(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 January 2007 (11.01.2007)

PCT

(10) International Publication Number
WO 2007/004110 A2

(51) International Patent Classification:
G11B 27/034 (2006.01) *H04N 9/802* (2006.01)
G11B 27/028 (2006.01) *H04N 5/265* (2006.01)

(21) International Application Number:
PCT/IB2006/052088

(22) International Filing Date: 26 June 2006 (26.06.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/695,654 30 June 2005 (30.06.2005) US

(71) Applicant *(for all designated States except US)*: **KONINKLIJKE PHILIPS ELECTRONICS N.V.** [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **AGNIHOTRI, Lalitha** [IN/US]; 345 Scarborough Road P.o. Box 3001, Briarcliff Manor, NY 10510-8001 (US). **BARBIERI, Mauro** [IT/US]; 345 Scarborough Road P.o. Box 3001, Briarcliff Manor, NY 10510-8001 (US). **DIMITROVA, Nevenka** [US/US]; 345 Scarborough Road P.o. Box 3001, Briarcliff Manor, NY 10510-8001 (US).

(74) Common Representative: **KONINKLIJKE PHILIPS ELECTRONICS N.V.**; INTELLECTUAL PROPERTY & STANDARDS, C/o Bram, Eric M. 345 Scarborough Road, P.o. Box 3001, Briarcliff Manor, NY 10510-8001 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) **Title:** SYSTEM AND METHOD FOR THE ALIGNMENT OF INTRINSIC AND EXTRINSIC AUDIO-VISUAL INFORMATION

(57) **Abstract:** A system and method are provided for the alignment of intrinsic and extrinsic audio-visual information using classifications of the information and is particularly useful for aligning intrinsic and extrinsic information of different languages. The system comprises an intrinsic content analyser (2,3) that is communicatively connected to an audio-visual source (1), such as a film source that may be in a dubbed language. The film is searched for intrinsic data that is extracted and classified. Further, the system comprises an extrinsic content analyser (5,6) communicatively connected to an extrinsic information source (4), such as an original language film screenplay available through the Internet. The system searches the extrinsic information source, retrieves and then classifies the extrinsic information. The intrinsic and extrinsic information is aligned by correlating (7) the classifications to provide a multi-source data structure. The correlation is independent of the language of the intrinsic and the extrinsic information.

WO 2007/004110 A2

SYSTEM AND METHOD FOR THE ALIGNMENT OF INTRINSIC AND EXTRINSIC AUDIO-VISUAL INFORMATION

The invention relates to the alignment of intrinsic and extrinsic audio-visual information, more specifically it relates to analysis and correlation of features in e.g. a film with features not present in the film but available e.g. through the Internet.

People who are interested in films throughout the World were for many years obliged to consult books, printed magazines or printed encyclopaedias in order to obtain additional information about a specific film. With the appearance of the Internet, a number of Internet sites were dedicated to film related material. An example is the Internet Movie Database (http://www.imdb.com) which is a very thorough and elaborated net site providing a large variety of additional information to a large number of films. Even though the Internet facilitates access to additional film information, it is up to the user to find his or her way through the vast amount of information available though out the Internet.

With the appearance of the Digital Versatile Disk (DVD) medium, additional information relating to a film is often available in a menu format at the base menu of the DVD film. Often interviews, alternative film scenes, extensive cast-lists, diverse trivia, etc. are made available. Furthermore, the DVD format facilitates scene browsing, plot summaries, bookmarks to various scenes etc. Even though additional information is available on many DVDs, it is the provider of the film that is the person who selects the additional information, the additional information is limited by the available space on a DVD disk and it is static information created during an authoring process. For the traditional broadcast of films, even this static information is not available.

The amount of films available and the amount of additional information available throughout the World concerning the various films, actors, directors, etc. are overwhelming, and users suffer from "information overload". People with interest in films often struggle with problems relating to how they can find exactly what they want, and how to find new things they like. To cope with this problem various systems and methods for searching and analysis of audio-visual data have been developed. Different types of such systems are available, for example, there are systems for automatic summarisation, such a system is described in US application 2002/0093591. Another type of system are systems for targeted search based on e.g. selected image data such as an image of an actor

in a film, such a system is described in the US application 2003/0107592. A system offering a significant improvement for the consumer has been presented in the literature and describes the text based alignment of screenplays with closed captions to extract high level semantic information of films that is not available, or difficult to extract, via other means. Enabling people to find exactly what they want is especially difficult for situations wherein the original language of the film has been changed, for instance, by dubbing the audio track. Therefore, limitations of the prior art systems constrain the usage of such systems to selected geographical areas.

The inventors have appreciated that a system being capable of integrating intrinsic and extrinsic audio-visual data, such as integrating audio-visual data on a DVD-film with additional information found on the Internet, irrespective of the languages of the intrinsic and extrinsic audio-visual data, is of benefit, and have, in consequence, devised the present invention.

It is an object of the present invention to provide an improved system for alignment of audio-visual data that is independent of the languages of the intrinsic and extrinsic audio-visual data.

Accordingly there is provided, in a first aspect, a system for alignment of intrinsic and extrinsic audio-visual information, the system comprising an intrinsic content analyser, the intrinsic content analyser being communicatively connected to an audio-visual source, the intrinsic content analyser being arranged to classify intrinsic information extracted from content sourced by the audio-visual source resulting in intrinsic classifications; an extrinsic content analyser, the extrinsic content analyser being communicatively connected to an extrinsic information source, the extrinsic content analyser being arranged to classify extrinsic information extracted from extrinsic information sourced by the extrinsic information source resulting in extrinsic classifications; and an intrinsic information and extrinsic information correlator being communicatively connected to the intrinsic content analyser and to the extrinsic content analyser and being arranged to correlate the intrinsic classifications with the extrinsic classifications, thereby providing a multi-source data structure.

An audio-visual system, such as an audio-visual system suitable for home-use, may contain processing means that enables analysis of audio-visual information. Any type of audio-visual system may be envisioned, for example such systems including a

Digital Versatile Disk (DVD) unit or a unit capable of showing streamed video, such as video in an MPEG format, or any other type of format suitable for transfer via a data network. The audio-visual system may also be a "set-top-box" type system suitable for receiving and showing audio-visual content, such as TV and film, either via satellite or via cable. The audio-visual system may also be a personal audio-visual storage/communication portable device. The video could be broadcast or streamed. The system comprises means for either presenting audio-visual content, i.e. intrinsic content, to a user or for outputting a signal such that audio-visual content may be presented to a user. The adjective "intrinsic" should be construed broadly. Intrinsic content may be content that may be extracted from the signal of the film source. The intrinsic content may be the video signal, the audio signal, text that may be extracted from the signal, etc.

The system comprises an intrinsic content analyser. The intrinsic content analyser is typically a processing means capable of analysing audio-visual data. The intrinsic content analyser is communicatively connected to an audio-visual source, such as to a film source. The intrinsic content analyser is arranged to search the audio-visual source by using an extraction algorithm to extract intrinsic information. The intrinsic content analyser is further arranged to classify the intrinsic information extracted from the content sourced by the audio-visual source.

The system also comprises an extrinsic content analyser. The adjective "extrinsic" should be construed broadly. Extrinsic content is content, which is not included in or may not, or only with difficulty, be extracted from the intrinsic content. Extrinsic content may typically be such content as a film screenplay, storyboard, reviews, analyses, etc. The extrinsic information may also contain timestamps and could be, for example, a time-stamped screenplay. The extrinsic information source may be an Internet site, a data carrier comprising relevant information, etc. The extrinsic content analyser is further arranged to classify the extrinsic information extracted from the content sourced from the extrinsic information source.

