences of audio properties from the sets of audio frames, and determining a

- correlation between difference sequences of the set of difference sequences. Different window sizes are first used to yield a set of difference sequences; each difference sequence is then normalized to [0, 1] through dividing difference values by the maximum of each sequence; the most likely audio scene change is determined by
- 5 locating the highest accumulated difference values derived from the set of difference sequences. A set of uniform difference peaks associated with the true audio scene change has been located with around 240 ms delay; the offset is identified from a correlation between difference sequences. After identifying the offset, the boundary classifier aligns the audio scene change with the candidate commercial boundary.
- 10 According to offset statistics in Figure 10, the shift of adjusted change point is currently confined to the range of [-500ms, 500ms]. Audio features are further extracted and arranged within adjusted 4-sec feature windows to be fed into two HMM-based classifiers associated with ASC and Non-ASC, respectively. That is, boundary classifier extracts audio features from the audio segment, trains first and second statistical models
- 15 for audio scene change and for non-audio scene change from the audio features extracted from the audio segment and classifies a candidate audio segment associated with the candidate commercial boundary from the first and second statistical models.

An alignment procedure seeks to locate the most likely ASC point within the

20 neighbourhood of a shot change as illustrated in Figure 5. Let W_i and W_j be two audio analysis windows, and their difference denoted by $d(W_i, W_j)$. By utilising Kullback-Leibler (K-L) distance metric [15], the difference can be written as

$$d(W_i, W_j) = \int_x [p_i(x) - p_j(x)] \ln p_i(x) / p_j(x) dx$$

where $p_i(x)$ and $p_j(x)$ denote the probability distribution functions (*pdf*) estimated by

25 the features extracted from W_i and W_j . One-scale is considered firstly. Let $W_i, i = 1, 2, \dots, N$ be a series of analysis windows with an overlap of INT ms. We then form the sequence $\{D_i\}_{i=1,2,\dots,N-1}$ as $D_i = d(W_i, W_{i+1})$. An ASC from W_l to W_{l+1} is declared if D_l is the maximum within a symmetric window of WS ms. Window size is important for good modeling. The difference curves in Figure 11 indicate different

change peaks in the case of different window sizes. Since one does not know a priori what sound one is analysing, multi-scale computing is used. The K-L metric first makes use of multiple window sizes $\{Win_{scale}\}_{scale=1,\dots,S}$ to yield a cluster of difference value series which is denoted by

$$5 \quad \{Distance_{scale}\}_{scale=1,\dots,S} = \{D_{i,scale} | i = 1, \dots, N_{scale}\}_{scale=1,\dots,S}; \text{ each series of}$$

$Distance_{scale}$ is then normalised to $[0,1]$ through dividing difference values $D_{i,scale}$ by the maximum of each series $Max(Distance_{scale})$; the most likely ASC point ω is finally determined by locating the highest accumulated values.

- 10 The probability $p(\omega\lambda)$ of the candidate window position $\omega\lambda$ being an ASC point is calculated as:

$$p(\omega\lambda) = \frac{1}{S} \sum_{scale=1}^S \left(\frac{Distance_{scale}(\lambda)}{Max(Distance_{scale}(\lambda))} \right)$$

$$\omega = \arg \max_{\lambda} (p(\omega\lambda)), \lambda = 1, \dots, M$$

- where M denotes the total number of candidate window positions, ω denotes the
15 window corresponding to an ASC point.

- Based on offset statistics, the shift of adjusted change point is confined to the range of $[-500\text{ms}, 500\text{ms}]$, i.e., $WS = 1000$. Audio features are extracted and arranged within the adjusted 4-second feature windows. 11 Scales are employed, i.e., $S = 11$, where the
20 window sizes $Win_{i=1,\dots,11} = 500 + 100 \cdot (i - 1)$ ms. At all scales, the overlap interval is set to $INT = 100$ ms. A single Gaussian *pdf* is used. 20 ms sliding window with an interval of 10 ms is applied.

- The Kullback-Leibler distance metric is a formal measure of the differences between
25 two density functions. The normal density function is currently employed to estimate the probability distribution of 43-dimensional audio features for each sliding analysis window. For the minimum window of 500 ms in total 499 samples of 20 ms unit with a

10 ms overlap result. Figure 11 shows at the sliding window level an overlap of 100 ms has been uniformly employed for multi-scale computing.

Figure 12 shows the Kullback-Leibler distances of a small set of ASC and Non-ASC samples which are illustrated to indicate the effectiveness of low-level audio features. The duration of each audio sample is 2 seconds. Two probability distributions are computed for two symmetric windows of one second. The same sampling strategy is applied, i.e., 20 ms unit with a 10 ms overlap. The audio samples are selected to cover diverse audio classes such as speech, different kinds of music, speech with music background, speech with noise background, etc. Two clusters of Kullback-Leibler distances can be delineated clearly. This indicates selected low-level audio features' capability in discriminating ASC samples from Non-ASC samples.

Although the Kullback-Leibler distance metric can explicitly provide a quantitative measure of audio signal change, the temporal context is not utilized unlike an HMM-based modelling. HMM is a powerful model to characterize the temporally non-stationary but learnable and regular patterns for the speech signal especially when utilised in conjunction with the Kullback-Liebler distance metric. As shown in Figure 13, a performance comparison is illustrated between a Kullback-Leibler based approach (with or without actual alignment of the offset with the K-L metric) and an HMM-based approach (with or without the K-L based alignment). The audio data set comprises 2394 Non-ASC samples and 1932 ASC samples. A Half-and-Half training/testing partition is applied. A left-to-right HMM consisting of 8 hidden states is employed. A diagonal covariance matrix is used to estimate the mixture Gaussian distribution comprising 12 components. The forward-backward algorithm generates two likelihood values of an observation sequence.

