(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2005-506643
(P2005-506643A)
(43)公表日　平成17年3月3日(2005.3.3)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G11B 27/034	G11B 27/034		5C053
H04N 5/91	H04N 5/91	N	5D110

審査請求 未請求　予備審査請求 有　（全 129 頁）

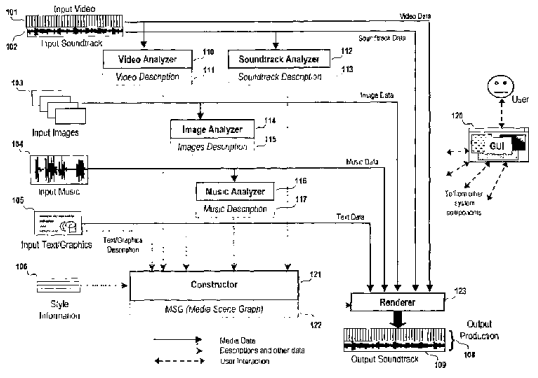
(21)出願番号　特願2002-553775(P2002-553775)
(86)(22)出願日　平成12年12月22日(2000.12.22)
(85)翻訳文提出日　平成15年6月20日(2003.6.20)
(86)国際出願番号　PCT/SG2000/000197
(87)国際公開番号　WO2002/052565
(87)国際公開日　平成14年7月4日(2002.7.4)
(81)指定国　AP(GH,GM,KE,LS,MW,MZ,SD,SL,SZ,TZ,UG,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,CH,CY,DE,DK,ES,FI,FR,GB,GR,IE,IT,LU,MC,NL,PT,SE,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BY,BZ,CA,CH,CN,CR,CU,CZ,DE,DK,DM,DZ,EE,ES,FI,GB,GD,GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NO,NZ,PL,PT,RO,RU,SD,SE,SG,SI,SK,SL,TJ,TM,TR,TT,TZ,UA,UG,US,UZ,VN,YU,ZA,ZW

(71)出願人　503225618
ミュビー　テクノロジーズ　ピーティーイー　エルティーディー
ＭＵＶＥＥ　ＴＥＣＨＮＯＬＯＧＩＥＳ　ＰＴＥ　ＬＴＤ
シンガポール国　シンガポール　１８９６４６　ザ　ベンコーレン　＃０２－２１　ベンコーレン　ストリート　１８０
１８０　Ｂｅｎｃｏｏｌｅｎ　Ｓｔｒｅｅｔ，　＃０２－２１　Ｔｈｅ　Ｂｅｎｃｏｏｌｅｎ，　Ｓｉｎｇａｐｏｒｅ　１８９６４６　（ＳＧ）
(74)代理人　100095577
弁理士　小西　富雅

最終頁に続く

(54)【発明の名称】メディアプロダクションシステムとその方法

(57)【要約】
【課題】出力制作物を生成するために入力データを自動的にまたは半自動的に編集する編集システムが提案される。入力素材は、入力素材を記述しかつ入力素材から入手されるメディア記述子のセットによって注が付けられるかまたはこのメディア記述子のセットを入手するために分析される。編集のスタイルは、そうしたい場合にはユーザから入手されるスタイル・データにより制御される。入力素材は、モーション・ビデオと静止イメージと音楽と音声と音響効果とアニメ化されたグラフィックスおよびテキストとのうちの１つまたはそれ以上を含むことができる。スタイル・データおよび記述子は、入力データに対して実行された場合、編集済みの出力制作物を生成する作業のセットを生成するために使用される。



【特許請求の範囲】
【請求項１】
出力データを形成するために入力データを編集するための方法であって、前記入力データおよび出力データの両方が少なくとも１つの視覚的データおよびオーディオ・データを含み、該方法が、
前記入力データの複数の領域のそれぞれを特徴付ける１つまたはそれ以上の記述子を生成するために前記入力データを分析するステップと、
前記入力データの編集を制御するためにスタイル情報を定義するステップと、
前記入力データに対して行う編集作業のセットを指定する編集決定のセットを生成するために、（ｉ）前記入力データ、（ｉｉ）前記記述子および（ｉｉｉ）前記スタイル情報を使用するステップと、
前記入力素材に対して前記編集作業のセットを実行することにより前記出力データを生成するステップと、
を含む、方法。
【請求項２】
前記記述子に、外部ソースから受信した前記編集決定のセットを生成する前記ステップで使用する追加の予め生成した記述子を供給するステップを含む、請求項１に記載の方法。
【請求項３】
前記追加記述子が、前記入力データを記録したときの計測により生成した記述子を含む、請求項２に記載の方法。
【請求項４】
前記追加記述子が、マニュアル的に生成した記述子を含む、請求項２または３に記載の方法。
【請求項５】
前記追加記述子が、音楽制作物中に生成した音楽記述子を含む、請求項２、３または４に記載の方法。
【請求項６】
出力データを形成するために入力データを編集するための方法であって、前記入力データおよび出力データの両方が少なくとも１つの視覚的データおよびオーディオ・データを含み、該方法が、
外部ソースから、前記入力データの複数の領域のそれぞれを特徴付ける１つまたはそれ以上の予め生成した記述子を受信するステップと、
前記入力データの編集を制御するためにスタイル情報を定義するステップと、
前記入力データに対して行う編集作業のセットを指定する編集決定のセットを生成するために、（ｉ）前記入力データ、（ｉｉ）前記記述子および（ｉｉｉ）前記スタイル情報を使用するステップと、
前記入力素材に対して前記作業のセットを実行することにより前記出力データを生成するステップと、
を含む、方法。
【請求項７】
前記出力データが、関連するサウンドトラックをモーション・ビデオ・データに加えたものを含む、先行する請求項の何れかに記載の方法。
【請求項８】
前記出力データが、関連するサウンドトラックを一連のイメージに加えたものを含む、先行する請求項の何れかに記載の方法。
【請求項９】
前記作業のセットが、次のタイプ、すなわち、細分化、選択的導入、順序付け、変形または結合のうちの少なくとも１つの作業を含む、先行する請求項の何れかに記載の方法。
【請求項１０】
前記入力データが視覚的データを含み、前記変形作業が、前記入力データにより定義され

たイメージの１つまたはそれ以上の部分の色の修正を含む、請求項９に記載の方法。
【請求項１１】
前記変形作業が、前記入力素材の１つまたはそれ以上の部分の再生速度の修正を含む、請求項９または１０に記載の方法。
【請求項１２】
前記結合作業がビデオ遷移を含む、請求項９乃至１１の何れかに記載の方法。
【請求項１３】
前記スタイル情報を定義するステップが、前記入力データの前記記述子に基づいて、複数の予め定義されたスタイル情報のセットの１つを選択することにより実行される、先行する請求項の何れかに記載の方法。
【請求項１４】
前記スタイル情報が、前記出力データ内に挿入された前記入力データのセグメントの持続時間に影響を与える好適なセグメント持続時間パラメータを含む、先行する請求項の何れかに記載の方法。
【請求項１５】
前記スタイル情報が各記述子に対する１つまたはそれ以上の対象値を含み、前記作業のセットを生成する前記ステップが、前記出力データに挿入するために、ａ）前記１つまたは複数の対象値およびｂ）前記各領域に対する前記記述子の近接度の計算に従って前記入力データの複数の領域の１つまたはそれ以上を選択するステップを含む、先行する請求項の何れかに記載の方法。
【請求項１６】
前記計算が、前記入力データの前記各領域の前記記述子値の正規化を含む、請求項１５に記載の方法。
【請求項１７】
前記計算が前記記述子の加重を使用し、それにより、いくつかの記述子が他の記述子よりも前記計算において有意なものになる、請求項１６に記載の方法。
【請求項１８】
前記出力データの領域の順序が、前記入力データの対応する領域の前記入力データの順序と等しいか少なくとも相関している、先行する請求項の何れかに記載の方法。
【請求項１９】
前記スタイル情報が、前記出力データ内の位置に関連する位置データを含み、該位置データが、前記関連する位置のところに前記出力データを生成する前記作業のセットを生成するために使用される、先行する請求項の何れかに記載の方法。
【請求項２０】
前記位置データが複数のデータ・セクションを含み、各データ・セクションが、前記出力データの１つまたはそれ以上のセクションと関連し、前記出力データの各セクションまたは複数のセクションを生成する前記作業のセットを生成するために使用される、請求項１９に記載の方法。
【請求項２１】
前記位置データが、前記出力データ内の位置の関数として変化する少なくとも１つのパラメータを含み、それにより、前記編集決定が、該決定により影響を受ける前記セクションの前記出力データ内の位置により影響を受ける、請求項２０に記載の方法。
【請求項２２】
前記位置データが、前記出力データ内の位置により周期的に変化する、請求項２１に記載の方法。
【請求項２３】
前記スタイル情報が、確率分布から生成されたデータを含む、先行する請求項の何れかに記載の方法。
【請求項２４】
ユーザから、前記入力データの１つまたはそれ以上の要素を識別し、前記各要素に対して

該要素を前記出力データ内に編集する方法の１つまたはそれ以上の態様を指定するマニュアル的入力を受信するステップをさらに含む、先行する請求項の何れかに記載の方法。
【請求項２５】
ユーザから、前記出力データのセグメントを置換すべきことを指定し、この置換が行われる修正した出力データを生成する目的で修正作業のセットを生成するために前記作業のセットを修正するマニュアル的入力を受信するステップをさらに含む、先行する請求項の何れかに記載の方法。
【請求項２６】
前記出力データの前記セグメントに似ている前記入力データのユーザ・セグメントを置換すべきことを示唆するために前記記述子を使用し、それにより、ユーザが、前記出力データのこれらのセグメントを前記入力データのこれらのセグメントで置換することを決定するステップをさらに含む、請求項２５に記載の方法。
【請求項２７】
ユーザから、前記出力制作物の音楽内の特定の時間と整合すべき時間的に重要な視覚的イベントを示す入力を受信し、前記記述子を用いて前記整合を行うステップをさらに含む、先行する請求項の何れかに記載の方法。
【請求項２８】
前記作業のセットを表す、本質的にツリーの構造を有するデータ構造を発生するステップをさらに含む、先行する請求項の何れかに記載の方法。
【請求項２９】
ユーザにデータ構造を表示し、対応する前記作業のセットを修正するためにユーザから前記データ構造の領域を示す入力を受信するステップをさらに含む、請求項２８に記載の方法。
【請求項３０】
ユーザが、一時的に修正が行われないデータ構造の領域を示す、請求項２９に記載の方法。
【請求項３１】
前記記述子が、前記入力データの複数の要素のそれぞれに対する人間確率記述子を含み、該人間確率記述子が、前記入力素材の各要素内に人間が存在する確率を表し、作業のセットを生成する前記ステップが、前記人間確率記述子の値が高い前記入力データの要素が前記人間確率記述子の値が低い要素よりも前記出力データ内にもっと頻繁に挿入される作業を生成する、先行する請求項の何れかに記載の方法。
【請求項３２】
前記記述子が、移動するイメージ・データを表す前記入力データの複数の移動するイメージ要素のそれぞれに対する少なくとも１つのカメラ移動記述子を含み、該カメラ移動記述子が、その要素が収集された場合に、各要素に対して前記要素を収集したカメラが移動した程度を表し、作業のセットを生成する前記ステップが、前記カメラ移動記述子の値が低い前記入力データの要素が前記カメラ移動記述子の値が高い要素よりも前記出力データ内にもっと頻繁に挿入される作業を生成する、先行する請求項の何れかに記載の方法。
【請求項３３】
ユーザから、スタイル情報を定義する前記ステップを実行するために、前記入力データを決定し、該決定のセットを生成する前記ステップおよび前記出力データを生成する前記ステップをスタートさせるために、信号を受信する予備ステップをさらに含む、先行する請求項の何れかに記載の方法。
【請求項３４】
前記出力データが少なくとも１つのオーバレイを含み、該オーバレイが少なくとも１つのテキストおよびグラフィックスを含む、先行する請求項の何れかに記載の方法。
【請求項３５】
オーバレイがアニメ化される、請求項３４に記載の方法。
【請求項３６】

前記入力データが音楽を含み、前記オーバレイのアニメーションの少なくとも１つのパラメータが、前記音楽の特徴を表す音楽記述子により決定される、請求項３５に記載の方法。

【請求項３７】
前記スタイル情報を定義する前記ステップ、前記作業のセットを生成するステップおよび前記出力データを生成するステップのうちの少なくとも２つが、異なる別の場所にいるユーザによりスタートされる、先行する請求項の何れかに記載の方法。

【請求項３８】
前記スタイル情報を定義する前記ステップおよび前記決定のセットを生成するステップが第１のユーザにより実行され、前記決定のセットが前記入力データにアクセスまたはそのコピーにより装置を操作している第２のユーザに送信され、該第２のユーザが前記セットを用いて前記出力データを生成する前記ステップをスタートし、それにより、前記第２のユーザが、前記第１のユーザから前記第２のユーザにメディア・データを送信しなくても、前記第１のユーザが生成した出力データをチェックすることができる、先行する請求項の何れかに記載の方法。

【請求項３９】
前記記述子が前記入力データの少なくとも一部の短いセクションと関連するミクロ記述子を含み、該ミクロ記述子が、前記スタイル情報と組合わせてまたは該スタイル情報打ち消すように、前記入力データの対応するセクションに関連する前記編集作業を生成するステップにおいて使用する編集ヒントを入手するために使用される、先行する請求項の何れかに記載の方法。

【請求項４０】
前記入力データが音楽データを含み、前記ミクロ記述子が音楽バーまたはもっと短い時間の尺度上の前記音楽のセクションに関連する、請求項３９に記載の方法。

【請求項４１】
前記入力データが音楽データを含み、前記記述子が音楽の完全な一部を記述しているマクロ記述子を含み、前記音楽データ上で実行される前記作業のセットが、前記マクロ記述子を用いて選択した前記スタイル情報の領域を用いて生成され、ミクロ記述子が前記音楽の一部のセクションを記述する、請求項１乃至３８の何れかに記載の方法。

【請求項４２】
１つまたはそれ以上の前記作業のセットが、前記ミクロ記述子から入手した時間により変化する値のセットに、前記スタイル情報により支配される時間に依存するしきい値を適用することにより決定される、請求項４０または４１に記載の方法。