The system also comprises a means for correlating the intrinsic and extrinsic information in a multi-source information structure. The rules dictating this correlation may be part of the extraction and/or the retrieval algorithms. A correlation algorithm may also be present, the correlation algorithm correlating the intrinsic and extrinsic information in the multi-source information structure. The correlation algorithm correlates the intrinsic

and extrinsic information based upon the classification of the intrinsic and extrinsic information. Correlation based upon the classification of the intrinsic and extrinsic information rather than the content of the intrinsic and extrinsic information *per se* renders the system more tolerant to the use of different languages of the sources for the intrinsic and extrinsic information. Such language differences often occur when films are dubbed, for example. The multi-source information structure may be a low-level information structure correlating various types of information e.g. by data pointers. The multi-source information structure may not be accessible to a user of the system, but rather to a provider of the system. The multi-source information structure is normally formatted into a high-level information structure, which is presented to the user of the system.

The system of claim 2 has the advantage that it relies upon the classification of audio related features of the intrinsic and extrinsic information that are easy to compute upon resource limited machines and yet still reach the object of aligning the intrinsic and extrinsic information irrespective of the languages of the information.

Advantageously the system of claim 3 identifies the location or duration of the classifications identified in the intrinsic and extrinsic information allowing the correlation of the classifications to be performed in a straightforward manner by aligning locations or durations.

The system of claim 4 is arranged such that both the intrinsic content analyser and the extrinsic content analyser can be arranged to classify the audio related features as silence and or speech. This is advantageous since both the intrinsic content analyser and the extrinsic content analyser only need to identify a limited number of classifications in the audio related information further reducing the resources required to achieve the object of aligning the intrinsic and extrinsic information irrespective of the languages of the information.

Favourably, the alignment of the intrinsic and extrinsic information can be further improved by identifying numerous speakers from the various voices detected. Detecting the individual speakers also leads directly to the identification of speaker changes and both of these forms of information can be taken into account during the correlation phase for improved alignment. This can lead to an improved correlation between the intrinsic and the extrinsic information independent of the language of the intrinsic and extrinsic information.

Advantageously, the system of claim 6 is arranged to align the intrinsic and extrinsic information irrespective of the languages of the information when the extrinsic information does not include timestamps. This is achieved by estimating the location or duration of classifications based on, for example, the duration of the film and the location or durations of classifications in the intrinsic and extrinsic information.

Beneficially, by identifying names and the location or point in time of such names a further improvement in the alignment of the intrinsic and extrinsic information is achieved by the use of such locations or points in time during the correlation phase. This is facilitated by the fact that some names, such as characters in a film, often remain the same even after language translation or dubbing.

Advantageously, the intrinsic information comprises a film and the extrinsic information comprises a screenplay allowing a high level of understanding of the context of a film to be recognized by a system with limited processing resources even when the languages of the intrinsic and extrinsic information sources are different.

Favourably, the intrinsic information comprises a film and the extrinsic information comprises a time-stamped screenplay allowing the context of the film to be aligned with the content of the film by a system with further limited processing resources.

According to a second aspect of the present invention the object is realized by a method as claimed in claim 10. Further advantageous measures are defined in claims 11 through 14.

A third aspect of the present invention provides a computer-readable recording medium containing a program to realize the object of the invention as defined in claim 15.

According to a fourth aspect of the present invention the object is realized by providing a program for controlling an information processing apparatus as claimed in claim 16.

These and other aspects, features and/or advantages of the present invention will be apparent from and elucidated with reference to the embodiments described hereinafter.

Preferred embodiments of the invention will now be described in detail with reference to the drawings in which:

FIG. 1a is a schematic diagram of a first embodiment of the present invention;

FIG. 1b is a diagram showing the alignment of intrinsic and extrinsic information based on classification of the audio;

FIG. 2 is a flowchart illustrating individual method steps and the interconnections of said method steps of the invention;

FIG. 3a is a schematic diagram of a second embodiment of the present invention;

FIG. 3b is a diagram showing the alignment of intrinsic and extrinsic information based on classification of the audio and changes in the speaker;

FIG. 4a is a schematic diagram of a third embodiment of the present invention;

FIG. 4b is a diagram showing the alignment of intrinsic and extrinsic information based on classification of the audio, changes in the speaker and scene detection;

FIG. 5a is a schematic diagram of a fourth embodiment of the present invention;

FIG. 5b is a diagram showing the alignment of intrinsic and extrinsic information based on classification of the audio, changes in the speaker, scene detection and name spotting within the intrinsic and extrinsic information; and

FIG. 6 is a schematic illustration of results created during the correlation phase used for intrinsic and extrinsic information alignment.

FIG.1a shows a system 8 for integrated analysis of extrinsic and intrinsic audio-visual information according to the present invention that operates independent of the languages of the intrinsic and extrinsic audio-visual data. A video signal source 1 is the source of intrinsic audio-visual information, for example, this could be a feature film on a data carrier such as a DVD disk or a television broadcast. These two examples exemplify suitable sources. The intrinsic information is information that may be extracted from the audio-visual signal directly, i.e. from image data, audio data and/or transcript data. Transcript data may be in the form of subtitles, closed captions or teletext information. The extrinsic audio-visual information is here exemplified by extrinsic access to the screenplay of the feature film from a screenplay source 4, for example via an Internet connection.

Further, extrinsic information may also be the storyboard, published books, additional scenes from the film, trailers, interviews with director and/or cast, reviews by film critics, etc. All such extrinsic information may be obtained through the Internet connection. These further forms of extrinsic information may like the screenplay undergo analysis.

The intrinsic information is processed using an intrinsic content analyser comprising an audio feature extraction unit 2 and an audio classification unit 3. The intrinsic content analyser may be a computer program adapted to search and analyse intrinsic content of a film. This would require a processor, memory for the program and the data to be processed and suitable input/output connections. The audio content is extracted from the video content originating from video signal source 1 and is processed initially by the audio feature extraction unit 2. The audio feature extraction unit 2 may use time based or frequency based analysis to extract the audio features and is well known in the prior art. For example, the analysis can be based upon low level signal properties, Mel Cepstral Frequency Coefficients (MFCCs), psycho-acoustic features including roughness, loudness and sharpness or by modelling temporal envelope fluctuations in the auditory domain.

After audio feature extraction the audio processing further includes audio classification by the audio classification unit 3. The classification of audio is also well known in the prior art. Typically a quadratic discriminate analysis is used. Features are normally calculated by segmenting the audio into frames, where a frame is usually around one-half to one second in length. The frame-to-frame distance, or hop size, is generally less than the frame length resulting in overlapping frames, which generally improves the classification process. The feature vectors resulting from the audio feature extraction process are grouped into classes based on the type of audio and are used to parameterise an N-dimensional Gaussian mixture model, where each Gaussian distribution has its own mean and variance for each class. N is the length of the feature vector resulting from the audio feature extraction process. The model is trained as is usual in the prior art. Such training methods could be arranged to use the so called ".632+ bootstrap" method, or the "leave-one-out bootstrap" method which are typically known to the skilled person. The audio classification unit 3 outputs the classification of the audio for each frame of audio, for example, each frame can be classified as speech, silence, music, noise or combinations thereof, such as, speech and speech, speech and noise, speech and music, etc. Further processing is performed on classifications not defined as silence or music. The output of

the audio classification unit 3 is shown diagrammatically in the lower portion of FIG. 1b, noted by the term "Audio Signal Classification". Referring also to the flowchart of FIG. 2 the audio classification is denoted by method step 21.