If the audio scene change is not aligned with the candidate commercial boundary, the probability/likelihood scores for each of these can be fused later to provide what may be acceptable results. However, by performing the early fusion, the co-occurrence of some features an effectively indicate the boundary and performance may be improved.

With an alignment process, F1 or overall accuracy is increased by 3.9% - 4.6%. Against the Kullback-Leibler based approach alone, the HMM-based method improves the F1 or overall accuracy by 2.9% - 4.2%. Comparatively, the alignment plays a more important role in performance improvement. An emphasis should be put on the overall accuracy of ASC and Non-ASC, since two generated probabilities for ASC and Non-ASC jointly contribute to the boundary classification. According to simulation results, a promising accuracy of 87.9% has been achieved by HMM with an alignment process.

Silence is detected by examining the audio energy level. The short-time energy function is measured every 10 ms and smoothed using an 8-frame FIR filter. The smoothing implicitly imposes a minimum length constraint on the silence period. A threshold is applied, and the segment that has its energy below the threshold is decided as Silence. A black frame is detected by evaluating the mean and the variance of intensity values for a frame. A threshold method is applied. A series of consecutive black frames (say 8) is considered to indicate the presence of Black Frames.

The use of Silence & Black Frames is limited by editing techniques at TV commercial boundaries and their frequent occurrences within an individual commercial. However, Silence and Black Frames can be combined with FMPI and ASCI to form a complete feature set useful for detecting TV commercial boundaries.

As shown in Figure 4, when fusion of FMPI, ASCI, Silence, and Black Frames is implemented, this is accomplished by a supervised learning algorithm. In the example of Figure 4, a binary classification is carried out. An SVM-based classifier is used in an assessment of system performance.

The boundary classifier classifies the candidate commercial boundary as a commercial boundary from a fusion of likelihood scores for frame marked with product information (FMPI), audio scene change (ASC) and, optionally, audio silence and video black frame. ASCI yields two probability values $p(\text{ASC})$ and $p(\text{Non - ASC})$, Silence and Black Frames yields two values $p(\text{Silence})$ and $p(\text{Black Frames})$ to indicate the

presence of Silence and Black Frames, respectively. In terms of FMPI, $2 \cdot n$ video shots within the symmetric neighbourhood of a candidate boundary (Left n shots, Right n shots) produce $2 \cdot n$ values $\{p_i(\text{FMPI})\}_{i=1, \dots, 2n}$ to indicate the presence of FMPI shots. In this instance,⁶ the candidate video frame comprises a frame of a plurality of video frames of a candidate commercial window associated with the candidate commercial boundary, and the boundary classifier determines a commercial boundary probability score for video frames of the candidate commercial window determines the likelihood the candidate commercial boundary is a commercial boundary from a plurality of the commercial boundary probability scores. An overall likelihood score is derived from one or more of the probability scores. Hence the complete feature is $2 \cdot n + 4$ dimensional. In the performance assessment, we set $n = 2$.

Machine learning is used to complete the fusion of the probability scores because it is not a trivial task to construct manually the heuristic rules to fuse the probabilities. With these probabilities as the feature vector, a SVM is used to learn the patterns associated with (true) commercial boundaries or false commercial boundaries in terms of those probabilities, from a series of manually labelled true or false boundary examples. The fusion can be linear or non-linear. In some apparatuses, the boundary detection problem is transformed into a binary classification problem.

A commercial video database is built for assessment, which consists of 499 clips of individual TV commercial videos covering 390 different commercials. The TV commercial video clips come from heterogeneous video data set of 169 hours of news video taken from 6 different sources, namely, LBC, CCTV4, NTDTV, CNN, NBC, and MSNBC. These commercials have extensively covered three concepts: namely, Ideas (e.g. education opportunities, vehicle safety), Products (e.g. vehicles, food items, decoration, cigarettes, perfume, soft drink, health and beauty aids), and Services (e.g. banking, insurance, training, travel and tourism).

Figure 14 shows the statistics in terms of the number of video shots and the duration within a single TV commercial clip. Three major modes about the duration are

observed to be roughly located at 15 seconds, 30 seconds, and 60 seconds. The 30 seconds mode is often used and claims to cut costs as well as gaining reach. The 60 seconds mode is considered as a media idea featuring the substance, tone, and humour of a creative idea. The 15 seconds mode is the saviour of the single-minded idea. The number of video shots features a larger variance. This may be related to various types (e.g. Problem-Solution Format, Demonstration Format, Product Alone Format, Spokesperson Format, Testimonial Format, etc.) of TV commercials.

Figure 15 shows performance results of an individual TV commercial's boundaries detection. Using different features and different SVM parameters yields a set of recall-precision curves. A promising accuracy of $F1 = 89.22\%$ have been achieved through half-and-half training/testing on the basis of the fusion of FMPI and ASCI only. This performance has provided a basis of a reliable system and method for detecting boundaries, since FMPI and ASCI are completely intra-commercial content-based and independent of post-editing techniques. A further improvement of performance from $F1 = 89.22\%$ to $F1 = 93.7\%$ is obtained by fusing FMPI, ASCI, SILENCE, and BLACK FRAMES. Comparatively, using traditional BLACK FRMAES yields a poor result of $F1 = 81.0\%$ (Recall = 87.0%, Precision = 75.8%).

The performance of "FMPI+ASCI+SILENCE+BLACK FRAMES" may vary with different video data streams due to non-uniform post-editing techniques. However, in present simulation experiments, a heterogeneous video data set has been employed aiming at a fair performance evaluation.