【請求項４３】
前記作業が、第１のメディア・タイプに関連する前記入力データ内のデータに対して実行されるべき作業を含むとともに、第２のメディア・タイプに関連する前記入力データ内のデータに依存して入手される、先行する請求項の何れかに記載の方法。

【請求項４４】
前記第１のメディア・タイプがモーション・ビデオであり、前記第２のメディア・タイプが音楽である、請求項４３に記載の方法。

【請求項４５】
前記入力データがモーション・ビデオおよび音楽に関連するサウンドトラックを含み、前記作業のセットが、
そのオーディオ特徴に従ったサウンドトラックの一部の選択、
音楽記述子の値に従った前記サウンドトラックの一部内でミックスする時間の決定、および
前記サウンドトラックの一部がミックスされる場合の前記音楽の音量の低減
のうちの少なくとも１つを実行するために前記サウンドトラックの一部を前記音楽とミックスする、先行する請求項の何れかに記載の方法。

【請求項４６】

コンピュータ装置が読むことができ、該コンピュータ装置に先行する請求項の何れかに記載の方法を実行させるプログラム命令を含む記録媒体のようなコンピュータ・プログラム製品。

【請求項４７】

出力データを形成するために入力データを編集するための編集システムであって、前記入力データおよび前記出力データの両方が視覚的データおよびオーディオ・データのうちの少なくとも一方を含み、該システムが、

前記入力データの複数の領域それぞれを特徴付ける１つまたはそれ以上の記述子を生成するために前記入力データを分析するための分析手段と、

前記入力データの編集を制御するためにスタイル情報を定義するためのスタイル定義手段と、

前記入力データに対して実行すべき編集作業を指定する１つまたはそれ以上の編集決定のセットを生成するために（ｉ）前記入力データ、（ｉｉ）前記記述子および（ｉｉｉ）前記スタイル情報を使用するための構成手段と、

前記入力素材に対して前記作業のセットを実行することにより前記出力データを生成するためのレンダリング手段と、

を含む、編集システム。

【請求項４８】

出力データを形成するために入力データを編集するための編集システムであって、前記入力データおよび前記出力データの両方が視覚的データおよびオーディオ・データのうちの少なくとも一方を含み、該システムが、

前記入力データの複数の領域のそれぞれを特徴付ける１つまたはそれ以上の記述子を受信するための手段と、

前記入力データの編集を制御するためにスタイル情報を定義するためのスタイル定義手段と、

前記入力データに対して行う編集作業を指定する１つまたはそれ以上の編集決定のセットを生成するために（ｉ）前記入力データ、（ｉｉ）前記記述子および（ｉｉｉ）前記スタイル情報を使用するための構成手段と、

前記入力素材に対して前記作業のセットを実行することにより前記出力データを生成するためのレンダリング手段と、

を含む、編集システム。

【発明の詳細な説明】

【０００１】

発明の技術分野

本発明は一般的にコンピュータで生成されるメディア制作物（ｍｅｄｉａ　ｐｒｏｄｕｃｔｉｏｎ）に関する。特に、この発明は、動画ビデオ、静止画イメージ、音楽、スピーチ、音響効果、アニメグラフィックス及びテキストの一つ又は２以上を含む制作の全自動若しくは半自動編集に関する。

【０００２】

発明の背景

現在、アナログメディアは徐々にデジタルメディアに置きかえられつつある。オーディオの分野において、この移行は既に大きく起こっている。そして、この移行はイメージ、ビデオ、グラフィックスアニメーション及び他のメディアでも進行中である。これらのメディアがデジタル化するにつれ、コンピュータ資源における容量／コストの比は増加の一途をたどる。デジタルメディア制作物に関する新しいユーザとマーケットが開拓されつつある。この発明の特に関連するところは簡易メディア制作物、特に簡易ビデオ制作物の新規マーケットを開拓することにある。つまり、専門的ビデオ制作物の使用は除かれ、最近までは必要な機材のコストは非常に高かった。これには、ホームビデオ制作物（例えば、休暇、結婚式等）、非公式組織の使用（例えば、内部通信やチーム作り）、社会や他の組織による使用等が含まれる

【０００３】
簡易や“デスクトップ”という概念は１０年ほど前から知られていたが、種々の課題から広く導入されるには至らなかった。理由として、
１．技術的インフラの課題：カメラからビデオをデジタル化するときの不便さや品質の低下、ハードディス容量の制限、不充分な処理能力、等
２．簡便で低コストの流通機構の欠如：最近まで汎用のフォーマットはビデオテープしかなかったが、重複や流通に関連するコストや時間が多くのアプリケーションの可能性を不可能としてきた。
３．時間や専門知識が、特に編集及び制作前段階において、満足にたる品質の制作物を作るのに要求されていた。
第１及び第２のこれらの問題は現在ではＤＶカメラ、ＩＥＥＥ　１３９４（“ファイヤワイヤ”）インターフェース及びワールドワイドウエブ上のビデオ配信によりなくなっている。

【０００４】
この発明は第３の問題の解決を目的とし、自動若しくは半自動のデジタルメディア、特にビデオの編集にある。

【０００５】
今日、ビデオ編集に使用される主たるツールは“非線形ビデオ編集（Ｎｏｎ－Ｌｉｎｅａｒ　ｖｉｄｅｏ　Ｅｄｉｔｏｒ”）若しくはＮＬＥである。これらはコンピュータプログラムであって、フィルムカットやビデオテープ装置を用いた線形ダビング編集などの従来型の編集方法から手法を採用している。これらは編集のためのマニュアル的な手法を用いており、それによればユーザが経験してその結果が高品質なビデオ制作物となるようなシナリオに適合する。このようなタイプの製品は多数存在し、アドブ社のＰｒｅｍｉｅｒｅやアップル社のｉＭｏｖｉｅを挙げられる。

【０００６】
ＮＬＥは初期のテクノロジーにおいて多大な進展をみせている。しかしながらいまだ多くの問題がのこっている。その問題とは、ユーザがメディアの専門家ではないこと、専門家の要求する品質が必ずしも要求されていないこと、若しくは素材の編集を素早く行う必要のあることである。非専門家に適用するようにされたＮＬＥであったとしても、それは満足な制作を得るのにかなり深刻な学習曲線と多大な時間が要求される。ユーザは１分の出力ビデオを形成するのに１時間を費やすことが一般的な許容範囲と考えられている。即ち、制作時間とプレイバック継続時間との間には６０：１の割合が許容されている。この発明の目的は自動化によりこの割合を劇的に小さくすることにあり、この点により、ユーザ自身に何ら手間をかけさせること無く満足のいく結果が得られる場合がある。

【０００７】
実時間の画像表示や音声進行に同期したテキスト文などユーザが制作物を創造することを許容する既存のツールが幾つか存在する。これらはアニメーションツール（例えばマイクロメデイア社のＦｌａｓｈ等）、スライドショウツール（マイクロソフト社のＰｏｗｅｒＰｏｉｎｔ等）及びストリーミングメデイアのための著作ツール（リアルネットワーク社のＲｅａｌＰｌａｙｅｒ等）。しかしながら再度考えるべきことは、数分間の単純な制作物を形成するのに数時間が必要なことを時としてユーザは知ることになる。

【０００８】
発明の概要
この発明は入力されるメディア素材からメディア制作物を形成する新規有用な装置と方法を提供することを目的とする。
一般的にいえば、この発明より入力された素材が編集されて出力制作物が構築される。このプロセスはメディア記述子のセットを派生させ、このメディア記述子は入力された素材を表現する。この入力された素材は分析により若しくは外部ソースから提供されることより又はその両者により得られる。このことに引き続きコンピュータによる構築プロセスが生じる。この構築プロセスは（ｉ）メディア記述子のセットにより、及び（ｉｉ）編集ス

タイルを決めるための、例えばユーザが決めるスタイルデータのような、スタイルデータに基づく編成決定の形成を含む。

【０００９】

インプットされた素材はモーションビデオ、静止画像、音楽、スピーチ、音響効果、アニメ画像及びテキストの少なくともひとつを含む。

メディア記述子のセットは予め形成されるか（例えば、この発明品の外部において）若しくはインポートされたデスクリプタにより、例えば入力された素材と一緒になって補充されることがある。

【００１０】

決定論的な及び確率論的な（蓋然論的な）処理の一方若しくは両方を含むプロセスによりスタイルデータは形成される。

編集は入力された素材に適用される下記のプロセスの少なくとも１つを含むことがある。そのプロセスとは、（ビデオ／音声）の細分化、選択的導入、順序付け、変形と結合である。これらのプロセスはときとしてユーザの介入により補足される。これは２つのステージよりサポートされている。一つは自動的な構築プロセスに先立つ事前選択ステージであり、他の一つは構築後のタッチアップステージである。

【００１１】

この発明の極めて好ましい局面は音楽による制作物を形成することにある。ここにおいて、入力される素材はａ）モーションビデオ素材及び／又は画像のセット、及びｂ）録音された音楽である。この発明の装置はビデオ／画像と音楽の両者を分析し、それぞれのためのメディア表現データを形成する。そしてこの情報を利用して出力制作物を形成し、この出力制作物は音楽の構成により影響され若しくは決められる。

【００１２】

本発明の標準的な応用は家庭用、企業用、趣味に生きる人々用のビデオやその他の時間によるメディアの制作物、音楽に同期したスライドショウの制作物、リッチメディア電子グリーティングカードの制作物、ＷＷＷ用のメディアの制作物、リッチメディアオンラインカタログの制作物、消費者間の取引アプリケーションに関するリッチメディアオンラインコンテンツの制作物を含む。なお、消費者間の取引アプリケーションにはオンラインアクション、分類された宣伝、カラオケビデオ等の制作物のような専門的なビデオアプリケーションの幾つかが含まれる。

【００１３】

この発明は、方法と装置の局面をともに含んでいるが（即ち、方法のステップを実行する各手段から装置が構成される）、種々のハードウエアの中で具現化することができる。このハードウエアには一般目的用のコンピュータ、個人用デジタル補助具、専用のビデオ編集ボックス、セットトップボックス、デジタルビデオレコーダ、テレビジョン、ゲームコンソール、デジタル静止画カメラ、デジタル動画カメラが含まれている。

【００１４】

この発明の実施例を、例示のためのみの図面を参照して以下に説明する。

【００１５】

実施例の詳細な説明

図１はこの発明の実施例の全体構造を示す。

図１を参照にして、実施例の装置に入力される素材は次のものの１又は２以上を含む。

“入力ビデオ”［１０１］、即ち、デジタルビデオストリームや１以上のデジタルビデオフィルムのようなモーションビデオ。典型的には、カメラやビデオカメラにより捕らえられたビデオのような何ら編集されていない“生の映像”である。これはまたサウンドトラック［１０２］を含むことができる。

“入力イメージ”［１０３］、即ち、デジタル画像フィルムのような静止イメージ。これらはモーションビデオの代わりに、若しくはモーションビデオに追加して使用される。

デジタル音声ストリームや１又は２以上のデジタル音声フィルムのような様式内の“入力音楽”［１０４］。この実施例においては、音楽は出力制作物のためにタイミングとフレ

ームワークを提供する。入力視覚素材は種々の方法で編集される。この編集の方法は音楽による制作物を形成するための音楽の構成に関連する。
入力テキスト及び／又はグラフィックス［１０５］は典型的にはタイトル、クレジット、サブタイトル等に用いられる。
“スタイル情報”［１０６］、即ち、装置により使用されるデータや論理であって自動生成プロセスの特性を制御し若しくは影響を与える。換言すれば、“編集スタイル”を指す。ユーザは所定のスタイルの中から選択することができるし、及び／又は個々のスタイルのパラメータにアクセルすることもできる。実施例においては、スタイルは装置の外にあり又は装置の一部をなす。
【００１６】
この明細書において“入力素材”なる用語は装置に入力されるメディアの１又は２以上のものを意味する。支持されるメディアのタイプはモーションビデオ、静止イメージ、音楽、スピーチ、音響効果、静止した若しくはアニメのグラフ、及び静止した若しくはアニメのテキストを含む。“入力視覚素材”なる用語はビデオ、イメージ、アニメーション、グラフィック若しくはテキストのような視覚型の入力素材をいう。
【００１７】
出力
図１を参照して、装置により形成された“出力制作物”［１０８］はビデオ、アニメーション、若しくはイメージの時間列のようなものであり；これは、補助的なサウンドトラック、出力サウンドトラック［１０９］、音楽の形成、スピーチ及び／又は他の音を含むことができる。この出力制作物は入力素材の全部又は一部により形成され、ここにおいて入力素材は以下に説明する１又は２以上のプロセスにより装置により処理される。
【００１８】
“細分化（Ｓｅｇｍｅｎｔａｔｉｏｎ）”。即ち、入力ビデオは視覚的な若しくは聴覚的な特徴に基づいて細分化される。例えば、無作為なショット、ショットの部分、特別な声若しくは背景音を含むセグメントなどである。ショットとはビデオの連続体であり、中断やカットを持たず、中止やポーズをすることなくビデオカメラに収録されたビデオの一つのセグメントなどを指す。
“選択的導入”（Ｓｅｌｅｃｔｉｖｅ　ｉｎｃｌｕｓｉｏｎ）。即ち、ビデオのセグメント、音楽若しくはサウンドトラック、選択されたイメージ、若しくはイメージやビデオフレーム中の領域のような入力素材の要素が出力制作物に含まれる。他方、他のものは含まれない。典型的には、従来的なメディア制作物として、多数のものが除かれる。
【００１９】
“順序付け（Ｓｅｑｕｅｎｃｉｎｇ）”。入力素材の要素を順序付けすることができ、その結果、出力制作物を有する要素の時間配列は入力素材中のそれらの要素の時間配列に対応し、またはデスクリプタの相同性のような他の要件に基づいてそれらは順序付けされる。
“変形（Ｔｒａｎｓｆｏｒｍａｔｉｏｎ）”。入力素材の要素を変形することができる。例えば、周知の“特殊効果”、色の変化（例えばモノクロ及びフラッシュ効果）、速度（例えば、スローモーション）、大きさ（例えば人工的な拡大）、位置（例えば人工的なパン撮り）、形状（例えばラッピング）等を含むプロセスがある。
“結合（Ｃｏｍｂｉｎａｔｉｏｎ）”。入力素材は同時にまた順番に結合される。例えば、入力素材からのイメージとビデオセグメントは入力音楽と同時に提供することができ、また、入力テキスト／グラフィックスはそのビデオに重ねることができる。イメージとビデオのセグメントは結び付けて重ね合わせることができる。これは周知のワイプ（ｗｉｐｅｓ）やディソルブ（ｄｉｓｓｏｌｖｅｓ）のような移行（ｔｒａｎｓｉｔｉｏｎｓ）の使用を許容する。入力サウンドトラックのセグメントは入力音楽のセグメントにミックスすることができる。マルチのイメージ及び／又はビデオセグメントは同時に出力制作物のフレームの異なる領域に提供することができる。また、それらは混合されて合成画像（“ミキセージ（ｍｉｘａｇｅ）”）を提供する。