The extrinsic information is processed using an extrinsic content analyser and comprises an audio related feature extraction unit 5 and an audio related classification unit 6. The extrinsic content analyser may be adapted to search the extrinsic information based on the extracted intrinsic data from the intrinsic content analyser. The extracted intrinsic data may be as simple as the film title, however the extracted intrinsic data may also be a complex set of data relating to the film. The extrinsic content analyser may include models for screenplay parsing, storyboard analysis, book parsing, analysis of additional audio-visual materials such as interviews, promotion trailers etc. The output of the extrinsic content analyser is a data structure that contains the audio related classification of the extrinsic information and timestamps within the film for which the classification is valid. Typical classifications are again speech, silence, music, noise or combinations thereof, such as, speech and speech, speech and noise, speech and music, etc. Advantageously, long lines of dialogue are used as anchors in order to segment the film into smaller sections for alignment. The extrinsic information may be further analysed to extract high-level information about scenes, cast mood, etc, as is known from the prior art. As an example, high level structural parsing can be performed on the original language screenplay with timestamps from the aligned original language screenplay source 4. The characters can be determined and cross-referenced with actors e.g. through information accessed via the Internet, e.g. by consulting an Internet based database such as the Internet Movie Database. The scene locations and the scene descriptions may also be extracted again with timestamps.

Referring again to FIG. 1a the extrinsic information is an aligned original language screenplay from the aligned original language screenplay source 4. The term aligned is meant to indicate that an external service provider or system has already aligned the original language screenplay to the original language film. The term "aligned" is taken to be equivalent to the term "time-stamped" in this description. This alignment will not be valid for a dubbed version of the film in another language and is improved by the present invention. The extrinsic information will in most cases not contain audio information from which audio features can be extracted directly in the manner known to the prior art. For

example, the aligned original language screenplay will probably be text based, however, even in this case the audio related feature extraction unit 5 in combination with the audio related classification unit 6 can still determine the classifications of silence, speech, music, noise and combinations thereof by textually parsing the screenplay and studying, for example, the timestamps of the dialogue of each actor or actress. The term "related" is used in the naming of the audio related feature extraction unit 5 and the audio related classification unit 6 to make a clear distinction between audio based feature extraction based upon the intrinsic audio samples and audio related feature extraction based upon extrinsic information. An example of the output of the audio classification unit 3 is shown diagrammatically in the upper portion of FIG. 1b, noted by the term "Aligned Screenplay Timeline". Referring again to the flowchart of FIG. 2 the audio related classification is denoted by method step 26.

The intrinsic and extrinsic information are correlated in order to obtain a multi-source data structure by the alignment unit 7. The alignment unit 7 correlates the classifications and timestamps of the classifications. Using the multi-source data structure a further high-level information structure may be generated by the system, for example, by using a model for actors, compressing plot summaries and by detecting scene boundaries. The model for actors may include audio-visual person identification in addition to character identification from the multi-source data structure. Thus the end user may be presented with a listing of all the actors appearing in the film, and may be able to select an actor and be presented with additional information concerning this actor, such as other films in which the actor appears or other information about a specific actor or character. A compressed plot summary module may include plot points and story and sub-story arcs. These are the most interesting points in the film. This high-level information is very important for the summarisation of the film. The user may thereby be presented with a different type of plot summary than what is typically provided on the DVD or by the broadcast, or may choose the type of summary that the user is interested in. During semantic scene detection, shots for scenes and scene boundaries are established as is known in the prior art. The user may be presented with a complete list of scenes and correspondent scene from the screenplay in order to compare the director's interpretation of the screenplay for various scenes, or to allow the user to locate scenes containing a specific character. A typical example of the output of the alignment unit 7 is shown in FIG. 1b by

successful alignment points 10. In the flowchart of FIG. 2 the related method step is that of coarse alignment, step 25.

FIG. 3a shows a second embodiment of the invention leading to more precise alignment of the intrinsic and extrinsic information by using speaker identification known in the prior art to identify sentence boundaries. Since the audio classification boundaries can have some lag/lead/overlap/overrun when compared to the timing of the original film it is beneficial to adjust the coarse alignment produced by step 25 of FIG. 2. This can be achieved because correlation between sentence boundaries will always occur, even when the languages are different. In FIG. 3a the intrinsic information is processed using an intrinsic content analyser further comprising a speaker identification unit 31 and a speaker change detector 32. Generally, voice models are used to identify individual speakers from only intrinsic data. Further methods of speaker identification known from the prior art are those using voice fingerprints and face models.

The audio content is again extracted from the video content originating from video signal source 1 and is processed initially by the audio feature extraction unit 2. Speaker identification is preferably achieved by the extraction of the Mel Cepstral Frequency Coefficients (MFCCs) in the audio feature extraction unit 2. The audio classification unit 3 takes the audio features, classifies the audio as described earlier and outputs the classification of the audio for each frame of audio. The output of the audio classification unit 3 is shown diagrammatically in the lower portion of FIG. 3b, noted by the term "Audio Signal Classification". Referring also to the flowchart of FIG. 2 the audio classification is denoted by method step 21. In parallel to audio classification, the speaker identification unit 31 also use the audio features to identify the individual speakers, see step 22 of FIG. 2. Once individual speakers are identified the speaker change detector 32 easily detects the boundaries between individual speakers, i.e. sentence boundaries, in step 23 of FIG. 2. The outputs of the speaker identification unit 31 and the speaker change detector 32 are shown in the middle portion of FIG. 3b. It is possible that during dubbing one voice may be used for multiple characters in the original movie. However, the original screenplay information coupled with the timestamps provides enough information to resolve this problem.

In the second embodiment shown in FIG. 3a the extrinsic information is extracted in the method as described for the first embodiment, i.e. that of FIG. 1a. The

aligned extrinsic information again contains timestamps. This is denoted by method step 26 in FIG. 2. The intrinsic and extrinsic information are again correlated in order to obtain a multi-source data structure by the alignment unit 7 of FIG. 3a. The alignment unit 7 correlates the classifications and the timestamps of the classifications to get a coarse alignment, as shown in step 25 of FIG. 2. The changes in speakers, or sentence boundaries, are used to provide the maximum correlation between the original language and the dubbed language films. The related method step is step 27 of FIG. 2. A typical example of the output of the alignment unit 7 is shown in FIG. 3b by improved alignment points 10 over that of the first embodiment.

In the third embodiment of FIG. 4a is provided a system that can achieve the object of the invention without the requirement that the original language screenplay has timestamps available. In such a situation is it advantageous to have estimate durations relevant to the film. For example, a rough timeline of the original screenplay can be estimated based upon knowledge of the length of the film, available from the extrinsic or intrinsic information. As known in the prior art, visual shot and scene changes can also be aligned with high-level information in the screenplay to the film. Such alignments serve as anchors for alignment of the screenplay where the relative durations of dialogues in the original screenplay can be estimated. For example, a sentence with twice as many words as another probably lasts twice as long. It is further advantageous if for each word, a statistical model is available to estimate how long it takes to speak each word via training on labeled data. For example, in the original language screenplay a statistically trained word duration estimator can estimate how long each dialogue is spoken for. For statistically training a word duration estimator, ground truths can be obtained from duration of the words in many films from which an estimate of how long any particular word, e.g. bottle, takes to utter on average, along with a standard deviation. On a coarse level, matching of the longest line in the screenplay and finding the longest monologue in the film can provide adequate alignment. Optionally, an estimate of the duration of each sentence can be made and matching portions for each sentence can be located. Also, very short lines can be located and aligned to short audio classifications taking into account the knowledge of the duration of the film. The word duration estimator 44 of FIG. 4a can use any of the methods stated above to provide timestamps to the screenplay. The related

method step is 29 in FIG. 2 and uses as input the audio related classifications of the original language screenplay from step 26.