The apparatus of Figure 2, comprising separable boundary and commercial classifiers can be considered as an apparatus for identifying a boundary of a commercial broadcast in a video broadcast and classifying the commercial broadcast in a pre-defined category. The apparatus comprises a video shot transition detector configured to identify a candidate commercial boundary in the video broadcast, a boundary classifier configured to verify the candidate commercial boundary as a commercial boundary and a commercial classifier configured to classify the commercial in a pre-defined category.

- A commercial classifier apparatus for classifying a commercial video broadcast in a pre-defined category will now be described. The commercial classifier may be used in conjunction with the apparatus for determining a likelihood that a candidate commercial boundary of a commercial broadcast in a segmented video broadcast is a commercial boundary described above. Use of the two apparatuses together (as illustrated in, say, Figure 2) may be particularly advantageous; if a candidate commercial boundary can be determined to be a commercial boundary with any level of certainty, this facilitates identification of a commercial broadcast for its classification.
- 10 The architecture of classifier 68 is shown in more detail in Figure 16. The commercial classifier 68 comprises, optionally, a video processor 200 for extracting video and/or audio data from a frame of the video broadcast commercial and converting the video and/or audio data to text data, a classifier model 202, and a proxy document identifier 204 for identifying a proxy document as a proxy of the commercial video broadcast.
- 15 The proxy document identifier may identify the proxy document as a document related to a keyword identified by First keyword derivation module 206. first text pre-processing module 208, test word vector mapper 210 and training module 212. Training module 212 is for the compilation of training data from a corpus of training documents and may comprise second keyword derivation module 214, second text pre-processing module 216 and training data vector mapper 218. The classifier module is trained by data from the training data, and classifies the commercial video broadcast from an examination of proxy data from the proxy document.
- 20 The classifier module may be a support vector machine module.
- 25 The proxy document identifier 204 is configured to interface with a document index/database 220 which may be a remote external resource, as shown in Figure 16, from commercial classifier 68.
- 30 A process flow of a first commercial classifier 68 is described as follows with respect to Figure 17. The classification process starts at step 230 and, at step 232, video processor 200 parses a commercial video broadcast for video and/or audio data. At step 234,

proxy document identifier 204 identifies a proxy document from the video/audio data. As described below, this may be done by converting the video/audio data to text data and identifying the proxy document from the text data with ASR and OCR modules of the video processor. At step 236, the classifier model 202 is trained with training data
5 from training module 212. At step 238, the classifier model 202 classifies a commercial broadcast from an examination of proxy data from the proxy document identified by proxy document identifier 204.

The architecture and process flow for a second commercial classifier is described with
10 respect to Figure 18. Where appropriate, like reference numerals denote like parts when compared with Figure 16.

The Commercial Video Processing Module (COVM) 200 aims to expand the deficient and less-informative transcripts from ASR 252 and OCR 254 with relevant proxy
15 articles searched from the world-wide web (WWW) at step 268 like Google and encyclopaedia webs. For each incoming TV commercial video $TVCOM_i$ 250, the module first converts the video/audio data to text by extracting the raw semantic information via ASR 252 and OCR 254 on the key frame images. Key frames can be extracted at the local minima of motion as described above for FMPI recognition. The
20 accuracy of OCR depends on the resolution of characters in an image. It is empirically observed that text of a larger size contains more significant information than small text. As shown in the right upper image in Figure 22, it is comparatively easy for an OCR module to recognise the text of large size "Free DSL Modem, Free Activation", which contains more category related semantic information than the small and difficult-to-
25 recognise text "after rebates with 12 months commitment". Therefore, the failure of an OCR module in recognising small texts may not necessarily degrade the final performance significantly. It is also the reason why the n nouns and noun phrases with largest font size from OCR are selected to form keywords. Subsequently, spell checking and correction 258 are applied to the transcripts of the ASR and OCR modules
30 by a text-correction module. Any misspelled vocabulary terms are corrected and the terms not found in dictionaries are removed. Both English dictionary and

encyclopaedias are used as the ground truth for spell checking, as a normal English dictionary may not include non-vocabulary terms like brand names. Based on the corrected transcript S_i , the proxy article d_i is obtained. With the word feature space derived from the TRFM module, the testing document vector is generated from d_i .

5

Potential keywords and keyword selection are then made at steps 262, 264. Keyword expansion 268 is made with respect to, for example, the internet 268 and a proxy document assignation step 270 then takes place. Steps 264, 266, and 270 are described in more detail with respect to Figure 19. (Note that the same or similar process may be applied when identifying training keywords by the training data and word feature processing module 212 of Figure 18.)

10

The proposed approach firstly preprocesses the output transcripts of ASR and OCR in TV commercial video $TVCom_i$ with spell checking at step 258 to generate corrected transcript S_i at step 300. A list L_i of nouns and noun phrases from S_i are extracted by a natural language processor at step 302. A set of keywords $K_i(kw_{i1}, \dots, kw_{in})$ are selected by applying the steps below:

15

- a) Check S_i for an occurrence of a brand name or a dictionary of brand names at step 302.
- 20 b) If the result returns that the brand name(s) is are found in S_i at step 306 the brand is selected as a keyword kw_i and searched on the online encyclopedia Wikipedia (<http://en.wikipedia.org/wiki>). The keyword derivation module therefore identifies a keyword by querying the text data for an occurrence of a brand name identifier word; and in dependence of detecting an occurrence of the identifier word, identifying the identifier word as a keyword
- 25 c) If the result returns "No" at step 306, from L_i , another word, such as 1...n nouns and/or noun phrases with largest font size from OCR and the last m from ASR are heuristically selected at step 266 as keywords.