【００２０】
出力制作物は入力素材を参照せずにその装置により形成された素材を含むことができる。背景として用いられる色や風合い、静止画及びアニメグラフ要素等である。
【００２１】
構成要素の分析と説明
図１を参照して、実施例は下記の構成要素を有し、この構成要素は入力素材の分析と説明に関連している。
ビデオアナライザ［１１０］。これは入力ビデオを分析し、１又は２以上の記述子（デスクリプタ、（ｄｅｓｃｒｉｐｔｏｒ））を含むビデオ記述（Ｖｉｄｅｏ Ｄｅｓｃｒｉｐｔｏｒ）［１１１］形成する。このビデオアナライザは信号分析技術やマルチフレームや個別フレームに対する他の種類の処理方法を応用し、記述子を形成する。典型的な記述子は輝度やカラーヒストグラムなどの色合いの指標、テクスチャーの指標、形状の指標、モーションアクティビティの指標、ショット時間、入力ビデオ中の他のセグメントの境界を規定する記述子、分類別の相似性指標（例えば、入力ビデオの１つのセグメントが人の顔を含む可能性、自然の情景である可能性等）、他の記述子の統計的特性と変化割合の指標、２若しくはそれ以上のデスクリプタを組み合わせることにより形成された記述子、等である。これら多くの記述子とテクニックは当業者によく知られており、新しいものは継続的に定義される。
【００２２】
サウンドトラックアナライザ［１１２］。これは入力ビデオの入力サウンドトラックを分析し、１又は２以上の記述子を含む１つのサウンドトラック記述［１１３］を形成する。このサウンドトラックアナライザは信号分析テクニックや入力サウンドトラックに対する他の種類の処理を適用し、記述子を形成する。典型的な記述子は音声の強さ若しくは大きさの指標、スペクトル中心のような周波数に関する指標、明るさ及びシャープさ、分類別の相似性指標（例えば、入力サウンドトラックの１つのセグメントが人間の声を含む可能性）、他の記述子の統計的特性と変化割合の指標、２若しくはそれ以上のデスクリプタを組み合わせることにより形成された記述子、等である。これら多くの記述子とテクニックは当業者によく知られており、新しいものは継続的に定義される。
【００２３】
イメージアナライザ［１１４］。これは入力イメージを分析して１又は２以上の記述子を含む１つのイメージ記述［１１５］を形成する。このイメージアナライザは信号分析テクニックや個別のイメージやイメージのグループに対する他の種類の処理を適用し、記述子を形成する。典型的な記述子は明るさ若しくはカラーヒストグラムのような色の指標、テクスチャーの指標、形状の指標、分類別の相似性指標（例えば、イメージが人の顔を含む可能性、それが自然情景である可能性等）、他の記述子の統計的特性と変化割合の指標、２若しくはそれ以上のデスクリプタを組み合わせることにより形成された記述子、等である。これら多くの記述子とテクニックは当業者によく知られており、新しいものは継続的に定義される。
【００２４】
音楽アナライザ［１１６］。これは入力音楽を分析して１又は２以上の記述子を含む１つの音楽記述［１１７］を形成する。この音楽アナライザは信号分析テクニックや音楽のセグメントに対する他の種類の処理を適用し、記述子を形成する。典型的な記述子は強さや大きさの指標、ビート強度、音楽的なリズム及びテンポの指標、スペクトル中心、明るさ及びシャープさのような周波数に関する指標、ルートノート音程、協和音、音楽的キーメンバー及び和音のような音楽的音程の指標、他の記述子の統計的特性と変化割合の指標、２若しくはそれ以上のデスクリプタを組み合わせることにより形成された記述子、等である。この音楽アナライザは種々のタイムスケールにおいて入力音楽の構成の表現を提供することができる。この時間軸は導入部分、節及びコーラスのような“マクロ”なタイムスケールから小節、ビート及びサブビートのような“ミクロ”なタイムスケールまである。音楽構成を表現する手段は音楽家、音楽理論化及び他の人には周知であり、このようなタ

イプの情報を信号分析により抽出するテクニックはコンピュータ音楽分析の分野では周知である。

【００２５】

この明細書において、前述の分析要素［１１０、１１２、１１４及び１１６］は“メディアアナライザ”として集合的に知られており、記述［１１１、１１３、１１５及び１１７］は“メディア記述”として知られている。

【００２６】

メディア記述はその後の使用のために、例えばこれをディスクや不揮発メモリにセーブすることにより、保存することができる（簡素化のため図１には記載されていない）。これはユーザが、再分析の必要なく、入力素材から異なる出力制作物を形成することを許容する。よって、多数の代替的な制作物を見るための処理時間が縮小される。

【００２７】

信号分析に対して追加的に又は代替的に、記述子を装置にインポートしメディア記述に保存することができる（簡素化のため図１には記載されていない）。かかる記述子は少し前の時間に形成されており、典型的には入力素材中に内蔵されまた何かの方法で入力素材にリンクされている。このような記述子はビデオ記述を含み、このビデオ記述は撮影時間、焦点距離、カメラに取付けられたＧＰＳにより形成される地理的位置などのカメラ用測定装置により形成される。これらはまた音楽記述を含むことができ、この音楽記述は音楽シーケンサやＭＩＤＩ（Ｍｕｓｉｃａｌ Ｉｎｓｔｒｕｍｅｎｔ Ｄｉｇｉｔａｌ Ｉｎｔｅｒｆａｃｅ）データから抽出若しくは提供されるエレメントのような音楽制作プロセスの間に形成される。音楽シーケンサやＭＩＤＩは音楽制作に広く用いられており記述的な情報を生成するのにも使うことができる。この記述的な情報はミックスダウンされた後の音楽音声信号から提供するものと異なる：例えば、音楽的な音程、測定、音楽の繰り返し構造等である。

【００２８】

インポートされた記述子はマニュアルや半自動化プロセスから発生させることができる。例えば、入力素材及びその記述子を装置内にインポートする前にユーザが入力音楽、ビデオ若しくはイメージに注釈する。かかる記述子は信号分析により生成された記述子に近い関係にある。例えば、記述子を生成してこれを装置を用いてマニュアル的に訂正若しくは修正することが望ましい。そしてその結果、装置において他のモジュールによる処理のための基礎として修正された記述子が用いられる。

【００２９】

インポートされた記述子はメディア記述に直接保存することができる。若しくはそれらにはインポートした後に更なる分析、変換及び解釈が要求されることがある。この機能はメディアアナライザにより提供される。

【００３０】

他のコンポーネント

図１に記載されているように装置は次のコンポーネントを備えている。

グラフィカルユーザインターフェース即ちＧＵＩ［１２０］。これはユーザと装置の仲介手段として機能し、装置の他の複数のモジュールと通信する。ユーザの関与として典型的には次の事項がある。

入力素材を含むファイルの選択及び出力制作物のためのあて先ファイルの選択のような完全な制御。制御は他の面では分析の開始やタスクの構築を含む。

スタイル情報を伴うユーザの関与－例えば、所定のスタイルの選択、若しくは新しいスタイルの生成、若しくは既存のスタイルの変性がある。

マニュアル的な関与はプリセレクションステージとタッチアップステージの双方において関係する。

【００３１】

ＧＵＩの特性と変形態様を更に以下に説明する。

コンストラクタ［１２１］。これは装置の主たるロジックの多くを含む。これは入力され

た１又は２以上のメデイア記述子を受取りそしてスタイル情報［１０５］を受け取る（若しくはそれ自身中に含む）。その主たる機能はこれらの入力を用いて全ての編集決定を作ることにあり、この編集決定は出力制作物［１０８］の様式を特定するため及び“メディアシーングラフ”即ちＭＳＧと呼ばれる構成内に当該出力制作物の様式の特定物を保存するために必要とされる。このＭＳＧは出力制作物の様式の完全な表現としてまた出力制作物を作るための指令の完全なセットとして考えられている。これは入力素材（ビデオ、音楽若しくはサウンドトラック、選択されたイメージ、又はイメージやビデオフレーム中の領域など）の全ての要素のソースやタイミングを含み、当該入力素材は出力制作物、変形のタイプやこれらの要素に適用される特殊効果、出力制作物中で用いられる変形効果のタイプ、出力制作物中で用いられるテキストやグラフィックのような全てのオーバーレイのソース及び表現、これら全ての要素のタイミング等に使用される。このＭＳＧはレンダラ（ｒｅｎｄｅｒｅｒ、下記参照）を制御しそしてマニュアルタッチアップの間に重要な役割を果たす。それは、一次的に基礎となるデータ構成であり、ユーザはこのデータ構成をこのステージで操作する。またこのデータ構成は全ての時間おける現在の制作物の全ての表現でありかつユーザにより作られた変化を反映するために更新される。

【００３２】

このＭＳＧは後の使用、最終制作物の進行性のタッチアップを許容するために任意的にセーブされまたリロードされる。そして、ＭＳＧの一部（例えば編集情報の一時的な領域や所定のタイプ）を“ロック”し、残りを“アンロック”することができる。これは進行的な修正により出力制作物が作られることを許容する。これにより、ユーザは装置にコンストラクタ（及びレンダラ）が作動するように指示し、結果の出力制作物を観察し、彼／彼女の好む特徴や領域をロックし、他の出力制作物を観察し、他の領域／特徴のセットをロックし、かかる作業を繰り返す。

【００３３】

コンストラクタのロジック及びＭＳＧの構成は以下に詳述する。

レンダラ［１２３］。これはＭＳＧ内の情報に基づいて出力制作物を形成する。換言すれば、これはＭＳＧのデータをインストラクションとして解釈する。そしてインストラクションにおいて、入力素材のエレメントの選択、選択されたもののシーケンシング、変形、結合及び集中のような応用プロセス、これらをファイルや音声画像モニタのような出力に移送若しくはコピーする。結果は出力制作物である。レンダラの動作の種類は周知でありこれ以上の説明は不要であり、多くの非線形ビデオエディタにおいて見られ、マイクロソフト社のＤｉｒｅｃｔＳｈｏｗ及びアップル社のＱｕｉｃｋＴｉｍｅのような汎用的なビデオアーキテクチャにより一般的にサポートされる。このレンダラは圧縮モジュールを有することができる。この圧縮モジュールはデジタルビデオ圧縮やデジタル音声圧縮のような技術を用いて出力制作物を圧縮する。これらの技術としてＭＰＥＧ（Ｍｏｔｉｏｎ Ｐｉｃｔｕｒｅ Ｅｘｐｅｒｔｓ Ｇｒｏｕｐ） スタンダード体として規定されるようによく知られている。

【００３４】

明細書において代表的に、発明はメディアアナライザ、コンストラクタ及びレンダラを含む単一の装置として記載されている。しかしながら、それは分配されたシステムでもよく、当該分配されたシステムにおいて、各モジュールは分離されたプログラム、異なる部分による異なるロケーションにおける異なる時間での可能な実行である。コンストラクタにとって必要なときはメディア記述が保存されそしてインポートされることは既述の通りである。いずれかの部分によるいずれかのロケーションにおけるいずれかの先の時間に形成されたメディア分析モジュールによってこれらのメディア記述は生成される。

【００３５】

同様に、ＭＳＧは出力制作物を作るための完全なインストラクションのセット若しくは出力制作物の様式の完全な表現である。そのため、レンダラはコンストラクタ若しくはアナライザから独立して動作することができる。出力制作物を観察しながらこれを実時間で動作させられる。換言すれば、出力制作物を急いで生成することにおいて、レンダラは優れ

たプレバックエンジンの効果を奏する。この可能性を作るために要求されるすべてはＭＳＧと入力素材がレンダラ動作時間において入手できることにある。

【００３６】

例えば、２つの集団が入力素材の共通部分に対するアクセスをシェアするとき、若しくは入力素材に２つの同一コピーがあるとき、一つの集団がアナライザとコンストラクタを動作させてＭＳＧを生成し、このＭＳＧを第２の集団に送信し、ここにおいて第２の集団はレンダラを実行させて出力制作品を早急に生成し、彼／彼女がそれを観察する。他の例では、人々のコミュニティが入力素材の共通部分と補助された既生成のメディア記載のコピーを最初に要求することができる。その結果、異なるＭＳＧの単なる移送によりそれらが相互にシェアする外部制作物が個別に形成される。この利点は次のとおりである。即ち各ＭＳＧは典型的なメディアデータに比較して小さな量であり、そしてそのため迅速かつ簡単に移送することができる。メディアの共通部分はＣＤ－ＲＯＭやＤＶＤ等の媒体によって分配されることに適している。当該ＣＤ－ＲＯＭ／ＤＶＤを所有する人々のコミュニティはその制作物を共有することができ、例えば、ＭＳＧをイーメールの添付物として他者に送ることができる。

【００３７】

自動的な生成のプロセスが図２から８を参照して説明される。

ビデオ編集例

図２は出力制作物が入力素材から生成される典型的な例を示し、ここにおいて前述の構成プロセスのアプリケーションが用いられる。この構成プロセスには、細分化、選択的導入、順序付け、変性及び結合である（この図は純粋に視覚的な例であり、音声は示さない）。従来的な線形及び非線形編集において、これらのプロセスは周知でありまたマニュアルでなされていた。この発明の主たる目的はこれらの全部又は一部を自動化することにある。この発明がいかにして当該自動化を達成するかを説明する前に、図２に記載の例の幾つかを検討することが有効である。