The intrinsic content analyzer of FIG. 4a may optionally further comprise a video feature extraction unit 41 and a scene detection unit 42. These units work substantially in the video feature domain and are common building blocks known to the skilled person. The outputs of these units are indicated in FIG. 4b as scene alignments and shot changes. The alignment unit 7 of FIG. 4a uses the estimated timeline for the screenplay, the audio classifications and timestamps from the intrinsic information, the speaker identification and speaker changes to correlate the intrinsic and extrinsic information. A similarity matrix can be created for aligning the duration, estimated or not, of sections of dialogue. For example, every dialogue duration i in the screenplay within two long dialogues is compared to every duration j in the speaker change of the entire film. A matrix is thus populated:

$$SM(i, j) \leftarrow \text{screenplay}(i) \approx \text{speakerchange}(j)$$

In other words, $SM(i,j)=1$ if word i of the average estimated dialogue duration is proportionally the same as the duration of the speaker changes in the dubbed film, and $SM(i,j)=0$ if they are different. Here the term proportionally means that the speaker duration lies within the standard deviation of the estimated dialogue duration. This is because certain languages have longer words on an average, for example, German versus English, however they have to fit into the specific time slot within the scene. Screen time progresses linearly along the diagonal $i=j$, such that when the lines of dialogue from the screenplay line up with speaker durations it is expected that a solid diagonal line of 1's is noted. FIG. 6 shows an example segment of a similarity matrix for the comparison of the estimated durations of the screenplay and for speaker changes. In the similarity matrix estimated durations of the screenplay and of the speaker changes may be characterized according to whether a match is found. Thus every matrix element may be labelled as a mismatch 61 if no match is found or as a match 62 if a match is found. Favourably, a match is further analysed based on the criterion that the best match will follow a track in the similarity matrix. Naturally many matches may be found, but a discontinuous track may also be easily detected and a best path through this track can be established. The words on this best track that do not match may be labelled accordingly 63. Thus, even though the alignment does not follow a diagonal in the similarity matrix it may still be

taken into account in the alignment of the extrinsic and intrinsic information. The final output of this process, method step 27, is shown in FIG. 4b as the alignment points 10.

The fourth embodiment, shown in FIG. 5a, extends that of the third embodiment by additionally performing name spotting in the audio and the extrinsic information. A name spotter unit 51 is adapted to identify names in the intrinsic information known to be important in the film. For example, the extrinsic information can contain character names extracted from the Internet Movie Database directly, or can textually parse the extrinsic information as part of the general extraction of audio related features in the audio related feature extractor 5 of FIG. 5a, or method step 26 of FIG. 2. Such character names are generally not translated even in dubbed films. In case where the names are translated, we rely on the similarity to the original language and the repetitiveness in the movie itself. For example, "John" and the corresponding Italian version of the same name "Giovanni" would appear in the analogous time locations in the movie .The intrinsic information can, for example, be directed through a speech recognition system. The output of which can be analysed for character names. The timestamps of any such character names can be used as further alignment information, or anchor points, for the correlation phase. For situations where the original language screenplay does not contain timestamps, the character names can be used to improve the estimated timestamps accorded to the screenplay. The name spotting process is identified as step 24 in FIG. 2 and the alignment process making use of the extra information is identified as step 28 in the flowchart of FIG. 2. The output of the alignment unit 7 is identified at alignment points 10 in FIG. 5b.

In any of the preceding embodiments performing the known method of face-speech matching can assess the quality of the alignment. Such a method normally operates on video features contained within intrinsic information in the video content. For example, if the face speech matching says that there is a "talking face" but no voice is detected, this information can be used in the estimate of how long a sentence should have been. This information may then be used to compensate for the time a sentence is actually spoken for. This information can also give a measure of the quality of the dubbing and can then be used to recommend a dubbed movie to the viewer. A high quality of dubbing leads directly to a viewer enjoying the movie. Low quality dubbing can detract significantly from the

viewing experience. If it is necessary to constantly overrun or under run dialogues, then a low dubbing quality rating can be assigned.

It will be apparent to a person skilled in the art that the invention may also be embodied as a computer program product, storable on a storage medium and enabling a computer to be programmed to execute the method according to the invention. The computer can be embodied as a general-purpose computer like a personal computer or network computer, but also as a dedicated consumer electronics device with a programmable processing core.

In the foregoing, it will be appreciated that reference to the singular is also intended to encompass the plural and vice versa. Moreover, expressions such as "include", "comprise", "has", "have", "incorporate", "contain" and "encompass" are to be construed to be non-exclusive, namely such expressions are to be construed not to exclude other items being present.

Although the present invention has been described in connection with preferred embodiments, it is not intended to be limited to the specific form set forth herein. Rather, the scope of the present invention is limited only by the accompanying claims.

CLAIMS:

1. A system (8) for alignment of intrinsic and extrinsic audio-visual information, the system comprising:

an intrinsic content analyser (2, 3), the intrinsic content analyser being communicatively connected to an audio-visual source (1), the intrinsic content analyser being arranged to classify intrinsic information (3) extracted from content sourced by the audio-visual source resulting in intrinsic classifications;

an extrinsic content analyser (5, 6), the extrinsic content analyser being communicatively connected to an extrinsic information source (4), the extrinsic content analyser being arranged to classify extrinsic information (6) extracted from extrinsic information sourced by the extrinsic information source resulting in extrinsic classifications; and

an intrinsic information and extrinsic information correlator (7) being communicatively connected to the intrinsic content analyser (2, 3) and to the extrinsic content analyser (5, 6) and being arranged to correlate the intrinsic classifications with the extrinsic classifications, thereby providing a multi-source data structure.

2. The system of claim 1, wherein

the intrinsic content analyser is arranged to classify audio related features (3) extracted from the content sourced by the audio-visual source resulting in intrinsic audio classifications; and

the extrinsic content analyser is arranged to classify audio related features (6) extracted from the extrinsic information sourced by the extrinsic information source resulting in extrinsic audio classifications.

3. The system of claim 2, wherein

the intrinsic content analyser is further arranged to identify a location or duration of at least one of the intrinsic audio classifications (3);

the extrinsic content analyser is further arranged to identify a location or duration of at least one of the extrinsic audio classifications (6); and

the intrinsic information and extrinsic information correlator (7) is arranged

to correlate the intrinsic audio classifications with the extrinsic audio classifications further based upon the location or duration of the intrinsic audio classifications and the location or duration of the extrinsic audio classifications.

4. The system of claim 2, wherein

the intrinsic content analyser is arranged to classify the audio related features (3) extracted from the content sourced by the audio-visual source as silence and/or as speech; and

the extrinsic content analyser is arranged to classify the audio related features (6) extracted from the extrinsic information sourced by the extrinsic information source as silence and/or as speech.

5. The system of claim 2, wherein

the intrinsic content analyser is arranged to identify at least one speaker (31) within the audio related features extracted from the content sourced by the audio-visual source resulting in identified speakers;

the intrinsic content analyser is further arranged to identify a change of speaker (32) within the audio related features extracted from the content sourced by the audio-visual source resulting in identified speaker changes; and

the intrinsic information and extrinsic information correlator (7) is arranged to correlate the intrinsic audio classifications with the extrinsic audio classifications further based upon the identified speakers and the identified speaker changes.

6. The system of claim 5, wherein

the intrinsic content analyser is further arranged to identify a location or duration of at least one of the intrinsic audio classifications;

the extrinsic content analyser is further arranged to provide an estimated location or duration (44) of at least one of the extrinsic audio classifications based upon a duration of the audio related features (6) extracted from the extrinsic information sourced by the extrinsic information source and a duration extracted from the extrinsic information sourced by the extrinsic information source; and

the intrinsic information and extrinsic information correlator (7) is arranged

to correlate the intrinsic audio classifications with the extrinsic audio classifications further based upon the location or duration of the intrinsic audio classifications and the estimated location or duration of the extrinsic audio classifications.

7. The system of claim 1 or 2, wherein

the intrinsic content analyser is arranged to identify intrinsic names (51) contained within the intrinsic classifications;

the intrinsic content analyser is further arranged to provide a location or point in time of the intrinsic names (51);

the extrinsic content analyser is arranged to identify extrinsic names contained within the extrinsic classifications;

the extrinsic content analyser is further arranged to provide a location or point in time of the extrinsic names; and

the intrinsic information and extrinsic information correlator is arranged to correlate the intrinsic classifications with the extrinsic classifications further based upon the location or point in time of the intrinsic names and the location or point in time of the extrinsic names.