The document identifier identifies the proxy document as a document related to the keyword by querying an external document index or database with the keyword as a query term and assigning a most relevant result document of the query as the proxy document. This is done by searching at step 312 via a document index or database

5 through, for example, a Web Search Engine. The keyword derivation module identifies another word in the text data, for example a noun word, as a keyword. The Google search engine may be utilised at step 312 for its superior performance in assuring the searched articles' relevancy. Among returned articles, the one with the highest relevancy rating is selected at step 270 as the proxy document d_i , by proxy document

10 identifier 204 which we denote as the proxy article of TV commercial $TVCom_i$. By exploiting d_i , TV commercial classification is reduced to the problem of text categorisation. That is to approximate a classifier function $\Phi: D \times C \rightarrow \{T, F\}$ to assign a Boolean value to each pair $(d_i, c_i) \in D \times C$, where D is the domain of proxy article d_i and C is the set of predefined commercial category c_i . A value T assigned to

15 (d_i, c_i) indicates the proxy article d_i under c_i , while a value F assigned to (d_i, c_i) means d_i not under c_i . The values are calculated and assigned to each pair according to a multi-class supervised learning procedure. Given some categories of documents (i.e. different topics), the classifiers are trained based on manually labelled documents. For a testing document, the classifiers can determine whether the document belongs to a

20 category or not. Typically, the value T=1 indicates true, the value F=0 indicates false. Some learning algorithms may generate an output probability ranging from 0 to 1, instead of the absolute values of 1 or 0; thresholding may be applied for a final determination of the category.

25 The first IR Preprocessing Module (IRPM) 208 function at steps 272, 274 is a known vocabulary term normalisation process used in the setting-up of IR systems. It applies two major steps: the Porter Stemming Algorithm (PSA) 276 and the Stop Word Removal Algorithm (SWRA) 278 to rationalise proxy data. PSA is a process of removing the common morphological and inflexional endings from words in English so

that different word forms are all mapped to the same token (which is assumed to have essentially equal meaning for all forms). SWRA is to eliminate words of little or no semantic significance, such as “the”, “you”, “can”, etc. As shown in Figure 18, both testing and training documents go through this module before any other process runs on them.

Once rationalised text has been obtained, the training scripts are digitised and the test word vector mapper 210 forms the test vector at step 282 from proxy data for examination by the classifier model 202 at step 284.

10

The classifier model 202 is trained with training data from the training module 212.

The training module 212 is composed of a Training Data & Word Feature Processing Module (TRFM) which accomplishes two tasks. Firstly, a topic-wise document corpus 286 is constructed from available public IR corpora or related articles manually collected from the WWW 287 as the training dataset of a text categoriser. In this way, the training corpus can possess a large amount of training documents and wide coverage of topics. Such a training corpus can avoid the potential over-fitting problem, which may be caused if the textual information of a limited set of TV commercials only is taken as training data. In a proposed system, the categorised Reuters-21578 and 20 Newsgroup corpora are combined to construct the training dataset. The defined topics of these corpora may not exactly match the categories of TV commercials. One solution is to select the topics from these corpora that are related to a commercial category and combine them to jointly construct the training dataset for representing the commercial category. For example, the documents on the topics of “earn”, “money”, and “trade” in Reuters-21578 are merged together to yield the training dataset for the finance category.

Next the document frequency technique illustrated in greater detail in Figure 20 is employed to perform word feature selection on the training dataset. Document frequency is a technique for vocabulary reduction. Its promising performance and as the computational complexity is approximately linear in the number of training

documents, means it lends itself to the present implementation. The word feature selection process 292 measures the number of documents in which a term w_i occurs, resulting in the document frequency $DF(w_i)$. If $DF(w_i)$ exceeds a predetermined threshold at step 350, w_i is selected as a feature at step 354; otherwise, w_i is discarded and removed from the feature space at step 352. An example of a suitable threshold is 2 when 9107 word features are selected. The basic assumption is that rare terms are either non-informative for category prediction, or not influential in global performance. For each document, the number of occurrences of term w_i is taken as the feature value $tf(w_i)$ at step 356. Finally each document vector is normalised to unit length at steps 294, 296 so as to eliminate the influence of different document lengths.

The Classifier Module (CLAM) performs text categorisation of query articles based on the training corpus and determine the classification of commercial video. There are two principal reasons why SVM is utilised to accomplish the text categorisation task. Firstly, SVM is able to handle high dimensional input space. Text categorisation usually involves feature space with extremely high (around 10,000) dimensions. Moreover, the over-fitting protection in SVM enables it to handle such a large feature space. Secondly, SVM is able to tackle sparse document corpus. Due to the short length of document and large feature space, each document vector contains only a few non-zero entries. As both theoretically and empirically proved, SVM is suitable for problems with dense concepts and sparse instances.

Figure 21 shows the output script of ASR (at step 256 of Fig. 18) on a TV commercial of *Singulair*, which is a brand name of a medicine relieving asthma and allergic symptoms. The script is erroneous and deficient due to background music. By comparing ASR generated script and the actual speech script, it can be found that the innate noise of audio data encumbers the ASR techniques from delivering a semantically meaningful and coherent passage describing the advertised commodity. Any other relevant articles that fall into the same category can serve as the proxy of the TV commercial to serve in the semantic classification task. From the ASR generated scripts, certain noun or noun phrase can be extracted, like <allergy> for example, as

keywords. By searching these keywords in World Wide Web (step 260), an example of the relevant article is acquired which can be assigned as the proxy document. By replacing the ASR output scripts with such articles in text categorisation, it is expected to lead to a more satisfactory commercial classification result.

5

Figure 22 shows another source of potential keywords provided by key image frames of commercial videos. The examples shown present text significantly related to advertised commodity's category, such as <Credit Card> for finance, or even its brand names, such as <Microsoft>.