【００３８】

細分化

デジタルビデオファイルのような２つの入力ビデオ［２０１、２０２］が細分化されて５つの“ソース”セグメント、ｓＳｅｇ１　からｓＳｅｇ５［２１１、２１２、２１３、２１４、２１５］が形成される。その中の１つであるｓＳｅｇ５［２１５］は一つのフレームからなるセグメントである。

選択的導入

５つのソースセグメント［２１１－２１５］は出力ビデオ制作物に含まれ、他方入力ビデオの残りの素材は使用されない。一つのイメージとしてｓＩｍａｇｅ１［２１６］が含まれる。

順序付け

この例において、出力制作物を形成するセグメントの順序は入力素材のそれと同じではない。例えば、出力制作物において、入力ビデオＢ［２１１，２１４］からの最初の２つのセグメントは入力ビデオ［２１２、２１３］からの２つのセグメントにより離される。

変形

幾つかの変形の例が図２に示されている。セグメントｓＳｅｇ２は色彩の情報が削除されてもモノクロに変形され、その明るさ［２２０］だけが保存される。ｓＳｅｇ３はフラッシュ効果を付加することにより変形される。即ち、１又は２以上のフレーム内の領域の明るさが強調される。ｓＳｅｇ４は時間の変形に関し、例えばオリジナルのフレーム［２２２］の間に新しいフレームを生成することによりオリジナルのスピードの０．４倍に遅くする。ｓＳｅｇ５はさらに大きな時間の変形に関し、ここにおいて１つのフレームがフリーズ［２２３］を生成するために幾つかの進行性のフレームにコピーされる。ｓＩｍａｇｅ１もまた複数の進行性のフレームにコピーされ、その結果出力制作物［２２４］のセグメントを形成する。その他多くのビデオ変形が周知である。更には、重ねて使用されるテキストとグラフィック要素は種々の方法で変形することができる。例えば、アニメ化しそ

の結果それらは位置、大きさ、形、色その他を時間の経過とともに、また可能な場合以下に説明するように音楽のパラメータに対応して、変化させる（これらは図２において“ＡｎｉＴｅｘｔ”［２２５］及び“ＡｎｉＧｒａｐｈｉｃ”［２２６］として示されている）。テキストとグラフィック要素はフェードインし［２３５］またフェードアウトする［２３６］。

結合

図２はまた入力素材を結合する幾つかの方法を示している。変形されたセグメントｄＳｅｇ１及びｄＳｅｇ２は連結されて切断部を形成するか又は突き当て編集［２３０］を形成する。他のセグメントは部分的に重ねて結合され、ディソルブ［２３１］、ワイプ［２３４］及び他の周知の変形効果が許容される。テキスト及びグラフィック要素、静止［２２７］及びアニメ［２２５、２２６］はビデオ上で重ねられて最終制作物を形成する。

【００３９】

入力素材を用いずに装置により形成された素材の例が図２に示されている。それはブラックバックグランド［２２８］がテキスト［２２７］に重ねられることである。

【００４０】

上記の全ては出力制作物に関してタイミング参照をしている。それらは出力制作物のタイムライン［２４０］上に投影される縦方向の破線として示される。入力ビデオのセグメントはその入力ビデオソースファイルに関連する付加的なタイミング参照のセットを含む。例えば、ｓＳｅｇ４における開始時刻［２４１］及び終了時刻［２４２］である。

【００４１】

従来的なＮＬＥにおいては、これらのプロセスを何に対して適用するかまたどこで適用するかについてユーザが全て決定している。この発明では出力制作物を自動的に生成する。それは、それ自体の決定を形成しまた既述のプロセスをその通り動作させることにより行われる。コンストラクタ［１２１］は装置の心臓部であって、どのプロセスをどこに適用させるかを決定する。それに対し、レンダラ［１２３］は実際のプロセスを実行する。

【００４２】

構築プロセス

図３はこの発明の中心的な構築ロジックを示す。コンストラクタ内の構築ロジック［３０１］はスタイル情報［３０２］及び入力としてのメディア記述（ビデオ及び／又はイメージ［３０３］の記述及び任意の音楽記述［３０４］）を備え、これらの情報を用いて編集決定のセットを作る。ここにおいて、編集決定はＭＳＧ［３０５］に保存され出力制作物を特定する。スタイル情報は参照、提案若しくは構築ロジックの要求のセットとして考えることもできる。選択物に対する構築ロジックの動作はメディア記述内のデータの価値に依存する。その結果、特別な編集決定のセットはスタイル情報と入力素材の特性にともに依存することとなる。

【００４３】

このプロセスの幾つかの例を以下に詳細に説明する。最初はスタイルの特性から始める。

スタイル

スタイルはデータ、ロジック若しくはこれらの組合せより定義される。例えば、図３のスタイル情報［３０２］はマニュアル的に定義されたパラメータのセットとすることができる。ここにおいて、このパラメータは構築ロジックによりインポートされる。また、このパラメータはオブジェクト指向性プログラミングインプリメントに分類されるスタイルのようなプログラムされたスタイルのロジックにより形成されるパラメータのセットであることができる。この特質はこの説明若しくは後の説明において非常に重要というわけではない。なお、この説明及び後の説明は交換可能である。

【００４４】

スタイル情報は、例えば、パラメータの価値のセットをマニュアルで規定するプロセスによって、スタイルデザイナにより生成される。そしてスタイルデザイナの目的はスタイルを生成することである。ここにおいて、スタイルはシステムを生み出して高品質な出力制作物を形成する。この情報はスタイルを含み、このスタイルはそれが影響するプロセス構

築の部分のいずれかに対応して分類され、この分類は既述と同様の分類手法を用いる。例として、一実施例のスタイル情報は次の通りである。

【００４５】

“細分化パラメータ” これらの効果の幾つかは入力ビデオ若しくは入力サウンドトラックが細分化される方法に影響する。ビデオを細分化する多くのテクニックが知られており、それらは、カラーヒストグラムテクニックを用いるショットを形成する細分化、補助的なサウンドトラックの音響特性に基づく細分化等である。細分化は線形であり、入力素材のスタートからエンドまでのリストにおいて同じ重さのセグメントのセットの特定である。若しくはそれは階層的であってもよい、即ち、入力素材がセグメントに分割され、そのセグメントはセグメント持続の階層において他のセグメントを含んでいる。各スタイルは使うべきテクニックを特定し、またパラメータを特定する。このパラメータは閾値（ショット変化と考えられるカラーヒストグラムの変化の度合いのような）、セグメント長の最小及び最大値、特定すべきセグメント最小値等である。このように入力ビデオや入力サウンドトラックの細分化を制御するパラメータに加えて、望ましいセグメント持続を制御するパラメータがある。即ち、出力制作物を含むセグメントの望ましい持続時間である。これは“カッティングスピード”を制御し、これは出力制作物の重要な特徴である。

【００４６】

“選択的導入パラメータ” これらは入力素材の要素の選択を制御するパラメータのセットである。この入力素材（ビデオ、音楽若しくはサウンドトラック、選択されたイメージ、イメージやビデオフレーム中の領域のセグメントのような）は出力制作物において異なるポイントにおいて使用される。特に、この実施例において、それらは、明るさ（ビデオやイメージの平均照度）及び望ましい活動レベル（ビデオの平均的総モーション）を含むメディア記述の目的価値のセットである。他の実施例においては、既述（“構成要素の分析と説明”の欄における）の記述子のいずれの種類も使うことができる。

【００４７】

“順序付けルール” いずれのスタイルも順序付けの操作方法を特定している。例えば、パラメータは出力制作物を含む入力素材の要素を如何にして選択するかを特定する。選択の方法として、順序付けられて（入力素材に生じているのと同じ順序で）、順序付けられないで（入力素材の順序付けを考慮せずに）、若しくは部分的順序付けて（例えば、素材を順序付けて移動させる時間間隔のある長さ内をみれば、オリジナルの順序は巨視的には保存されるが微視的には順序付けられない選択が許容される）が挙げられる。

【００４８】

“変形パラメータ” これは各スタイルにおいて用いられる変形のセットを特定し、また出力制作物の異なるポイントにおいて適用されるべき変形の種類のためのルールを特定する。例えば、パラメータのセットがフラッシュ効果の特別なタイプを特定する。このフラッシュ効果は明るさ、範囲、持続時間等として表現される。そして“出力制作物の全ての第４のセグメント、しかし最後のフラッシュから１０秒が経過し現在のセグメント明るさが所定値以下”のようなフラッシュが適用されるときに、ルールのセットが特定されることがある。変形パラメータはまたある方法を特定する。この方法においてテキストとグラフィック要素が提供されてアニメ化される。ここにおいて位置、サイズ、形、色等の静的及び動的な価値が含まれる。

【００４９】

“結合パラメータ” これは入力素材（システムにより形成されたもの）のいずれの要素を結合すべきかを特定する。例えば、使用する変形（カット／ディソルブ／ワイプ）のタイプ、どの程度の頻度及びいかなる順序付けを各タイプに用いるか、変化の持続時間、ブラックバックグランドを形成する時期及び期間、テキスト／グラフィック要素の重ね合わせの時期といかなるタイプの素材がその上へ重ねられるか（例えば、ある値を超える明るさのビデオ素材若しくはホワイトテキストの重ね合せを禁止するために）等を挙げられる。

【００５０】

パラメータ及びその値の正確な選択は事実や部分的には主観に依存する。可能性の範囲は大きくそして入力素材の範囲やタイプのようなファクタに影響される。この入力素材はこの装置の人口統計的に好ましい対象ユーザにより順次処理されて、そして他のファクタに影響されることとなる。

【００５１】

編集決定における変化の形成

面白い制作物を生成するために、制作物のコースにおいて編集決定に関するいくつかの変化の導入が一般的に必要である。例えば、多くの場合、前述の好ましいセグメント持続時間を変えることが好ましい。同じ長さのセグメントを有する制作物のなかには速やかに単調になってしまうものがある。“編集リズム”の満足を生成するようにセグメントの持続時間は変化されなければならない。

【００５２】

一つの実施例において、これは変化の導入を必要とし、この変化は単一で若しくは複合した幾つかの方法により達成される。

“セクション”及び“サブスタイル”　出力制作物はセクションの一連として構築されている。各セクションには異なるサブスタイルが与えられている。これらのサブスタイルはある順序、即ちサブスタイル順序により用いられる。ここにおいて、サブスタイルは任意に繰り返される。各サブスタイルは幾つかの若しくは全てのスタイルパラメータ（及び／又はスタイル情報を形成するためのロジック）の値を含む。例えば、このスキームは出力制作物において３つのセクションを定義するスタイルを特定することを可能とする。ここにおいて、第１のセクションは低い明るさで、僅かな特殊効果を伴い、遅いディソルブ変化を伴う長いセグメントを含み、第２のセクションは高い明るさで、多くの特殊効果を伴い、シャープカットを伴う突合せ編集を含み。第３のセクションは第１のそれと同様の特徴を有する。

【００５３】

“徐々の変化”　スタイルパラメータのサブセットについて徐々の変化を特定することが可能である。例えば、以前の例の２つの対照的なセクションの代わりに、第１のサブスタイルの特徴から第２のサブスタイルのそれへの緩やかな変化が認められることがある。この例において、サブスタイル変化において突然に変化する２つの明確に特定されたセクションを持つことが可能である。また、出力制作物の過程において徐々に変化するパラメータの幾つかが許容される。

【００５４】

“確率的な形成”　制限偶然変化（Ｌｉｍｉｔｅｄ　ｒａｎｄｏｍ　ｖａｒｉａｔｉｏｎ）が出力ビデオの各セグメントの平均において導入されている。これは、各セグメントに対してパラメータ値の変化を伴うコンストラクタを提供する。例えば、あるサブスタイルは望ましいセグメント持続時間を特定する。このセグメント持続時間は０．２５Ｓの標準偏差を伴う正規分布を用いる１Ｓと２Ｓの間の偶然値とされる。この場合、各時間において、コンストラクタはサブスタイルからの値を要求し、与えられた値が異なることとなり、しかし、常に１Ｓと２Ｓの間に位置する。

【００５５】

“値サイクル”　これはまた出力ビデオのかくセグメントの平均において動作される。各パラメータは値の列となり、これらの値は繰返しの順序付けで使われる。例えば特殊なサブスタイルにおいて、望ましいセグメント持続時間は３つの値：４，２，２（秒）の順序付けを有してもよい。ここにおいて、このサブスタイルが用いられ、出力制作物におけるセグメントの持続時間はサイクル４，２，２，４，２，２，４，２，２，（続く）となる。異なるパラメータのためのサイクルの長さは同じであったり異なったりする。例えば、表１において、セグメント対象の明るさは明と暗を繰り返す（サイクル長さは２）。持続時間及び変化タイプのセグメントは３の長さのサイクルを有する。全ての第４のセグメントはモノクロに変化され、全ての第８のセグメントはフラッシュ効果を有する。すべてのパターンは全ての２４番目のセグメントのみにおいて複製される。これは変化を生成し、

また、出力制作物の編集リズム内にサイクルの質を導入する。多くの観察者はこれを明白に気づくことはない。これは意識下であり、しかし、確率論的な変化に対して異なる効果を生成し、そしてある場合に制作物の品質向上に気付くであろう。これは、以下に説明するように出力制作物が音楽による制作物であるときに特に真実である。

【００５６】

【表１】

	サイクル長さ	1	2	3	4	5	6	7	8	9	10	11	12
対象明るさ	2	暗	明	暗	明	暗	明	暗	明	暗	明	暗	明
持続時間(S)	3	4	2	2	4	2	2	4	2	2	4	2	2
移行型	3	カット	カット	ディソルブ	カット	カット	ディソルブ	カット	カット	ディソルブ	カット	カット	ディソルブ
色 or モノクロ	4	M	C	C	C	M	C	C	C	M	C	C	C
フラッシュ効果	8	No	No	No	No	No	No	No	Yes	No	No	No	No
時間		→		→		→		→		→			

Diss＝ディソルブ

【００５７】

入力素材の要素を選択して出力制作物を構築する

コンストラクタの中心的な機能として入力素材（ビデオ、音楽若しくはサウンドトラック、選択されたイメージ、若しくはイメージやビデオフレーム中の領域に関するセグメント）の要素を選択し順序付けることがある。この入力素材により出力制作物が形成されることとなる。以下、このことについて説明するが、そこにおいて入力素材はビデオであり、問題となる入力素材の要素はビデオセグメントである。イメージのセットのような他のメディアのプロセスにも関連しこれはより単純である。