8. The system of claim 2 wherein

the audio-video source provides a film; and

the extrinsic information comprises a screenplay of the film.

9. The system of claim 2 wherein

the audio-video source provides a film; and

the extrinsic information comprises a time-stamped screenplay of the film.

10. A method for alignment of intrinsic and extrinsic audio-visual information, the method comprising the steps of:

classifying intrinsic information extracted from content sourced by an audio-visual source resulting in intrinsic classifications (21);

classifying extrinsic information extracted from extrinsic information sourced by an extrinsic information source resulting in extrinsic classifications (26); and

correlating the intrinsic classifications with the extrinsic classifications, thereby providing a multi-source data structure (25).

11. The method of claim 10, further comprising the steps of:

identifying a location or duration of at least one of the intrinsic classifications (21);

identifying a location or duration of at least one of the extrinsic classifications (25); and

correlating the intrinsic classifications with the extrinsic classifications further based upon the location or duration of the intrinsic classifications and the location or duration of the extrinsic classifications (26).

12. The method of claim 10, further comprising the steps of:

identifying at least one speaker within audio related features extracted from the content sourced by the audio-visual source (22) resulting in identified speakers;

identifying a change of speaker within the audio related features extracted from the content sourced by the audio-visual source (23) resulting in identified speaker changes; and

correlating the intrinsic classifications with the extrinsic classifications (27) further based upon the identified speakers and the identified speaker changes.

13. The method of claim 10, further comprising the steps of:

determining an intrinsic feature duration of features extracted from the content sourced by the audio-visual source;

determining an intrinsic content duration of the content sourced by the audio-visual source;

identifying a location or duration of at least one of the intrinsic classifications based upon the intrinsic feature duration and the intrinsic content duration;

determining an extrinsic information duration from the extrinsic information sourced by the extrinsic information source;

determining an extrinsic feature duration of features extracted from the extrinsic information sourced by the extrinsic information source using the extrinsic

information duration;

estimating an estimated location or duration of the extrinsic classifications (29) using the extrinsic information duration and the extrinsic feature duration;

correlating the intrinsic classifications with the extrinsic classifications further based upon the location or duration of the intrinsic classifications and the estimated location or duration of the extrinsic classifications (27).

14. The method of claim 10, further comprising the steps of:

identifying intrinsic names contained within the intrinsic classifications (24);

identifying a location or point in time of the intrinsic names (24);

identifying extrinsic names contained within the extrinsic classifications (26);

identifying a location or point in time of the extrinsic names; and

correlating the intrinsic classifications with the extrinsic classifications further based upon the location or point in time of the intrinsic names and the location or point in time of the extrinsic names (28).

15. A computer-readable recording medium containing a program for controlling an information processing apparatus for alignment of intrinsic and extrinsic audio-visual information, said program enabling said information processing apparatus to perform the method steps of:

classifying intrinsic information extracted from content sourced by an audio-visual source resulting in intrinsic classifications (21);

classifying extrinsic information extracted from extrinsic information sourced by an extrinsic information source resulting in extrinsic classifications (26); and

correlating the intrinsic classifications with the extrinsic classifications, thereby providing a multi-source data structure (25).

16. A program for controlling an information processing apparatus for aligning intrinsic and extrinsic audio-visual information files, said program enabling said information processing apparatus to perform the method steps of:

classifying intrinsic information extracted from content sourced by an audio-visual source resulting in intrinsic classifications (21);

classifying extrinsic information extracted from extrinsic information sourced by an extrinsic information source resulting in extrinsic classifications (26); and

correlating the intrinsic classifications with the extrinsic classifications, thereby providing a multi-source data structure (25)