10

A system uses as an example 499 English TV commercials and extracted 191 distinct ones from TRECVID05 video database. Based on their advertised products or services, the 191 distinct TV commercials are distributed in eight categories, as illustrated in Figure 23. This system involves four categories: Automobile, Finance, Healthcare and IT. Though they do not exclusively cover all TV commercials, they count up to 141 and 74% of total commercials. Therefore, they should be able to demonstrate the effectiveness of the proposed approach. For each category, 1,000 training documents are selected from corpora Reuter and 20 Newsgroup. Altogether the training documents amount to 4,000. In word feature selection phase, the document frequency threshold is set to 2, and 9107 word features are selected. Prior to training SVM, these 4,000 documents were evaluated by a three-fold cross validation to examine their integrity and qualification as training data. The cross validation accuracy reached up to 96.9%, where Radial basis function (RBF) kernel was used and SVM parameter cost and gamma were determined to be 8,000 and 0.0005.

25

In keyword selection phase, the statistics show that in average, ASR and OCR can provide 2.8 and 2.3 potential keywords for each automobile commercial, 4.5 and 2 for finance, 6.4 and 2.5 for healthcare, and 5.7 and 2.3 for IT, respectively. We empirically set both keyword selection parameter n and m to be 2. The recognition of brand names from ASR and OCR plays an important role, as brand names may be the best keyword candidates. Figure 24 presents the number of commercials, in which OCR and ASR recognised their brand names successfully. It shows OCR can recognise brand names in

30

a considerable amount of commercials, especially in, for example, automobile ones. Overall, OCR can recognise brand names of 56% of all commercials. OCR recognises brand names from two major sources: the trade name text and the website address on frame image, as shown in upper left image of Figure 22.

5

The classification based on manually recorded speech transcripts of commercials is firstly performed. As Figure 26(a) shows, except IT, all other categories achieve satisfactory classification result and the overall classification accuracy reaches 85.8%. The reason of low accuracy in IT category lies in the mismatch of category definition
10 between the training data and testing commercials. In the training data, IT category mainly covers computer hardware and software. However, in testing commercials, it includes other IT products, like printers and photocopy machines. ASR transcripts are also applied to perform text categorisation. As Figure 26(b) shows, the ASR transcripts deliver bad results in all categories. Figure 26(c) shows the classification results with
15 proxy articles. Compared with ASR transcripts, the classification results have been improved drastically and the overall classification accuracy increases from 43.3% to 80.9%. Figure 25 displays the F1 values of classifications based on all three types of inputs. For most categories, the proxy articles deliver slightly lower accuracies than the manually recorded speech transcripts. The accuracy differences imply that the errors in
20 keyword selection and proxy article acquisition do occur, and however, they do not necessarily provoke serious degrades on the final performance.

It will be appreciated that the invention has been described by way of example only and various modifications may be made in detail without departing from the spirit and scope
25 of the invention. Features of one aspect of the invention may be provided in combination with features of another aspect of the invention.

References

- [1] J.V. Vilanilam and A.K. Varghese, Advertising basics! A resource guide for beginners. Response Books, New Delhi, 2004.
- [2] M. Mizutani, etc., "Commercial detection in heterogeneous video streams using
5 fused multi-modal and temporal features," Proc. ICASSP'05.
- [3] L. Agnihotri, etc., "Evolvable visual commercial detector," Proc. CVPR' 03.
- [4] R. Lienhart, C. Kuhmunch, and W. Effelsberg, "On the detection and recognition of television commercials," Proc. ICMCS'97, pp. 509-516.
- [5] H. Sundaram and S.-F. Chang, "Computable scenes and structures in films," IEEE
10 Tran. TMM, 4(4):482-491, 2002.
- [6] J. R. Kender and B.L. Yeo, "Video scene segmentation via continuous video coherence," Proc. CVPR'98, CA, USA, pp.367-373.
- [7] M. Yeung and B.L. Yeo, "Time-constrained clustering for segmentation of video into story units," Proc. ICPR'96, Vienna, Austria, pp.375-380.
- [8] A. Hanjalic, etc., "Automated high-level movie segmentation for advanced video-
15 retrieval systems," IEEE Tran. CSVT, 9(4):580-588, 1999.
- [9] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Scene determination based on video and audio features," Proc. ICMCS'99, pp.685-690.
- [10] A. G. Hauptmann and M. J. Witbrock, "Story segmentation and detection of
20 commercials in broadcast news video," Proc. Conf. ADL' 98.
- [11] L. Chaisorn, etc., "A two-level multi-modal approach for story segmentation of large news video corpus," Proc. TRECVID'03, MD, USA.
- [12] K. Matsumoto, etc., "Shot boundary determination and low-level feature extraction experiments for TRECVID 2005," Proc. TRECVID'05, USA.
- [13] T. Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data
25 segmentation and classification," IEEE Tran. Speech and Audio Processing, 9(4):441-457, 2001.
- [14] HTK toolkit. [Online] Available: <http://htk.eng.cam.ac.uk/>.
- [15] N. Babaguchi, etc., "Event based indexing of broadcasted sports video by
30 intermodal collaboration," IEEE Tran. TMM, 4(1):68-75, 2002.
- [16] LIBSVM. [Online] Available: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- [17] J. Yuan, etc., "Tsinghua Univeristy at TRECVID 2005," Proc. TRECVID'05.