【００５８】

既述のように（“順序付けルール”参照）、スタイルは出力制作物を含む入力素材の要素を如何にして選択するかを特定する。選択の方法として、入力素材から順序付けられて、順序付けられないで若しくは部分的に順序付けられて、が挙げられる。入力素材の要素の選択のプロセスは幾つかの複雑さを含む。この複雑さについて、順序付けられたケースと順序付けられていないケースとを参照して説明する。部分的に順序付けられるケースのような変形態様は、以下に説明するテクニックの併用により達成することができる。

【００５９】

セグメントの選択：順序付けられたケース

図４は一般的な順序付けられたケースを示す。ここにおいて、入力ビデオ［４０１］の一つの連続したものがある。この入力ビデオはD_iの持続時間を有し、この持続時間は出力制作物［４０２］の持続時間D_0より遥かに大きい。この入力及び出力持続時間の比は$R_{io}=D_i/D_o$である。入力ビデオはI_1-I_8と符号されたショットのようなセグメントに分割される。

【００６０】

コンストラクタはセグメントにより出力制作物セグメントを形成する。この例では、O_1-O_6のセグメントが既に形成されており、次のセグメントを形成しつつある。新しいセグメントを選択するために、コンストラクタは次に説明する図５のフローチャートに従って動作する。このプロセスを図４及び５を参照にして説明する。

【００６１】

コンストラクタは、図４に示されるように、t_0と符号された新しいセグメント［４０３］のための出力制作物におけるスタート時刻を第１にゲットする［５０１］。これは後に新しいセグメントのために要求されるパラメータのセットもゲットする［５０２］。例えば、要求されるセグメント持続時間d_0及び変化及び効果に関するデータを含むスタイル情報である。この持続時間は、入力ビデオから取得さるセグメントのための目的セグメント持続時間d_t［４０４］を形成するためにアジャストされ［５０３］、次の２つのこと

を許容する。
【００６２】
セグメント前及び／又は後のディソルブのようなオーバラップした変化があったならば、これらの持続時間は対象セグメント持続時間D_tに含められなくてはならない。
効果がスピード変化に適用されたならば、持続時間は計測されなければならない。例えば、出力セグメントが二倍の速度で演奏されるならば、対象セグメント持続時間d_Tは出力セグメントd_0の二倍の持続時間となる。
【００６３】
コンストラクタはその結果、入力ビデオの時間ｔを計算する［５０４］。ここにおいて、好ましいセグメントの発見が始まる。順序付けられたケースでは出力制作物は一般的に入力ビデオに対して殆ど線形なものとして表現される。そしてこれを達成するため、次のようにして計算されたタイムロケーションから入力ビデオセグメントが理想的には取得される。
【００６４】
$t_i = R_{i0} * t_0$
換言すれば、入力及び出力ビデオの相関位置は同じになる。
【００６５】
コンストラクタは少なくともd_Tの持続時間を有する新しいセグメントを形成するのに充分な長さのt_iセグメントのサブセグメントが存在するか否かをチェックする［５０５］。持続時間が＞＝d_Tであることに加えて、サブセグメントの選択は二つの制約が課せられる。
【００６６】
入力ビデオにおいてセグメント境界に近づかない。例えば、入力ビデオがショットに分割されるとセグメント境界に近づくことは好ましくない。なぜなら、そのようにすると出力作成物に意図しないカットを導入することになるからである。更には、生のビデオ素材においてショットの境界は一般的にきれいにカットされていない。例えば、再始動の後にビデオの再同期化として幾つかの悪いフレームが存在することがある。そしてそれはショット境界に近づく素材の使用を好ましくないものとする。図４を参照して、t_iと入力セグメントI_5［４０５］の間のビデオのサブセグメントが少なくともd_Tの持続時間であるか否かが問題となる。
これは狭義の順序付けられたケースであるので、出力素材は入力ビデオに現れる時間順序と同じ時間順序で提供され、それが繰り返されることはない。そのため、選択されたサブセグメントのために、以前に選択された素材よりも遅い入力ビデオのロケーションから始動されなければならない。検索ロジックはt_iから任意に後方へ検索することができ、しかし、以前に使われた素材のエンドのみへ出来る限り早く戻らなければならない（このことは図５には明確に記述されていない）。
【００６７】
もしこのようなものが入力セグメントt_iにおいて見つけることができないときは、コンストラクタは後のセグメントに対して前方へサーチを行い［５０６］、充分な長さ（持続時間＞＝d_T）のセグメントを探す。しかし前方へ遠過ぎて検索ポイントを得られない：入力ビデオにおいて現在のロケーションから遠く離れたセグメントを選択することは後のセグメントを順序付けることにならない。入力ビデオにおいて検索を中止する好適なロケーションは次の式で表現される。
$t_{i-stop} = R_{i0} * (t_0 + d_0)$
【００６８】
コンストラクタが上記からセグメント若しくはサブセグメントを見つけたとき、それはその中から持続時間d_Tのものを選んで［５０７］、出力セグメントとして使用する。この選択は単純であり、例えば、サブセグメントの最初の部分から選ぶことができる。若しくは、他の規則に適合する長さd_Tのものを見つけるように企てることがより好ましい。例えば、記述子対象値をマッチングさせることによる（以下の順序付けされていないケース

で説明されているものと同じ原理を用いる）、又はより面白いと評価されるもの若しくは周囲の素材に比べて質的に優れたもの（同様に以下参照）である。入力ビデオのセグメントが出力ビデオのセグメントよりも極めて長いときに、一般的な状況では最も有効である。

上記のアプローチを行ってもコンストラクタが好ましいセグメントを見つけられないときは、出力セグメントが入力ビデオのセグメントの周囲を含んではならないとする制限を緩和し、d_Tの持続時間の出力セグメントを入力ビデオにおいて２以上のセグメント／サブセグメントから形成する［５０８］。

【００６９】

セグメントの選択：順序付けられないケース

この順序付けされないケース（図６）において、幾つかのステップは既述の順序付けされるケースと同様である。

上記と同様に、コンストラクタは新しいセグメントのための出力制作物におけるスタート時刻をゲットする［６０１］。そして新しいセグメントに要求されるスタイル情報のようなパラメータのセットもゲットする。これには、セグメント持続時間d_0及び効果と変化に関するデータが含まれる［６０２］。この順序付けされないケースにおいて、スタイル情報から対象記述子値のセットをゲットしている［６０３］。これはこの値のセットに適合するセグメントが選択される。

【００７０】

持続時間はその結果、入力ビデオから取り出されるセグメントのために対象セグメント持続時間d_Tを形成するように調節されなければならない［６０４］。ここにおいて、既述の順序付けられたケースにおいて説明されていた方法と同様の方法により変化とスピード変化が許容される。

【００７１】

次のステップ［６０５］は入力ビデオにおいて候補となるセグメントとサブセグメントを見つける。これらは持続時間に少なくともd_Tの持続時間のセグメントである。これらはまた他の要件を満足する必要がある。例えば、ある順序付けられないケースにおいて再利用の素材が許されるかもしれないが（順序付けられるケースと全く異なって）、出力制作物において同一の素材の現れる回数を制限することが好ましい。これは、入力ビデオの各部分がどの程度用いられたかのカウントを継続することにより達成される。ここにおいて、候補となる（サブ）セグメントは入力素材の一つのセグメントの連続部分であり、これは最高許容回数よりその使用回数が少なくかつ少なくともd_Tの持続時間を有する。

【００７２】

そのような（サブ）セグメントが見つからないときは、コンストラクタは制限を緩和し、例えば既述の順序付けをするケースと同様に、入力ビデオの２以上のセグメント／サブセグメントより持続時間d_Tの出力セグメントを形成する（図示されていない）。

【００７３】

コンストラクタはメデイア既述からこれら“候補”（サブ）セグメントのための記述子の値をゲットする［６０６］。そして、候補ポイントと対象記述子値の記述子スペースの距離を評価する［６０７］（このプロセスは以下に更に説明し、図７以降に示されている）。最終的に、対象ポイントから最も小さい距離の候補ポイントに基づいてコンストラクタは候補のセグメントを選択する［６０８］。そして、出力制作物に使用される［６０９］。

【００７４】

記述子空間の近似性によるセグメントの選択

既述のように、候補（サブ）セグメントのセットからベストマッチの（サブ）セグメントを選択することが必要である。このベストマッチの（サブ）セグメントは次のとおりであり、即ち“記述子空間”（ｎ次空間であって、各ｎ記述子が表されている）の対象値のセットに最も近いものである。即ち、与えられたポイント（スタイル情報からの対象値により規定される座標）と候補のポイント（メディア記述の値のセットにより規定される座標

）の間の距離が最も小さい。
【００７５】
単純化の原則に拘わらず、このマッチングプロセスには考慮すべき幾つかの問題がある。これらを図７を参照して説明する。この説明は入力素材がビデオである場合に関係するが、この原理は他のメディアにも適用される。
【００７６】
１．距離の計算を確実にすることは人の予測によく適合した結果をもたらす。全ての記述子は知覚のスケールを用いることが大切である［７０１］。これはスケールであり、当該スケールにおいて記述子の値の与えられた相違は、全ての記述子の範囲内の位置に関係なく、所定の値の相違としてユーザによく知られた藻のである。多くのケースでは、これをある物理量のログ値で近似することができる。
２．一般的にいって、記述子はたくさんの異なるレンジを伴う異なるユニットであることができる。例えば、セグメント持続時間を０から３０秒とする一方他の記述子では１から１０００のスケールを用いる。距離の計算の影響をさけるために、ユニットを０から１のような共通のスケールで正規化する。かかる“ユニットの正規化”［７０２］は下記の１次線形方程式を用いることができる。
ｘ′＝　（ｘ－ｘｍｉｎ）／（ｘｍａｘ－ｘｍｉｎ）
ここにおいて
－　ｘはもとの（ユニットの正規化されていない）ユニットの値である。
－　Ｘｍｉｎはもとの値の最小値である。
－　Ｘｍａｘはもとの値の最大値である。
－　ｘ’はユニットの正規化された値である。
【００７７】
装置はユーザが提供する全ての種類の入力ビデオ素材に対して、何ら制限することなく、満足な出力制作物を提供することが望ましい。従って、装置はビデオ記述内において記述子値の広がりについて何ら制御することがない。例えば、分析により抽出された記述子のセットが入力ビデオ素材のセグメント一つでことを除いた全てに対して同様な値を有すケースを想定されたい。換言すれば、セグメントを表すポイントをのぞく全てが記述子空間のある小さな部分とクラスター化し、他のポイントは遠く離れている。このケースにおいて、一つの独立したポイントがスタイル情報により提供された対象値の全てのポイントに最も近いことが可能である。単純な距離のメモリが使用されたならば、それは毎回このセグメントの選択を導くこととなる。これは潜在的に非常に多くの繰返しの同一セグメントからなる出力制作物を生み出すおそれがあり、これは許容できない。
【００７８】
この問題を解決する一つのアプローチとして抽出された記述子のバラツキ、たとえそのバラツキが小さくても、を利用して出力ビデオにおける多様性を生成することがある。これは“分布の正規化”［７０３］により達成される。即ち、各ポイントにおいて記述子の値を線形に計測及びシフトさせ、もって記述のクラスタリングを除去若しくは少なくすることができる。分布を正規化するため、各記述子に対して順次次の式を適用する。
Ｘ’＝（（ｘ－ｍ）＊ｓ’／ｓ）＋ｍ’
ここに
ｘは以前に分布正規化された値
ｍは入力値の平均
ｓは入力値の偏差値＊
ｍ’は望ましい（出力）分布の平均
ｓ’は望ましい（出力）分布の偏差値＊
ｘ’は分布正規化されたユニットの値
＊例えば、これは標準偏差や平均偏差（一般的な統計学の定義において）とすることができる。標準偏差は多くの場合により正確であると一般的に考えられている。他方、平均偏差はより素早く計算できる。

【００７９】
分布の正規化は２つの方法で適用することができる。
ａ）ビデオ記述からの記述子値のセットとスタイル情報からの対象値のセットの両者を正規化し、もってそれらを一般的な標準分布に適合させる。即ち、ｍ’及びｓ’に値を固定する。（これを行う他の方法として、最終的な結果を同一にするには、値のセットを調節し、もって他と同じ分布を有するものとする）。
ｂ）　値のセットただ一つを正規化する。例えば一般的な標準偏差であるビデオ記述からの値のみであり、この場合、各セットの分布は必ずしも同じでなくてもよい。
【００８０】
これらの２つのアプローチは異なるケースにおいて用いられる。それぞれ利点と欠点を有し、異なるスタイルによってサポートすることができる。ａ）の利点はユニークな組合せの傾向にあることであり、これは分布が“相互のトップの上にある”からである。その欠点はスタイルにおける値の平均の意図的な全体の偏りを排除してしまうことにある。実際、スタイルの平均を極端な方向へ偏らせることは不可能になる。例えば、全ての明るさの対象値がハイであるスタイルがあるとき、ａ）では明るさセグメントのための優先順位が放棄され、値のセットを偏らせない例として同じ明るさ／暗さのミックスが与えられる。他方、ｂ）の利点はこのような偏りを維持できることにある。そしてその欠点はユニークな組合せを有効に提案できないことにある。それは２つの分布が“相互のトップの上にある”ことができないからである。（他方、スタイル情報からの値の分布はシステムデザイナがコントロール可能であるため、マニュアル的に同様なものを作ることができる。しかしこれは全てのケースにおいて容易というわけではない。）
【００８１】
４．　分布の正規化を適用した後、データ中の異常値は与えられた範囲の外にはずさる。距離の計算を容易にするには、かかる異常値を除去するか若しくは所定の制限値内に戻す必要がある［７０４］。
【００８２】
５．　知覚された相同性を規定することにおいて幾つかの記述子は他のものに比べてより重要となる。記述子を次のように重み付けすることが好ましい。
ある記述子を全てではないが幾つかの場合に無視できるようにすることが望ましい。例えば、特定のサブスタイルは明るさとセグメント持続時間にその対象を特定しているが、活性化レベルのような他の記述子は無視する。このサブスタイルは他のサブスタイルと同時に使用される必要があり、ここにおいて他のサブスタイルは活性化レベルを特定する。そして、各ケースにおいて形成される値の距離は相互に比較されなければならない。これは、“重要”な記述子、即ち無視されないもの、のみのために距離を加えることによる距離の計算において達成される。これは次のように言うことに等しい。即ち、重要でない記述子にとってはいかなる値であってもその対象値に完全にマッチする。
【００８３】
重み付けと記述子の無視を許容することを考慮した距離の計算は次のようになる。
$D = SQRT(SUM_D((|v_{gd} - v_{cd}|\ \hat{}\ 2) * W_d))$
ここにおいて
Ｄは一対のポイント（一方は与えられ、他方は候補）の距離
ＳＱＲＴはスクウェアルート計算
SUM_Dは重要な記述子（無視されたものを除いた）のセットの総合計
V_{gd}は与えられたポイントにおいてｄ番目の記述子
V_{cd}は候補ポイントにおけるｄ番目の記述子
^２は二乗計算
W_dは記述子ｄの重さ
【００８４】
候補のセグメント若しくは対象ポイントの最も近くにあるそれらからなるサブセットは対象ポイントに近い順にリストされる［７０６］。図６を参照する既述の例において、一つ