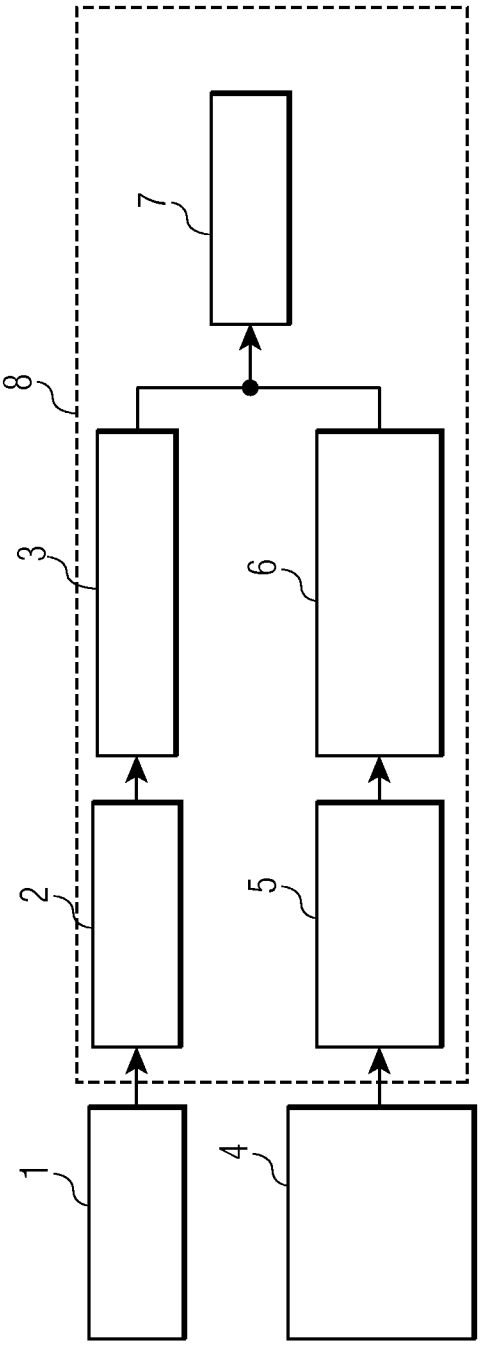

FIG. 1A

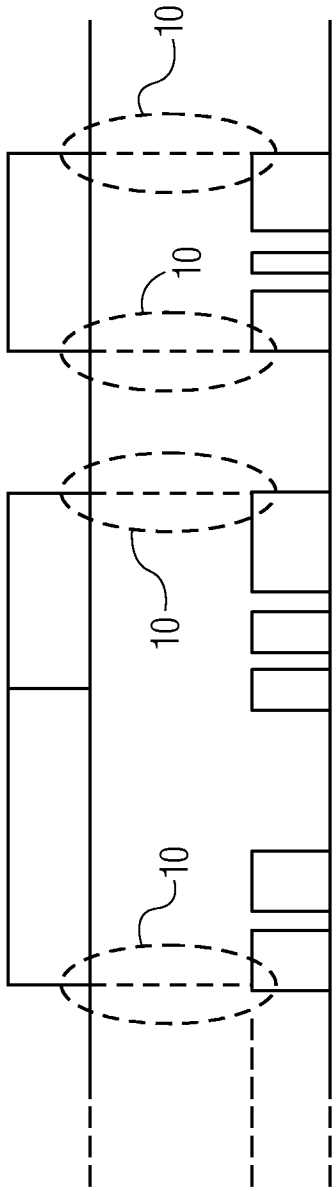

FIG. 1B

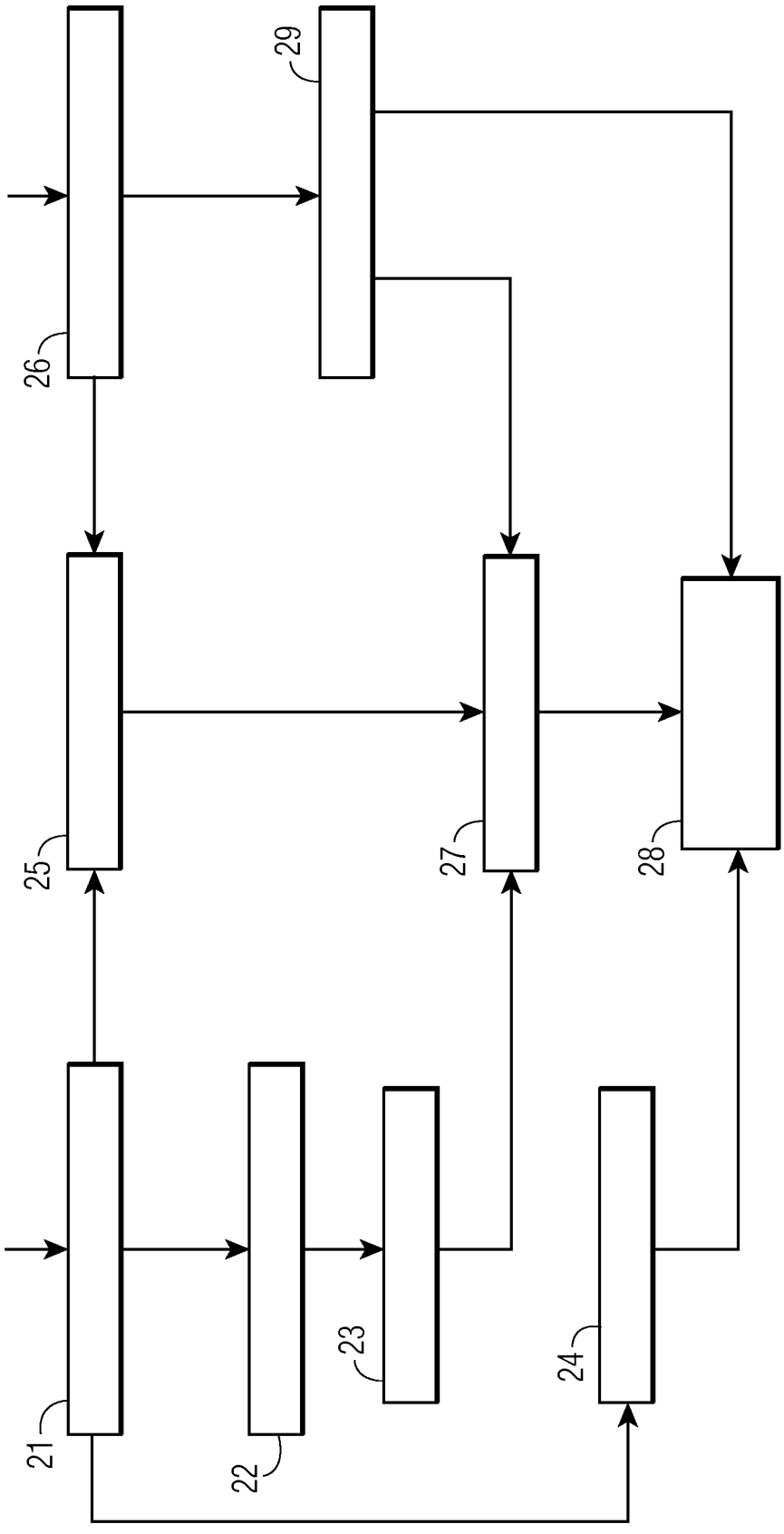

FIG. 2