[18] C. Colombo, A. Del Bimbo, and P. Pala, in a paper entitled "Retrieval of commercials by video semantics," published in *Proc. CVPR'1998*, Santa Barbara, CA, USA, 1998

Claims

1. Apparatus for determining a likelihood that a candidate commercial boundary of a commercial broadcast in a segmented video broadcast is a commercial boundary, the
5 apparatus comprising a boundary classifier configured to:
determine whether a candidate video frame associated with a candidate commercial boundary of a segmented video broadcast comprises product information;
and
determine a likelihood the candidate commercial boundary is a commercial
10 boundary in dependence of the determination the candidate video frame comprises product information.
2. Apparatus according to claim 1, wherein the apparatus further comprises a commercial detector configured to locate boundaries of video programmes and
15 commercial broadcasts in the video broadcast and to derive the segmented video broadcast.
3. Apparatus according to claim 1 or claim 2, wherein the apparatus comprises a video shot transition detector configured to identify candidate commercial boundaries in
20 the segmented video broadcast.
4. Apparatus according to any preceding claim, wherein the boundary classifier is a binary boundary classifier.
- 25 5. Apparatus for determining a likelihood that a candidate commercial boundary of a commercial broadcast in a segmented video broadcast is a commercial boundary, the apparatus comprising:
a commercial detector configured to locate boundaries of video programmes and commercial broadcasts in the video broadcast and to derive the segmented video
30 broadcast;
a video shot transition detector configured to identify candidate commercial boundaries in the segmented video broadcast

a binary boundary classifier configured to:

determine whether a candidate video frame associated with a candidate commercial boundary of a segmented video broadcast comprises product information; and

5 determine whether the candidate commercial boundary is a commercial boundary in dependence of the determination the candidate video frame comprises product information.

6. Apparatus according to any preceding claim, wherein the boundary classifier is
10 configured to be trained by training data derived from video frames of the segmented video broadcast which comprise product information.

7. The apparatus of any preceding claim, wherein the candidate video frame
comprises a frame of a plurality of video frames of a candidate commercial window, the
15 candidate commercial window being associated with the candidate commercial
boundary, and the boundary classifier is configured to determine a commercial
boundary probability score for video frames of the candidate commercial window and to
determine the likelihood the candidate commercial boundary is a commercial boundary
from a plurality of the commercial boundary probability scores.

20

8. The apparatus of claim 7, wherein the candidate commercial window is a first
symmetrical window of n video frames in the segmented video broadcast either side of
the candidate commercial boundary.

25 9. The apparatus of any preceding claim, wherein the apparatus is configured to
identify a candidate video frame from a motion measurement of a video frame
associated with the candidate commercial boundary.

10. The apparatus of claim 9, wherein the apparatus is configured to perform the
30 motion measurement from a calculation of average intensity of motion vectors in the
video frame.

11. The apparatus of any preceding claim, wherein the boundary classifier is configured to derive the training data by parsing video frames comprising product information and extracting a video frame feature for one or more portions of the video frame and/or for a complete video frame.

5

12. The apparatus of claim 11, wherein the apparatus is configured to extract a video frame feature for at least one of colour, texture and edge.

13. The apparatus of claim 13, wherein the apparatus is configured to extract video
10 frame features for colour, texture and edge.

14. Apparatus according to any preceding claim, wherein the boundary classifier is further configured to determine a likelihood the candidate commercial boundary is a commercial boundary in dependence of a determination the candidate video frame
15 comprises an audio scene change.

15. Apparatus according to claim 14, wherein the boundary classifier is configured to make the determination the candidate video frame comprises an audio scene change from a distance measurement of audio properties of first and second audio frames of an
20 audio segment of the video broadcast, the audio segment being associated with the candidate commercial boundary.

16. Apparatus according to claim 15, wherein the boundary classifier is configured to determine the candidate video frame comprises an audio scene change by:

25 partitioning the audio segment into a plurality of sets of audio frames, each set of audio frames having frames of equal length, the length of one set of audio frames being different from a length of another set of audio frames to determine a set of difference sequences of audio properties from the sets of audio frames, and

determining a correlation between difference sequences of the set of difference
30 sequences.

17. Apparatus according to claim 16, wherein the boundary classifier is configured to identify an offset between an audio scene change and a candidate commercial boundary from the correlation between difference sequences.
- 5 18. Apparatus according to any of claims 12 to 15, wherein the boundary classifier is further configured to align the audio scene change with the candidate commercial boundary.
19. Apparatus according to any of claims 14 to 18, wherein the boundary classifier
10 is configured to:
- extract audio features from the audio segment;
 - train first and second statistical models for audio scene change and for non-audio scene change from the audio features extracted from the audio segment; and
 - classify a candidate audio segment associated with the candidate commercial
15 boundary from the first and second statistical models.
20. Apparatus according to claim 19, wherein the boundary classifier is configured to classify a candidate audio segment by determining a probability value from at least one of the first and the second statistical models.
- 20 21. Apparatus according to any preceding claim, wherein the boundary classifier is configured to classify the candidate commercial boundary as a commercial boundary from a fusion of likelihood scores for frame marked with product information (FMPI), audio scene change (ASC) and, optionally, audio silence and video black frame.
- 25 22. Apparatus for classifying a commercial video broadcast in a pre-defined category, the apparatus comprising:
- a proxy document identifier configured to identify a proxy document as a proxy of the commercial video broadcast;
 - 30 a training module configured to compile training data from a corpus of training documents; and

a classifier module configured to be trained by data from the training data, the classifier module being further configured to classify the commercial video broadcast from an examination of proxy data from the proxy document.