の最も近い組合せのみが必要であることを説明した。しかしながら、この明細書のいずれかで記載したようにタッチアップマニュアルをサポートするために、近い順に並んだ他の候補セグメントのリストを有することが好ましい。

【００８５】

上記は記述子の直接マッチングを説明している。ここにおいて、スタイル情報の記述子のタイプはメディア記述の記述子のタイプと同一である。例えば、スタイル情報の明るさ記述子はメディア記述の明るさ記述子とマッチしている。これはまた非直接マッチングの使用を可能としている。ここにおいてスタイル情報で使用されている記述子のセットは、数学的若しくはアルゴリズムの関係を通して、メディア記述における記述子の異なるセットに対応付けられる。例えば、スタイル情報は次のように定義される“ＰｅｏｐｌｅＥｎｅｒｇｙ”を有することがある。

ＰｅｏｐｌｅＥｎｅｒｇｙ＝０．６＊Ｌｏｇ（Ａｃｔｉｖｉｔｙ）＋０．４＊ＰｅｒｓｏｎＰｒｏｂａｂｉｌｉｔｙ

ここにおいて“Ａｃｔｉｖｉｔｙ”はビデオセグメントにおける総平均行動の指標であり、“ＰｅｒｓｏｎＰｒｏｂａｂｉｌｉｔｙ”はそのセグメントが少なくとも一人のイメージを含む可能性（例えば周知の肌色検出アルゴリズムを用いて行う）の指標である。かかる記述子は、他の１又はそれ以上の記述子へ適用される数学的若しくはアルゴリズム的手法により規定され、“生成された記述子（ｄｅｒｉｖｅｄ　ｄｅｓｃｒｉｐｔｏｒｓ）”と呼ばれる。

【００８６】

かかる生成された記述子の観点からスタイル情報における対象値を定めることがときには有効になる。なぜなら、これは“ハイレベル”記述子の使用を許可し、この記述子はヒトにとってわかりやすり意味合いのある記述子の類型に近いからである。上記の例において、スタイル情報はＰｅｏｐｌｅＥｎｅｒｇｙの対象値を含み、他方、“Ａｃｔｉｖｉｔｙ”や“ＰｅｒｓｏｎＰｒｏｂａｂｉｌｉｔｙ”は入力ビデオの信号分析により抽出することができる。

【００８７】

生成された記述子を使用するとき、数学的若しくはアルゴリズム的な手法をメディア記述の低いレベルの記述子に適用することにより、コンストラクタロジックは生成された記述子の値を計算することができる。そして、記述子を生成された記述子の値にマッチさせる。

【００８８】

前述のように、入力素材の合計が出力制作物の長さよりも大きいことがしばしば生じる。従って、入力素材からより面白く若しくは品質に優れた部分を選択することが好ましい場合がある。これは前述のセグメント選択に関連し、同様なテクニックの幾つかを使用することができる。しかしながらその目的は相違している。即ち、前述のセグメント選択は基本的に出力制作物において入力素材が置かれるべき位置に関し、他方、素材選択の最適化は基本的に出力制作物において使われるべき入力素材の部分に関する。

【００８９】

本件のテクノロジは、単一の分析手法を用いている全種類の素材の全てについて確実に、ビデオやイメージの意味あるコンテントを決定するための手法を提供するものではない。従って、自動装置にとってヒトのビデオ編集者がしているように正確に素材を選択することは不可能である。更にこれは主観的なことであり、例えば、異なる編集者は異なる素材を選択する。そうであっても、素材の選択方法を偏らせ、入力素材の平均に比べてより面白く若しくは高品質であると多くのユーザが考えるような方向へ向かわせることができる。換言すれば、目的は、少なくともある種の素材において入力素材の中から取り出された偏っていないサンプルのセットよりも一般的にいって好ましい素材を自動的に選択することにある。

【００９０】

ここに、どのようにしてこれらの問題を解決できるかを示す２つの例がある。

１　多くの種類の内容を通じて、人々のイメージを含む素材が一般的に人々を含まない素材に比べてより面白いと考えられる。人間の存在を検出するための画像処理方法はよく知られている。例えば、肌の色、顔形、体形を用いる方法がある。これらの手法を用いることにより、一人又はそれ以上の人間を含むビデオのセグメント若しくはイメージの確率を示す記述子を計算することが可能になる。そのため、この記述子の高い値を有する素材はその低い値を有する素材よりも優先的に選択される。
２　“手持ち”の場面（即ち、三脚のような固定具ではなくカメラを手で支えて録画されたビデオ）においては、特に素人のユーザにとって、ファインダ内において何か特別に面白いものが見つかるまでカメラを動き回す傾向がある。換言すれば、手持ちで撮られて素材について、低いカメラ動作を伴ったセグメントは高いカメラ動作を伴ったセグメントに比べて面白い傾向にある。カメラの動作を予想する手法はよく知られており、例えば、動作ベクトルの抽出に基づく手法がある。したがって、入力ビデオ素材が手持ちで撮られたものであるかを特定することが可能である（これは、時間当たりの動作パターンを分析することにより決定することができる。また、プロンプトに対応してユーザが入力する情報によることもできる）。そして、それが手持ちで撮られたものであるとき、カメラ動作の低い部分が選択される。

【００９１】

これらの手法は、装置のユーザが望むときに使用できるオプションとして提供される。また、ユーザがあるスタイルを選択するときには使用でき、他のスタイルを選択したときには使用できないものとすることもできる。即ち、例えば、すぐ前で説明した手持ちで撮られたものについて低い動作を選択する手法は温和やくつろいだ雰囲気の出力制作物を形成する方向のスタイルに好適である。しかしながら、エネルギッシュで速いペースの制作物を形成する方向のスタイルには不適である。

【００９２】

メディアシーングラフ（Ｔｈｅ　Ｍｅｄｉａ　Ｓｃｅｎｅ　Ｇｒａｐｈ　（ＭＳＧ））
前述のように、ＭＳＧは出力制作物の様式を完全に表すコンストラクタによって形成されたデータ構造である。その意味では、これは周知の編集決定リスト（ｅｄｉｔ　ｄｅｃｉｓｉｏｎ　ｌｉｓｔ　（ＥＤＬ））に関係する。しかしながらＭＳＧはまた基本的には潜在的なデータ構造であり、ユーザはタッチアップの間にこれに接触する。他方、典型的なＥＤＬは線形な構造であり、この種の操作に適合していない。この種の操作により適合したＭＳＧ構造を図８を参照して以下に説明する。

【００９３】

構造は基本的にツリーであり、その中で出力制作物は根を構成する［８０１］。
ツリーにおいて幾つかの枝は定義に関係する。即ち、それらは出力制作物において使用されるある存在の特性を特定する。それらは出力制作物において使用される全種類の変形［８０２］のための定義を含む（例えば、あるフラッシュ効果がある明るさ、色、範囲、存続時間等を有することを特定する）。それらはまたディソルブのような移行［８０３］の定義、アニメ化されたテキスト含むテキスト［８０４］の定義、アニメ化されたグラフィックスを含むグラフィック要素［８０５］の定義、ブランクバックグラウンド［８０６］の定義等を含む。

【００９４】

ＭＳＧはまたタイムラインに関する１又は２以上の枝を有する。図８は、図２の２つのタイムラインと同様の目的のため、一つの主たるタイムライン［８０７］と一つの重複するタイムライン［８０８］を示している。主たるタイムラインは出力制作物を形成する各セグメント入力を含む。ここにおいて、出力制作物は入力素材［８０１］の要素と装置により構成されたブランク［８１１］からのセグメントを含む。これらのセグメントの変形［８１２］およびそれらの間の移行［８１３］もまた特定される。即ち、前述の変形と移行の定義を参照する様式が存在する。主たるタイムラインはセクション［８１４］の様式において追加的な構造のレベルを支持することができる。それらはそれぞれ単一のサブスタイルの使用に対応する（前述の“編集パラメータの変化の形成”を参照されたい）。これ

はユーザが選択するタッチアップ操作を促進させる。この操作は出力制作物の全てのセクションに適用することができる。最後に、重なったタイムライン［８０８］は、テキストの［８１５］及びグラフィックの［８１６］定義を参照することにより重なったものの順序を特定する。これらにはアニメ情報が含まれてもよい。

【００９５】

タイムラインを構成するセグメント、重なりその他の全ては時間データを有し、この時間データは出力制作物とある場合の入力素材の両者に関係する。ビデオセグメントの例によれば、ロケーション（スタートタイムのような）や出力制作物の持続時間の特定が必要である。即ち、スタートタイムや入力素材の持続時間のようなソースの特定が必要である。

【００９６】

グラフィカル・ユーザ・インターフェース（ＧＵＩ）

制作プロセスの高度な自動化により、本発明は、ある場合には、人間が介入しないでも、許容できる品質の出力制作物を生産することができる。それ故、本発明のある実施形態の場合には、ＧＵＩを非常に簡単なものにすることができるし使用しなくてもすむ。

【００９７】

図９は、非常に簡単であるが、実行可能なＧＵＩの一例を示す。このＧＵＩは、下記の機能を実行するために、ボタンのような３つの主要なユーザ制御装置を含む。

【００９８】

１．ユーザが入力素材を選択することができるようにする制御装置［９０１］。例えば、この制御装置は、ユーザに対して、入力素材を含む１つまたはそれ以上のビデオまたはイメージ・ファイルの名前を入力するようにプロンプトすることができる。

２．ユーザがスタイルを選択することができるようにする制御装置［９０２］。例えば、この制御装置を呼び出した場合、ユーザに対して、使用できるスタイルのリストを表示することができ、また、ユーザに対して、その中の１つを選択するようにプロンプトが行われる。

３．出力制作物を生成させる制御装置［９０３］。例えば、この制御装置は、ユーザに対して、出力製品を記憶するファイルの名前を入力するようにプロンプトすることができる。ユーザがこのファイル名を入力すると、出力制作物を生成するために、システムの主処理モジュール、すなわち、メディア・アナライザ、コンストラクタおよびレンダラが呼び出される。

【００９９】

プログラムを終了するための標準制御装置も設置されている［９０４］。

【０１００】

図１０はこのＧＵＩの修正例である。このＧＵＩは、下記の機能を実行するために、ボタンのような５つの主要なユーザ制御装置を含む。

【０１０１】

１．ユーザが入力視覚素材を選択することができるようにする制御装置［１００１］。例えば、この制御装置は、ユーザに対して、入力素材を含む１つまたはそれ以上のビデオまたはイメージファイルの名前を入力するようにプロンプトすることができる。

２．ユーザが入力音楽を選択することができるようにする制御装置［１００２］。例えば、この制御装置は、ユーザに対して、録音済み音楽を含む１つまたはそれ以上の音響ファイルの名前を入力するようにプロンプトすることができる。

３．ユーザがテキストを追加することができるようにする制御装置［１００３］。例えば、この制御装置は、ユーザに対して、テキスト情報をある形式で入力するようにプロンプトすることができる。テキストは出力制作物の上に重畳される（ｏｖｅｒｌａｉｄ）。オーバレイ・テキストの使用方法としては、タイトル、（制作に関連する人および組織に対する）クレジット、サブタイトル、説明のためのまたは宣伝のためのメッセージのようなメッセージ等がある。

４．ユーザがスタイルを選択または定義することができるようにする制御装置［１００４］。スタイルを選択するように、ユーザに対して使用できるスタイルのリストを表示する

ことができ、また、上記例のところで説明したように、ユーザに対して、その中の１つを選択するようにプロンプトすることができる。スタイルを定義するために、ユーザに対して、例えば、スタイル情報のすべてのパラメータの値を含む書式を表示することもできるし、ユーザに対して、値を入力するようにまたは変更するようにプロンプトすることもできる。

５．出力制作物を生成させる制御装置［１００５］。この制御装置は、上記例のところで説明したように、ユーザに対して、出力制作物を記憶するファイルの名前を入力するようにプロンプトすることができる。ユーザがこのファイル名を入力すると、システムの主処理モジュール、すなわち、メディア・アナライザ、コンストラクタおよびレンダラが呼び出される。この例の場合には、以下に説明するように、音楽をベースとする制作物を生成するために、視覚素材が音楽に編集され、音楽が入力サウンドトラックにより置き換えられるかまたはミックスされる。次に、出力制作物を制作するために、テキスト要素および図形要素のオーバレイが行われる。テキストおよび図形は、以下に説明するように、音楽に合わせてアニメ化することができる。

【０１０２】

プログラムを終了するための標準制御装置も設置されている［１００６］。

【０１０３】

上記例のいずれの場合も、出力制作物をメディア・プレーヤのような外部プログラムから見ることができる。別の方法としては、ユーザがシステム内から出力制作物を見ることができるように、上記ＧＵＩ素子をチェック・ウィンドウおよび当業者にとって周知の「移動制御装置」と一緒に供給することもできる。