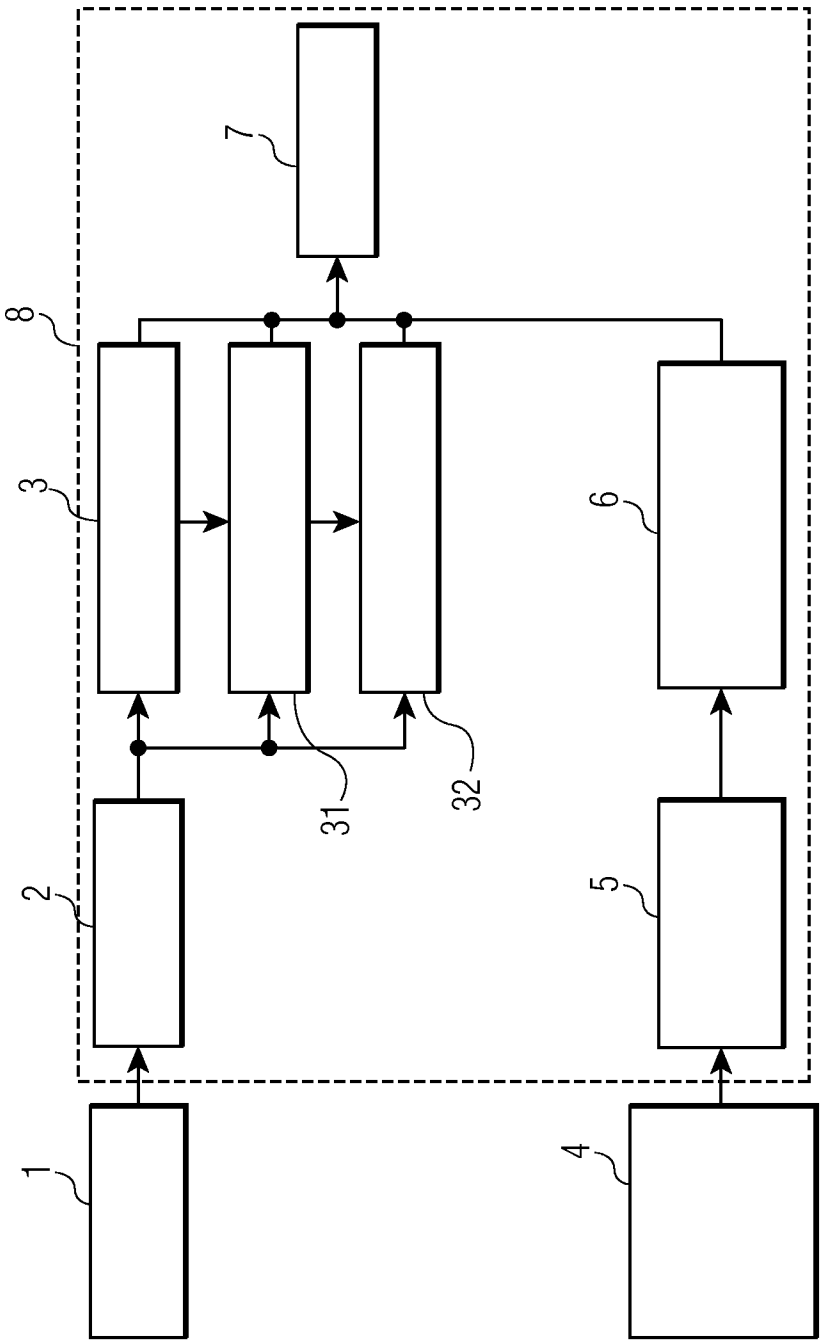

FIG. 3A

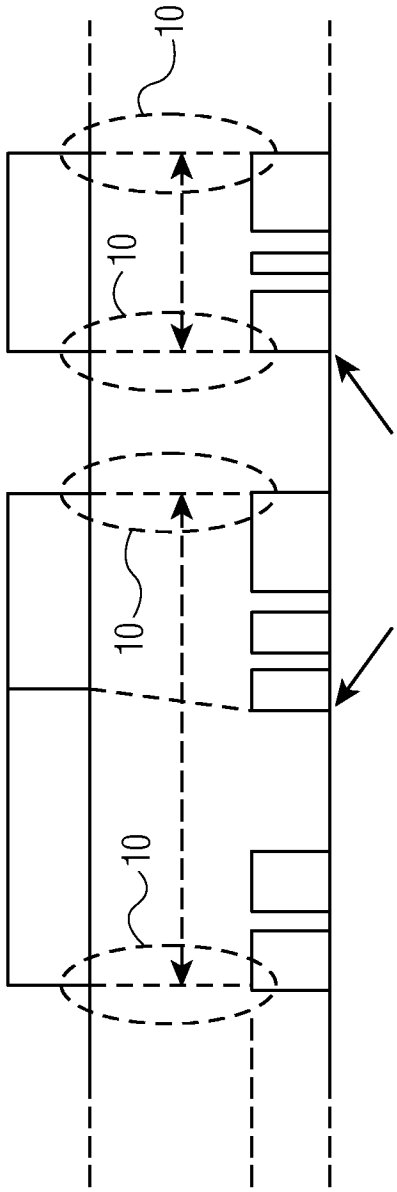

FIG. 3B

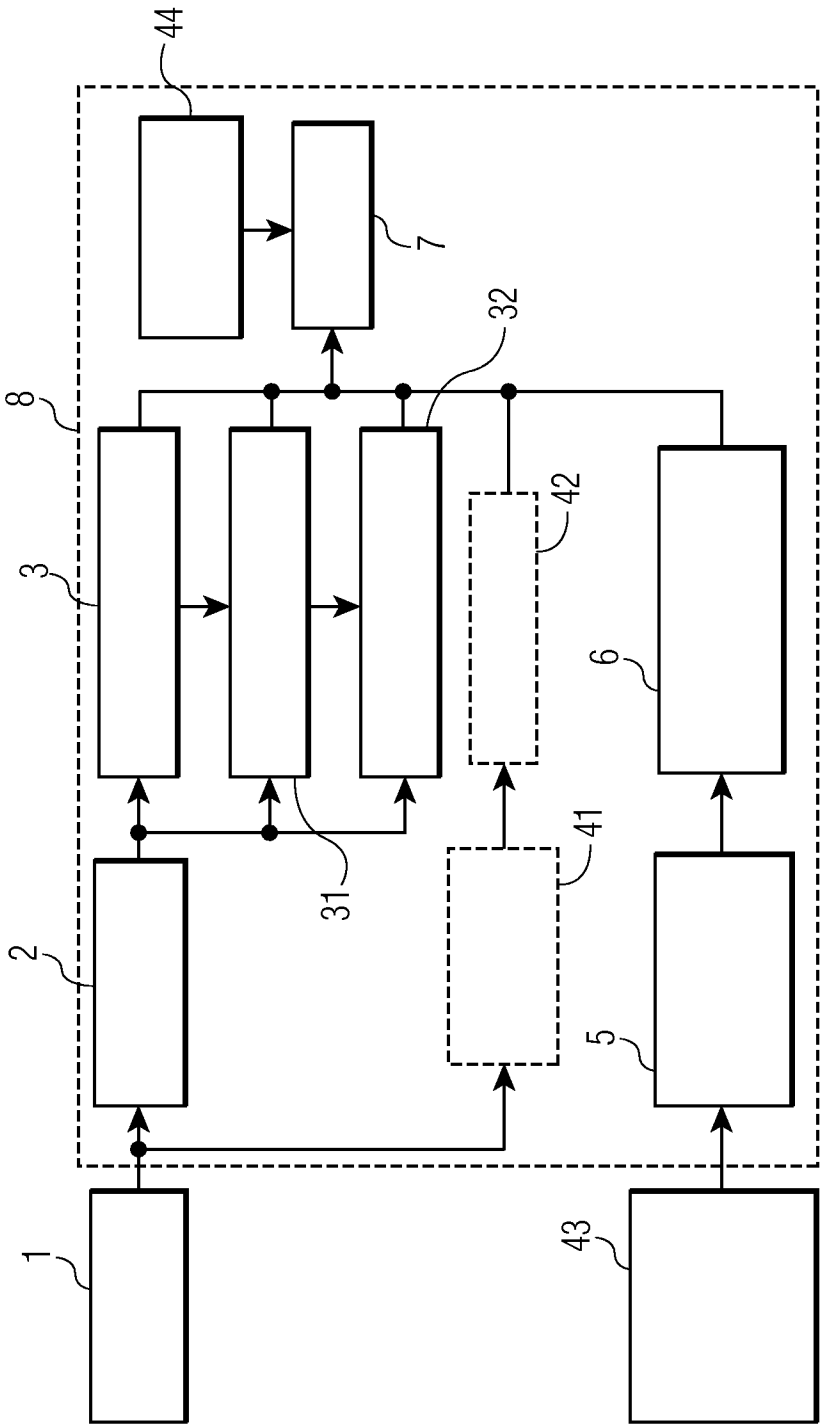

FIG. 4A

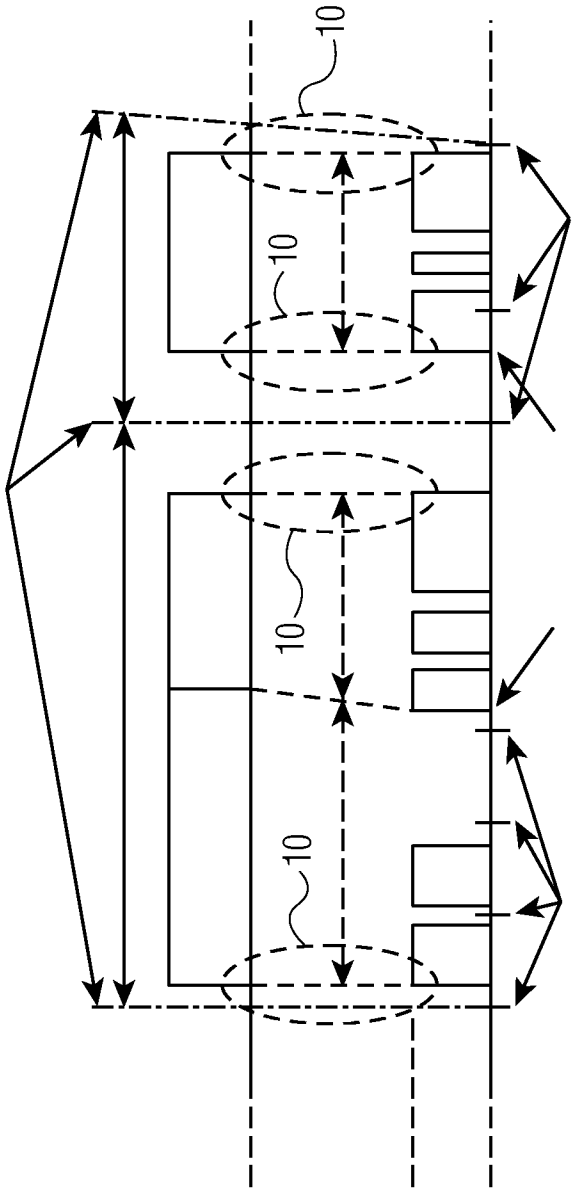