- 5 23. Apparatus according to claim 22, wherein the apparatus further comprises:
a video processor configured to extract video and/or audio data from a frame of
the video broadcast commercial and to convert the video and/or audio data to text data;
a first keyword derivation module configured to identify a keyword from the text
data; and
10 wherein the document identifier is configured to identify the proxy document as
a document related to the keyword.
24. Apparatus according to claim 22 or claim 23, wherein the training module is
configured to compile training data from external resources.
- 15 25. Apparatus according to any of claims 22 to 24, wherein the classifier module is a
support vector machine.
26. Apparatus according to claim 23, wherein the video processor comprises an
20 audio speech recognition module and an optical character recognition module.
27. Apparatus according to any of claims 22 to 26, further comprising a text
correction module for spell-checking and correcting the text data.
- 25 28. Apparatus for classifying a commercial video broadcast in a pre-defined
category, the apparatus comprising:
a video processor configured to extract video and/or audio data from a frame of
the video broadcast commercial and to convert the video and/or audio data to text data
using an audio speech recognition module and an optical character recognition module,
30 the video processor further comprising a text correction module for spell-checking and
correcting the text data;

a proxy document identifier comprising a first keyword derivation module configured to identify a keyword from the text data, the proxy document identifier being configured to identify a proxy document, as a proxy of the commercial video broadcast, by identifying the proxy document as a document related to the keyword;

5 a training module configured to compile training data from an external resource of a corpus of training documents ; and

a support vector machine classifier module configured to be trained by data from the training data, the classifier module being further configured to classify the commercial video broadcast from an examination of proxy data from the proxy
10 document.

29. Apparatus according to any of claims 23 to 28, wherein the first keyword derivation module is configured to identify a keyword by querying the text data for an occurrence of an identifier word; and in dependence of detecting an occurrence of the
15 identifier word, identifying the identifier word as a keyword.

30. Apparatus according to any of claims 23 to 29, wherein the first keyword derivation module is configured to identify a keyword by querying the text data for an occurrence of an identifier word and, in dependence of not detecting an occurrence of
20 the identifier word, identifying another word in the text data, for example a noun word, as a keyword.

31. Apparatus according to any of claims 23 to 30, wherein the document identifier is configured to identify the proxy document as a document related to the keyword by
25 querying an external document index or database with the keyword as a query term and assigning a most relevant result document of the query as the proxy document.

32. Apparatus according to any of claims 22 to 31, wherein the apparatus further comprises:
30 a first text preprocessing module configured to rationalise proxy data; and
a test word vector mapper configured to map proxy data to a proxy vector for examination by the classifier module.

33. Apparatus according to any of claims 22 to 32, wherein the training module comprises a second keyword derivation module configured to identify a training keyword by querying the training data for an occurrence of a training identifier word and, in dependence of detecting an occurrence of the training identifier word, identifying the training identifier word as a training keyword.

34. Apparatus according to any of claims 22 to 33, wherein the second keyword derivation module is configured to identify a training keyword by querying the training data for an occurrence of a training identifier word and, in dependence of not detecting an occurrence of the training identifier word, identifying another word in the training data as a keyword.

35. Apparatus according to any of claims 22 to 34, wherein the training module further comprises:
a second text preprocessing module configured to rationalise data in the training corpus; and
a training data vector mapper configured to map training data to a training data vector, for training of the classifier module.

20

36. Apparatus for identifying a boundary of a commercial broadcast in a video broadcast and classifying the commercial broadcast in a pre-defined category, the apparatus comprising:

a video shot transition detector configured to identify a candidate commercial boundary in the video broadcast;

a boundary classifier configured to verify the candidate commercial boundary as a commercial boundary;

a commercial classifier configured to classify the commercial in a pre-defined category.

30

37. A method for determining a likelihood that a candidate commercial boundary of a commercial broadcast in a segmented video broadcast is a commercial boundary, the method comprising, with a boundary classifier:
- determining whether a candidate video frame associated with a candidate
- 5 commercial boundary of a segmented video broadcast comprises product information; and
- determining a likelihood the candidate commercial boundary is a commercial boundary in dependence of the determination the candidate video frame comprises product information.
- 10
38. A method for determining a likelihood that a candidate commercial boundary of a commercial broadcast in a segmented video broadcast is a commercial boundary using the apparatus of any of claims 1 to 21.
- 15
39. A method for classifying a commercial video broadcast in a pre-defined category, the method comprising:
- identifying, with a proxy document identifier, a proxy document as a proxy of the commercial video broadcast;
- compiling, with a training module, training data from a corpus of training
- 20 documents; and
- training a classifier module with data from the training data, and classifying, with the classifier module, the commercial video broadcast from an examination of the proxy data from the proxy document.
- 25
40. A method for classifying a commercial video broadcast in a pre-defined category using the apparatus of any of claims 22 to 35.
- 30
41. A method of identifying a boundary of a commercial broadcast in a video broadcast and classifying the commercial broadcast in a pre-defined category, the method comprising:
- identifying a candidate commercial boundary in the video broadcast using a video shot transition detector;

verifying the candidate commercial boundary as a commercial boundary with a boundary classifier;

classifying commercial in a pre-defined category with a commercial classifier.

- 5 42. A system, for use in a video signal processor, for locating boundaries of individual TV commercials and categorising a TV commercial into one of predefined categories according to advertised product or service, said system comprising:
- a TV commercial detector for locating boundaries of video programmes and commercials;
- 10 a video shot transition detector for locating potential individual commercials' boundaries within a chunk of commercial segment;
- a binary boundary classifier for determining the true boundaries, wherein said classifier comprises a set of mid-level features to capture audio-visual characteristics significant for parsing commercials' video content, Black Frame-inclusive/exclusive
- 15 multi-modal feature vectors, and a supervised learning algorithm;
- a commercial categoriser for classifying advertised product or service into one of predefined categories, wherein said classifier comprises ASR and OCR modules for extracting raw textual information followed by spell checking and correction, keyword selection and keyword-based query expansion using external resources, machine
- 20 learning-based classifiers trained from external resources categorised according to different topics, and IR text pre-processing module; visual-based object recognition.

43. The system as claimed in claim 42, wherein the individual commercial boundaries' detection is reduced to a binary classification on the basis of audio-visual
- 25 features extracted within a symmetric window of each candidate boundary.