【０１０４】

他の実施形態の場合には、ＧＵＩはマニュアル的な相互作用用の追加機能を含むことができる。そうする理由は、本発明の第１の目的は編集プロセスを自動化することであるが、何時でもどの場合でも自動化を完全に行うことはできないからである。入力素材の性質および問題の用途によっては、完全に自動的に生成した出力制作物があらゆる細かい点でユーザの好みと一致しない場合もでてくる。それ故、下記のようなマニュアル的な相互作用のための機能をサポートするのが望ましい場合もある。

【０１０５】

・コンテンツの予備選択。コンテンツを予め選択すれば、自動構成を行う前に（ビデオ、音楽またはサウンドトラック、選択したイメージ、またはイメージまたはビデオ・フレーム内の領域のセグメントのような）入力素材の要素を自由に選択したり選択から外したりすることができる。ユーザは、入力素材の要素を識別し、それらの要素を制作プロセス中に使用するか否か何処で使用するかまたはどんな順序で使用するかを指定する。例えば、ユーザは、特定のセグメントＡを、出力制作物中に挿入しなければならないと指定することもできるし、最後の場面に使用しなければならないと指定することもできるし、制作中に発生する他の要因によりある挿入確率で他のセグメントＢを挿入することができると指定することもできるし、セグメントＢが含まれている場合に限って出力制作物中のセグメントＢより後に第３のセグメントＣを挿入しなければならないと指定することもできるし、第４のセグメントＤを挿入してはならないと指定することもできる。予備選択のこのプロセスをメディア記述により容易に行うことができるようにすることもできる。例えば、ビデオ記述内の細分化情報を、入力ビデオを一連のショットとしてユーザに表示するために使用することができる。これは、ユーザにとって、通常、ビデオの隣接部分よりも扱い易い。メディア記述からの情報は、ユーザの助けになる種々の方法で入力素材を分類したりまたはまとめたりするのにも使用することができる。例えば、入力イメージのセットまたは入力ビデオ・セグメントのセットを、それぞれがある点では類似しているイメージのセットを含む一組の「ビン」内でユーザに表示することができる。そうしたい場合には、ユーザは、ビンに項目を追加したりまたはビンから項目を除去したりして、マニュアル的にこの分類を洗練されたものにすることができる。次に、ユーザは上記の（「挿入」、「挿入禁止」等）のような命令をイメージの全ビンに適用する。・処理の予備選択。処理の

予備選択により、ユーザは、自動構成の前に入力素材の要素に適用される処理の種々の態様を自由に選択したり指定したりすることができる。例えば、ユーザは、入力音楽のあるセクション内で起こる出力制作物のすべての遷移効果はディゾルブのようなあるタイプにしなければならないと指定することができる。または、ユーザは、入力イメージのサブセットをマニュアル的に選択し、出力制作物においてはこれらのイメージをモノクロにするよう指定することができる。この場合も、ユーザを助けるために、メディア記述からの情報に基づく細分化および集合のような自動プロセスを使用することができる。例えば、システムは、輝度により入力ビデオ入力のセグメントを分類し、ある輝度しきい値以下のセグメントのセットと一緒にユーザに表示し、ユーザがこのセットにセグメントを追加したりセットからセグメントを除去したりできるようにし、その視覚的品質を改善するためにユーザにこれらのセグメントの輝度をある割合で明るくさせることができる。

・出力制作物の仕上げ。出力制作物の仕上げにより、ユーザは、例えば、セグメントに適用された持続時間および効果を維持しながら、出力制作物のビデオ・セグメントを入力素材からの別のセグメントで置換することにより、または、遷移効果のあるものを変更することにより、特殊効果を追加したり除去したりすることにより、また追加のテキストまたはグラフィックスを重畳することにより、自動制作の後で出力制作物を編集することができる。この場合もまた、メディア記述からの情報をこれらの作業でユーザを助けるために使用することができる。例えば、ユーザが出力制作物のビデオのセグメントを置き換えたい場合には、システムは、ユーザに、そこから選択が行われる別のセグメントのセットを表示することができる。これらのセグメントは、ビデオ記述からの類似性の基準により元のセグメントとの類似性の順序でリストの形にすることができる。この例を修正したものの場合には、ユーザに対して、「類似のセグメントによる置換」／「対照的なセグメントによる置換」のような２つのオプションを表示することができる。ユーザがこれらのオプションの１つを選択すると、システムは適当な別のセグメントを適用する。

【０１０６】

メディア記述内の情報のマニュアル的な仕上げプロセスを容易にするために使用する方法の全く別の例は、出力制作物が音楽をベースとする制作物である場合に関連する。経験を積んだビデオ編集者によりビデオを「音楽に合わせて編集する」場合、通常行われる方法は、ある視覚的要素をビートのような音楽のあるタイミング特性にマッチさせるという方法である。この例の場合、音楽記述からのタイミング情報は、カットおよびフラッシュのような時間が重要な役割を持つ視覚的イベントがビート、サブビートおよび音楽内の他の有意な時間と自動的に整合するように、出力制作物の視覚素材上で、ユーザがマニュアル的に行う仕上げ作業を修正するために使用することができる。例えば、ドラッグのような標準的ＧＵＩ操作により出力制作物の２つのセグメント間のカット点を変更した場合、音楽記述からの情報を、音楽信号の振幅が大きい、または強いビートがあるという他の表示がある音楽内の時点間でカット点をジャンプさせるために使用することができる。関連オプションは、イベントの境界が、それ自身音楽のビートに整合しているタイミング・グリッドに整合している音楽シーケンサの分野において周知の技術である量子化を使用する方法である。

【０１０７】

これらのマニュアル的な作業をサポートするためのＧＵＩは、リスト、（ファイル・マネージャで使用されるもののような）階層的表示、視覚的サムネイル、オーディオ波形ディスプレイ、タイムライン、移動制御装置を含むクリップ・ウィンドウ等を含む標準素子により組み立てることができる。これらの素子は、当業者にとって周知のもので、非線形ビデオ・エディタ（ＮＬＥ）、イメージ・エディタ、オーディオ・エディタ、および他のメディア処理ソフトウェアのようなツールで通常使用されている。

【０１０８】

本発明は、また、単に出力制作物を表示するためだけのものであって、通常の使用のためのＧＵＩ素子を含んでいない非会話型システムでも実施することができる（しかし、このようなシステムは、それを構成し管理するためにＧＵＩを必要とする）。図１１はこのよ

うな実施形態の一例のためのロジックを示す。このロジックは、例えば、「ウェブ・カム」（インターネットに接続しているカメラ）からのビデオまたはイメージのような連続して到着する入力素材から出力制作物を生成するのに適している。ある量または持続時間が集まるまで、素材はカメラから捕捉される［１１０１］。この時点で、スタイル、およびそうしたい場合には入力音楽の一部が自動的に選択される［１１０２，１１０３］。これらのものは、単に、多数のオプションからのランダムな選択であってもよいし、または、本明細書の他のところで記述するように、記述子マッチングのプロセスによりスタイルおよび音楽をビデオ記述／イメージ記述の特性とマッチさせることができる。現在、システムは、出力制作物を作成するのに必要な情報を持っていて、出力制作物を作成する［１１０４］。最後に、システムは、出力制作物をマルチメディア・コンピュータまたはテレビジョン・セットのようなオーディオ・ビジュアル・ディスプレイ・デバイスに送る［１１０５］。この出力制作物の生成および供給の間、このシステムは、他の制作のためにすぐ使用することができる素材を引き続き捕捉することができる。本発明のこの実施形態の１つの使用方法は、入力素材をライブ・カメラから捕捉している公共の場所にいる人々に、一定の間隔で自動的に制作したオーディオ・ビジュアル制作物を供給するために使用するという方法である。

【０１０９】

音楽をベースとする制作物

この実施形態は、視覚的要素の処理およびタイミングが基礎となっている、音楽トラックの特性およびタイミングにより支配される出力制作物を生成するのに特に適している。これは「音楽に合わせてのカッティング」と呼ばれることもあり、音楽ビデオ、アニメ制作物、販売促進およびマーケティング・ビデオ、テレビ・コマーシャルおよび多くの他の形でよく使用される。本明細書においては、このような制作物を「音楽をベースとする制作物」と呼ぶ。

【０１１０】

音楽をベースとする制作物の一般的な原理は、音楽が時間の基準として動作することである。視覚的要素は音楽と合うように操作されるが、音楽自身は変わらない。これが適用される視覚的要素はモーション・ビデオ、イメージ、アニメ、グラフィックスおよびテキストを含む。さらに、音声および音響効果のようなある種の非音楽的オーディオ要素は、音楽により影響を受ける種々の方法で、時間に従って操作したり位置させたりすることができる。一般的な言い方をすれば、音楽は「主人」であり、他の要素は音楽に「奉仕するもの」である。

【０１１１】

音楽をベースとする制作物は多数の技術により制作される。現在プロの編集者の技術により達成されるこれらの技術としては下記のもの等がある。

【０１１２】

－　視覚素材の編集「ペース」は、通常、テンポ（すなわち、ビートの速度）、音の大きさ、音楽および知覚した「エネルギー」レベルのような音楽のいくつかの一般的な特性により支配されるか影響を受ける。例えば、音楽がもっと速くその音がもっと大きい場合には、出力制作物はもっと短い平均持続時間のショットからなり、もっと急速なカットおよびもっと少ないゆっくりとしたディゾルブにより、ショット間の遷移はもっと速くなる。これを制御する音楽的特性は、音楽のある部分から他の部分の間で変化するばかりでなく、音楽の１つの部分中のセクション毎に変化する。例えば、多くのポップス内の「エネルギー」レベルは独唱の場合より合唱の場合のほうが高い。プロのビデオ編集者は、これを感知して、独唱の部分よりも合唱の部分により速い編集ペースを使用する。

－　視覚素材の選択も音楽の一般的特性により影響を受ける。例えば、より明るい色またはより速い動きのビデオはより大きなエネルギーで音楽を伴奏するように選択することができ、もっと暗い色またはもっと遅い視覚素材はもっと遅いかまたはもっと静かな音楽で伴奏するように選択することができる。

－　ビデオ内のカットのタイミングおよび他の遷移は、通常、音楽のビートまたは音楽の

有意の特徴のタイミングと同期している。これは、「ビートに合わせてのカッティング」と呼ぶことがあり、ビデオ素材が音楽的基礎に基づいて編集される場合に広く使用される。

－　変化する度合いに従って、モーション・ビデオのショット内のイベントのタイミングも、音楽のビートまたは音楽の有意の特徴のタイミングと同期させることができる。このことは、対象物間の衝突のような急激な減速を含む運動イベントの場合に特に当てはまる。例えば、プロの編集者が落下する物体が床に衝突するショットを処理している場合には、その編集者は、恐らく、この瞬間を強いビートまたは音楽の他の顕著なイベントと整合させるだろう。

－　さらに、ビデオに適用される特殊効果の選択およびタイミングは、多くの場合、音楽の特性により影響を受ける。例えば、時間を合わせて強烈なビートまたは他の顕著な音楽的イベント内にフラッシュを入れることができるし、または、短い凍結フレーム効果を音楽の静かな瞬間に適用することができる。もっと長い時間的尺度の場合、いくつかの視覚的効果を音楽全体に適用することができる。例えば、ポップスを伴奏する音楽ビデオの場合には、独唱部分の視覚素材はモノクロで表示することができ、一方、合唱部の視覚素材は全カラーで表示することができる。

－　テキストおよびグラフィックスのようなオーバレイは音楽の特性により影響を受ける場合がある。例えば、これらの要素が表示または消失する時間を強烈なビートまたは他の顕著な音楽的イベントにリンクさせることができる。その様子および動きが音楽に依存するように、上記要素を音楽に合わせて動かすことさえできる。例えば、各音楽的ビートに合わせて異なる位置の間をジャンプするようにまたは音楽的構造に関連してある時間に大きさおよび色が変化するように上記要素を動かすことができる。

【０１１３】

要するに、視覚素材を音楽とマッチするように編集する場合には、プロの編集者は、音楽的ビートの「ミクロ構造」またはビートのさらに小さな分割部分から音楽の部分からなる主要な部分の「マクロ構造」まで時間的尺度のある範囲を横切って使用できる技術のレパートリーを持つ。これに成功した場合には、視聴者に与える効果が強くなる。音楽およびビデオは一体化した制作物と知覚される可能性が高くなり、情緒的または劇的インパクトが強くなる。

【０１１４】

この実施形態は、以下に説明するいくつかの方法により、音楽をベースとする制作物の生成を自動的に行う。

【０１１５】

音楽をベースとする制作物のための自動化

音楽アナライザ［１１６］および音楽記述［１１７］の性質について説明してきたし、音楽をベースとする制作物の生成を自動的に行うことができるか容易に行うことができるいくつかの方法についても説明してきた。本発明のこの態様について以下にさらに説明する。

【０１１６】

編集スタイルを音楽の構造とマッチさせる１つの簡単な方法は、音楽記述のパラメータから、直接出力制作物の視覚的性質を定義しているパラメータの編集を制御する方法である。例えば、カッティング速度（平均セグメント持続時間の逆数）、ディゾルブするためのカットの比率を制御するために使用するビート強度、および入力ビデオから選択したセグメントの輝度を制御するために使用する音量を制御するために、音楽のテンポを使用することができる。この種の簡単なマッピングの場合には、ユーザが速いテンポの音楽部分を選択した場合には、高速カット出力制作物が出来上がる。または、他の例で説明すると、ユーザが対照的に大きくて静かな音楽部分を選択した場合には、出力制作物は対応する明暗の部分を持つことになる。

【０１１７】

場合によっては、このアプローチは効果的であり、本発明を使用することによりサポート

することができる。例えば、ユーザが、これらのスタイルを選択することにより、この動作モードを選択することができるように、いくつかのスタイルでこのアプローチを実施することができる。しかし、このアプローチにはいくつかの制限がある。何故なら、このアプローチは音楽に対するほとんどすべての制御を放棄するからである。例えば、音楽が非常に単調なものである場合には、出力制作物は単調なものになる可能性がある。何故なら、種々の変化を導入するための上記機構が動作しないからである。逆に、音楽が非常に高速な対照を持っている場合には、出力制作物は一貫性のないものになる恐れがある。それ故、このアプローチは異なる音楽部分に対してどちらかといえば脆い面がある。このアプローチは、いくつかの音楽部分に対して許容できる出力制作物を生成することができるが、広い範囲の音楽部分に対してうまくいくという保証はない。