FIG. 4B

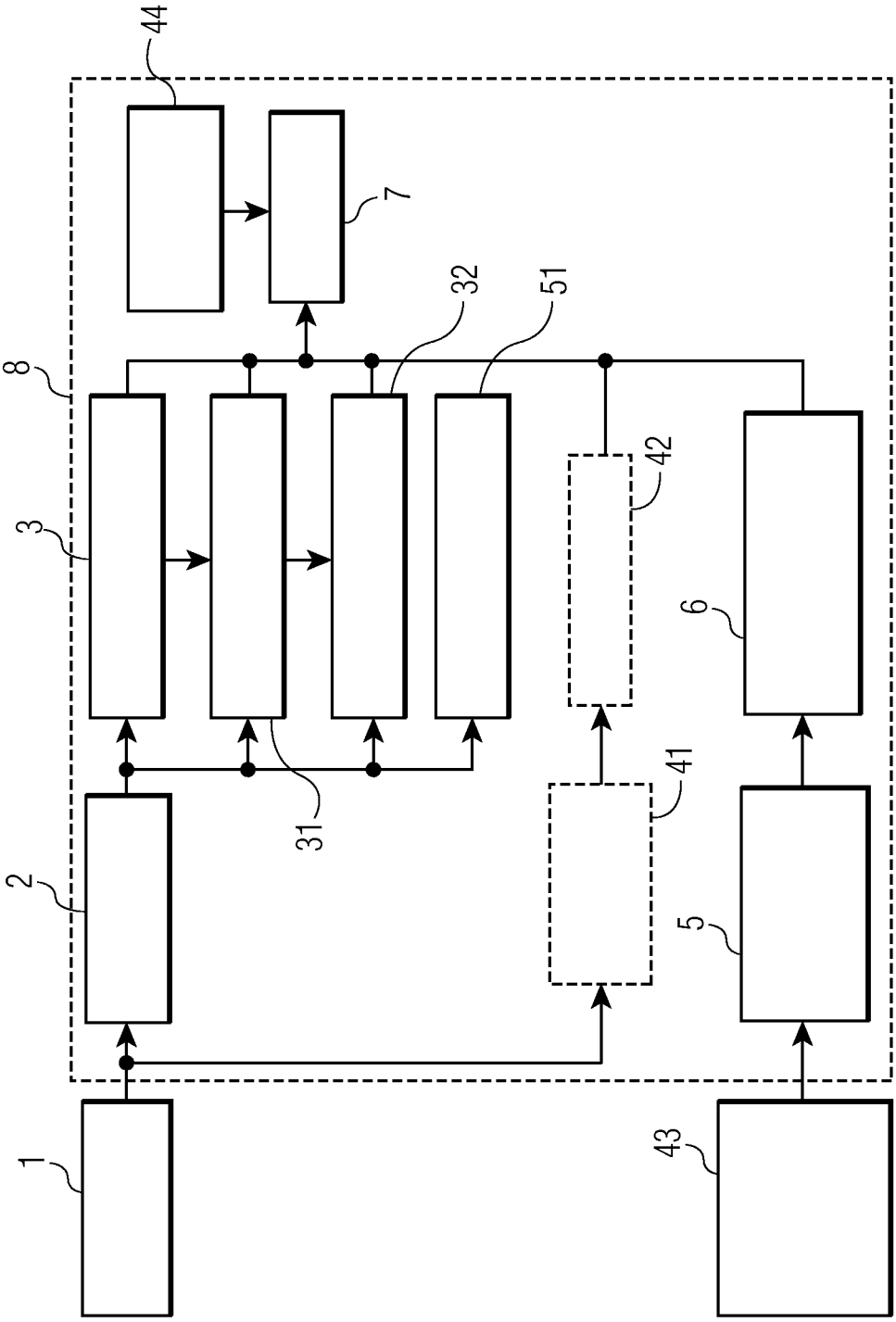

FIG. 5A

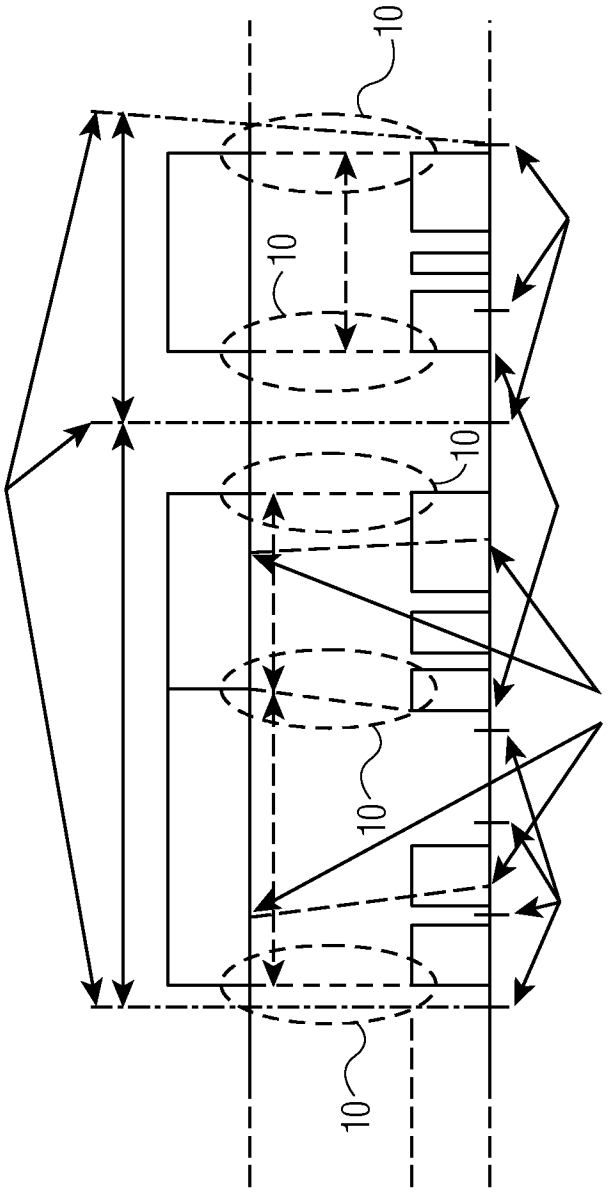

FIG. 5B

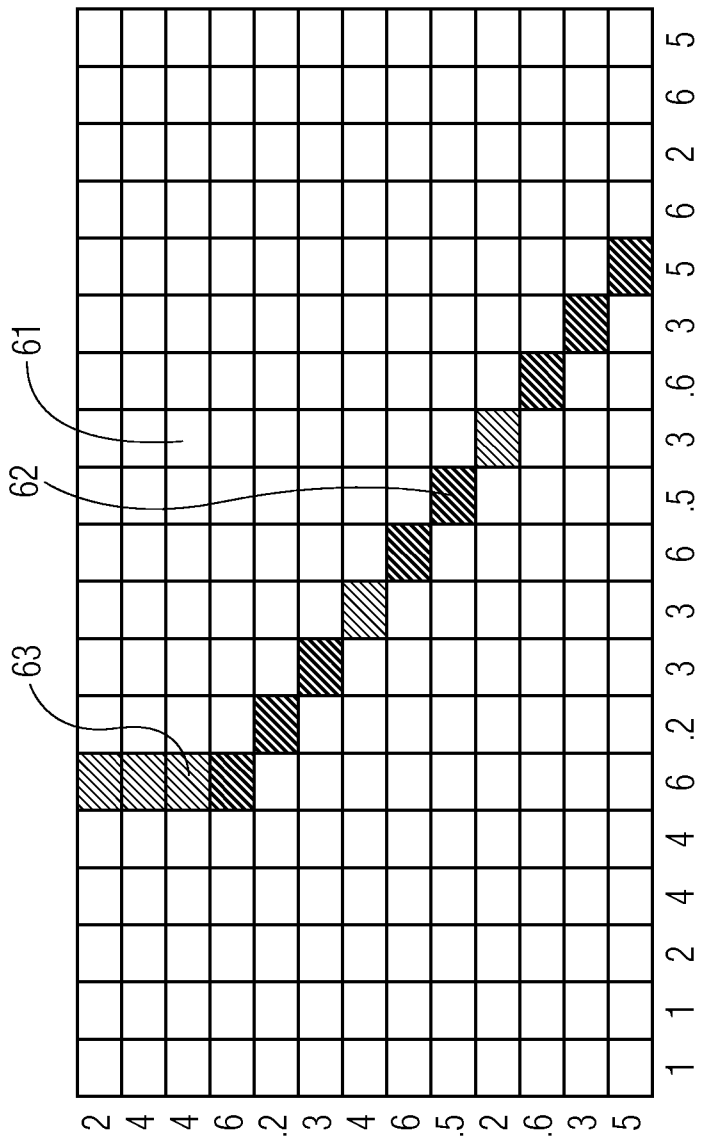

FIG. 6