44. The system as claimed in claim 43, wherein said candidate boundaries comprise video shot transition points within TV commercial segments, wherein said video shot transition has two types, which are cuts and gradual transitions, and wherein the
- 30 candidate boundary is located at the middle of the transition for gradual transition type.

45. The system as claimed in claim 42, wherein said mid-level features comprise Image Frame Marked with Product Information (FMPI), Audio Scene Change Indicator (ASCI), SILENCE, and BLACK FRAMES, wherein FMPI and ASCI are based on the TV commercial video content, and SILENCE and BLACK FRAMES are based on the post-editing techniques between consecutive TV commercials within a chunk of TV commercial.

46. The system as claimed in claim 45, said FMPI is referred to as those key image frames containing corporate symbols, brand names, product appearance, mild encouragement captions, contact information, and other product/service related textual information, wherein FMPI yields a set of visual features to be fed into a supervised learning algorithm for training a classifier.

47. The system of claim 46, comprising using a 4-dimensional feature wherein four neighbour video shots of each candidate boundary are considered, for left or right two shots within the left- or right-window of 2 seconds, two highest FMPI confidence values (1 or 0) are selected as features for the left side or the right side, respectively, otherwise, FMPI recognition is applied to one key frame selected from the middle point of each video shot.

20

48. The system as claimed in claim 45, said ASCI is referred to as two classifiers modelling the audio dynamic characteristics within a predefined temporal window for two categories of audio signal, namely, with and without the happening of an audio scene change, respectively; ASCI related feature is 2-dimensional, which comprises two probability values within a temporal symmetric window, say 4-second window, of a candidate commercial boundary.

49. The system as claimed in claim 45, said SILENCE is referred to as an audio segment, which persistently exhibits a short-time energy function below a threshold for above a predefined minimum time length; SILENCE related feature is 1-dimensional.

30

50. The system as claimed in claim 45, said BLACK FRAMES is referred to as a series of consecutive black frames with a length above a predefined frame number, wherein an image frame is declared as a black frame if the mean and the variance are below a threshold; BLACK FRAMES related feature is 1-dimensional.

5

51. The system as claimed in claim 43, wherein said audio-visual features are constructed by using FMPI, ASCI, SILENCE, and BLACK FRAMES; the fusion of FMPI and ASCI related features provides a system and method independent of post-editing techniques, and the fusion of FMPI, ASCI, SILENCE, and BLACK FRAMES
10 provides a system and method dependent on post-editing techniques.

52. The system as claimed in claim 42, wherein TV commercial classification is formulated as a text categorisation problem, wherein ASR and OCR techniques generate initial semantic information; external World Wide Web-based resources deliver
15 expanded query articles; the training based on external topic-wise document corpus yields a text categoriser matching the predefined category of advertised product or service, which accordingly classifies expanded query articles.

53. The system as claimed in claim 42, wherein ASR and OCR generated raw
20 textural information goes through spell checking and correction by external World Wide Web-based resources, and wherein a list of nouns and noun phrases are extracted from the corrected textual information as query keywords, which are passed to encyclopaedia and Google search engine to retrieve related articles, wherein those with high relevance ranking are selected as query articles.

25

54. A method of locating boundaries of individual TV commercials comprising the steps of:

partitioning an input video stream into commercial segments and programme segment;

30 detecting video shot changes including cuts, dissolves, and fade-out/-in to determine candidate boundaries of individual TV commercials;

computing mid-level features FMPI, ASCI, SILENCE and BLACK FRAMES within a symmetric window of each candidate boundary; and

utilising a supervised learning algorithm to fuse FMPI, ASCI, SILENCE and BLACK FRAMES to distinguish true boundaries of an individual TV commercial.

5

55. A method of constructing the mid-level feature of FMPI comprising the steps of: extracting colour-based, texture-based, and edge-based low-level features of an image frame to capture the local and the global visual characteristics; and

utilising a supervised learning algorithm including an SVM classifier to classify
10 an image frame into FMPI or Non-FMPI.

56. A method of constructing the mid-level feature of ASCI comprising the steps of: extracting low-level audio features including Mel-frequency cepstral coefficients (MFCCs) and its first and second derivatives, mean and variance of short time energy log
15 measure (STE), mean the variance of short-time zero-crossing rate (ZCR), short-time fundamental frequency (SF), mean of the spectrum flux and harmonic degree (HD) within the symmetric window of a candidate boundary point;

aligning the audio feature window with multi-scale Kullback-Leibler distance computing; and

20 utilising Hidden Markov Model (HMM) to train two classifiers to model the audio characteristics for two classes of audio segments, namely, with and without the happening of audio scene changes within the adjusted audio feature window.

57. The method as claimed in claim 56, wherein said alignment of the audio feature
25 window comprising the steps of:

using different window sizes to compute changes between successive audio analysis window with the Kullback-Leibler distance metric, yielding a set of difference sequences;

normalising each difference sequence to [0,1] though dividing difference values
30 by the maximum of each sequence; and

determining the most likely audio scene change point by locating the highest accumulated difference values derived from the set of difference sequences.

58. A method of classifying a TV individual commercial comprises the steps of:
constructing a training corpus that consists of a substantial amount of training documents with topics ranging over various fields including car, health care, finance,
5 IT;
aggrandising the deficient and less informative scripts from ASR and OCR by
questing relevant articles in World Wide Web , wherein spell checking and correction
are applied to the raw textual output from ASR and OCR;
performing IR preprocessing including the Porter's Stemming algorithm and the
10 Stop Word Removal algorithm on the testing and training documents;
employing document frequency techniques to select the word features for the
training dataset;
utilising SVM to train classifiers for those topics by using selected word
features; and
15 applying trained SVM-based classifiers to categorise the expanded query articles
selected by relevance ranking.