【０１１８】

もっと優れた代わりのアプローチは、音楽の特性によりスタイルおよび／またはサブスタイルを選択するという方法であるが、スタイル情報が個々の編集の決定を制御したりそれに影響を与えたりする恐れがある。このアプローチは任意の音楽入力に対してもっと予測可能で首尾一貫した結果を生成する。何故なら、すべての編集決定をスタイル情報が許可する範囲内に置くことができるからである。このアプローチを使用すれば、スタイル情報は、音楽が非常に単調なものである場合でも、例えば、上述した確率的発生および値循環の技術により種々様々なものを生成することができる。

【０１１９】

このアプローチは、図３のところで説明した本発明の中心である構成原理にもっと密接に適合している。音楽をベースとする制作物の場合に対して、このアプローチを、図１２を参照しながら、以下にさらに詳細に説明する。

【０１２０】

図３を参照しながら説明した前の例の場合のように、制作ロジック［１２０１］はスタイル情報［１２０２］、ビデオ／イメージ記述［１２０３］および音楽記述［１２０４］から情報を受け取る。これらの入力に応じて、制作ロジックはメディア・シーン・グラフ［１２０５］内に記憶する編集決定を生成する。この図は、それぞれが、かなり違った機能を実行する２つの部分、すなわち、マクロ記述［１２０６］およびミクロ記述［１２０７］から音楽記述を形成する方法を示す。

【０１２１】

音楽マクロ記述［１２０６］は導入部、独唱部、合唱部等のような音楽の主要な部分の時間的尺度における入力音楽の記述を含む。これらの部分の特性は、サブスタイル・シーケンス［１２０８］を生成するために使用される音楽セクション記述子のセットにより表示される。すでに説明したように、サブスタイル・シーケンスは、出力制作物を生成するためにサブスタイルを使用する順序を定義する。サブスタイル・シーケンスが確立されると、出力制作物内の任意の時間に対して対応するサブスタイルが存在する。それ故、出力制作中の特定の時間に対して編集情報が必要な場合には、この情報は正しいサブスタイルにより供給される。

【０１２２】

音楽ミクロ記述［１０２７］の役割について以下に説明する。すでに説明した入力音楽が存在しない場合に戻って説明すると、スタイル／サブスタイルから制作ロジック［１２０１］へ送られた情報は効果的に編集コマンドのセットであり、制作ロジックは、可能な場合には、これらのコマンドに従おうとする。（これは何時でも可能なわけではない。何故なら、ある種の決定はビデオ／イメージ記述に依存しているからであるが（ビデオ・セグメント選択に関する上記説明参照）、通常は可能であり、その場合には、制作ロジックはコマンドに従う。）

【０１２３】

しかし、音楽をベースとする制作物の場合には、サブスタイルが制作ロジックに送る情報は一組の優先順位である。これらの優先順位は音楽ミクロ記述［１２０７］からの音楽のローカル的特徴を考慮した後でだけそれに従うべきものである。ミクロ記述はバー、ビー

トおよびサブビートの時間的尺度のところの入力音楽の記述を含む。この記述は、一連の「編集ヒント」を含むこともできるし、一連の「編集ヒント」を生成するために使用することもできる。例えば、音楽振幅記述子から直接入手することができるある種の編集ヒントは、音楽の強いビートのところのようなある時点で出力制作物内のセグメント遷移を行うことが望ましいことを示す。

【０１２４】

サブスタイル・シーケンスが生成されると、制作ロジック［１２０１］は、出力制作物の冒頭のところからスタートし、出力制作物の終わりまで、下記のようにＭＳＧを形成することができる。

【０１２５】

－　この時点に対応するサブスタイルから出力制作物の現時点に関する編集の優先順位の入手。

－　音楽ミクロ記述［１２０７］から（出力制作物の現時点に直接関連する）入力音楽の現時点に関する編集ヒントの入手。

－　必要な場合には（セグメント選択に関する決定を行う場合）、ビデオ／イメージ記述［１２０３］からの記述子の値の入手。

－　これらの入力を結合することによる編集の決定とＭＳＧ［１２０５］内への編集の決定の記憶。

【０１２６】

例を挙げて上記２つの主要な態様について以下にさらに詳細に説明する。最初に、音楽のマクロ構造にマッチするサブスタイル・シーケンスの生成方法について説明し、２番目に、編集の決定を行うために構造体が編集の優先順位を編集ヒントと結び付ける方法について説明する。

【０１２７】

音楽のマクロ構造にマッチするサブスタイル・シーケンスの生成

音楽のマクロ構造にマッチするサブスタイル・シーケンスを生成するために使用する一般的な原理は、記述子の照合による入力ビデオ・セグメントの選択のところですでに詳細に説明した技術に類似の技術である記述子の照合の使用である。

【０１２８】

このプロセスの目標は、図１３に示す例のような音楽構造体にリンクしているサブスタイル・シーケンスを生成することである。これは、多くのポピュラー・ソングに存在する構造体、すなわち、導入部、独唱部１、合唱部等の後の一連の音楽セクション［１３０１］を示す。これらは、１対１の関係で一組のサブスタイル［１３０２］と照合される。この例におけるこれらのサブスタイルのシーケンスＳＳ３、ＳＳ２、ＳＳ４等はサブスタイル・シーケンスである。

【０１２９】

先に進む前に、この例の２つの特徴に注目したい。第１の特徴は、同じ音楽または類似の音楽に遭遇する度にその音楽は同じサブスタイルにリンクされることである。例えば、この場合、合唱部は何時でもＳＳ４にリンクされる。音楽セクションが非常に類似している場合には、このようなリンクは通常望ましいことであり、これから述べる手順は、同じような多くの場合に、このような結果になる。第２の特徴は、使用する特定のスタイルのすべてのサブスタイルに対して要件がないことである。この図には「ＳＳ１」がないが、それはこの特定の音楽部分に対してサブスタイル１が選択されなかったことを意味する。

【０１３０】

図１４は、音楽の構造体からサブスタイル・シーケンスを自動的に入手することができる１つの方法を示す。最初に、各音楽セクションに対して一組ずつ、一組の記述子の値が音楽記述［１４０１］から入手される。音楽セクションに対する適当な記述子は音楽セクションの持続時間、その平均的テンポ、音量およびビートの強度を含む。すでに述べた記述子のような多くの他の種類の記述子を使用することができ、すでに説明したように、これらの記述子は、マニュアル的に入力したか、任意の他の手段で生成した音楽制作物の副産

物として生成された信号分析により生成することができる。唯一の固定要件は、各音楽セクションに対する記述子のセットが音楽セクションのいくつかの知覚的に重要な品質の特徴を表すことである。

【０１３１】

次のステップ［１４０２］において、スタイル情報から、各サブスタイルに対して一セットずつ、対象記述子の値のセットが検索される。サブスタイルの対象値のセットが、このサブスタイルが特によく一致する音楽の特徴の記述である。通常、これらの値は、各サブスタイルに対する対象値のセットをマニュアル的に入力することにより、スタイル・デザイナーにより生成される。例えば、スタイル・デザイナーが、急速カット・サブスタイル（すなわち、すでに説明した、好適なセグメントの持続時間に対する小さな値を含んでいるかまたは生成するサブスタイル）を生成した場合には、スタイル・デザイナーは、このサブスタイルが、テンポおよびビート強度記述子に対しては高い値を示すが、音量に依存しない音楽セクションに最もよく適していると定義することができる。

【０１３２】

次のステップ［１４０３］においては、音楽セクションとサブスタイルとの間の記述子スペース内の距離のセットの計算が行われる。この計算は、シーケンシャルでない場合の入力ビデオ・セグメントの選択のところで説明したプロセスに類似していて、近接の計算を最適化するために導入した技術（図７参照）もこの場合に適用することができる。上記距離のセットから各音楽セクションに最も近いサブスタイルを割り当てることによりサブスタイル・シーケンスの「トライアル」バージョンを生成することができる［１４０４］。

【０１３３】

次のステップ［１４０５］において、望ましくない反復に対するサブスタイル・シーケンスのチェックが行われる。このチェックが必要な理由は、（図７のところで説明した）記述子分布正規化のような技術を適用した場合でもあまりに多くの音楽セクションに対して同じサブスタイルがマッピングされるという事態が起こり得るからである。連続しているが異なる２つの音楽セクションに同じサブスタイルがマッピングされた場合には、このようなことは特に望ましくない。図１３について説明すると、すでに説明した例の場合には、同じサブスタイルの連続発生だけが、発生するＳＳ４［１３０３］の３つの発生であることに留意されたい。何故なら、合唱部が３回繰り返されるからである。これは反復が必要な場合であるが、この例の中の任意の他の反復は恐らく望ましいものではない。このような望ましくない反復は、多くの場合、例えば、１つのサブスタイルの発生の全数がある値を超えたかどうかまたは連続反復の全持続時間がある時間の値を超えたかどうかをチェックすることにより検出することができる。

【０１３４】

このような望ましくない反復が発見された場合、それらの反復は、サブスタイル・シーケンス内のサブスタイルの中のあるものを、上記ステップ［１４０３］で発見した各音楽セクションに対するサブスタイルに次に近い別のサブスタイルにより置き換えることにより除去される［１４０６］。

【０１３５】

この技術は図６および図７を参照しながらすでに説明した入力ビデオ・セグメントを選択するための技術に類似しているので、上記の多くの詳細な点および別の技術もこの場合適用することができる。

【０１３６】

編集決定を生成するための編集優先と編集ヒントとの結合

図１５は、編集決定を生成するためにスタイル／サブスタイル情報からの編集優先順位を音楽マクロ記述からの編集ヒントと結合するためのある技術の図面である。この技術は音楽的ビートの時間の尺度のところで動作する。この技術はカットの決定（すなわち、セグメントにある変化をさせなければならない、出力制作物中の時間的な位置の識別）を行うための技術であるが、この技術またはそれを修正したものは、フラッシュまたは他の特殊効果を挿入する時間的位置の識別のような他の種類の編集決定を行うためにも使用するこ

とができる。

【０１３７】

この例の場合には、水平軸が時間軸であり、垂直方向の矢印［１５０１］は音楽マクロ記述から受信または入手した編集ヒント・パルスである。これらの矢印の高さは音楽の知覚的に重要な特徴に関連していて、その水平位置はスタート時間ｔ＝０からのそれらが発生した時間を示す。通常、問題の特性はオーディオ信号の振幅の変動からの信号のような音楽的ビートに密接に関連する特性である。音楽的ビートのこのような表現を自動的に抽出するための多くの技術は当業者にとって周知のものである。例えば、全振幅または信号のある周波数帯の振幅に対してしきい値超えテストを行うことができる。位相ロック・ループの使用のようなさらに改良した技術は、検出機構を、ほとんどのポピュラー音楽の場合のように、ビートが規則的である場合に発生する振幅変動の周期と同期させることができる。この技術を使用する場合は何時でも、編集ヒント・パルスは下記の傾向を持つことが望ましい。

【０１３８】

－　大多数がビートまたは１／２、１／４、１／３等のようなビートの分数と一致すること。

－　各バーの第１のビートのような強いビート上に発生するパルスがより大きい値を持つこと。

－　（主要なビート間に発生する）オフ・ビート・パルスの値が強いオフ・ビート音楽的イベントが存在する場所で大きな値を持つこと。例えば、「シンコペートした」と呼ばれる音楽のスタイルでのように、このようなことは多くの音楽で通常起こることである。

－　通常、人間が知覚するようなリズムにパルスが対応すること。

【０１３９】

この場合、制作ロジックは、各編集ヒント・パルスを、対応する時間のところでカットを行うようにとの要求と判断し、各パルスの高さを要求の強さであると判断する。パルスの高さは０～１のような範囲に制限することができる。図１５はそのような例を示す。

【０１４０】

しかし、制作ロジックはスタイル／サブスタイル情報も考慮に入れなければならない。スタイルが指定する１つのパラメータは、すでに説明したように、「カッティング速度」である。この例に関連しているものは、スタイル情報が出力制作物内の任意の瞬間に対して出力制作物の次のショットに対する好適なセグメント持続時間を指定することである。この好適な持続時間は、図１５においては、$t_{preferred}$で示してあるが、もっと一般的には、４本のライン・セグメント［１５０２，１５０３，１５０４および１５０５］で表示される。これら４つのセグメントは編集ヒント・パルスに適用されるしきい値を形成する。しきい値は$t_{preferred}$のところで最小になる。しきい値は、また、$t < t_{min}$および$t > t_{max}$に対して１の最大許容パルス値をとる。このことは、t_{min}とt_{max}との間に位置するパルスだけがしきい値を超えることができることを意味する。

【０１４１】

この機構の動作を完全に理解するためには、さらに２つの事実が必要になる。

【０１４２】

－　ゼロ時間、ｔ＝０は、前のカットに対応する。すなわち、ゼロ時間は現在のビデオ・セグメントのスタート時間である。制作ロジックがセグメント毎に出力制作物を生成すると、ゼロ時間は各セグメントに対してリセットされる。

－　選択したセグメント持続時間は、値$v_x = v_p - v_{th}$が最大であるパルスのｔ＝０からの経過時間である。ここで、v_pはパルスの値であり、v_{th}はパルスの時間のところのしきい値の値である。すなわち、最大値によりしきい値を超えるパルスの時間、または、しきい値を超えるパルスがない場合には、それに最も近いパルスである。図１５の場合には、このパルスはパルス［１５０６］である。パルス［１５０７］はもっと高い値を持つが使用されないことに留意されたい。何故なら、値v_xがパルス［１５０６］に対す

るものより大きいからである。
【０１４３】
上記要因すべてを考慮に入れた場合、このしきい値機構が下記の行動を行うことを理解することができる。
【０１４４】
－　強い編集ヒント・パルスに対応する持続時間に有利である。すなわち、すでに説明したように、ビートおよび音楽の他の特徴に関連するカットを行う傾向がある。
－　好適なセグメント持続時間の近くのパルスにとって有利である。特に、音楽が非常に静かで、その結果、編集ヒント・パルスが非常に弱い場合、または、音楽が相対的に特徴がなく、その結果、すべての編集ヒント・パルスが同じような強さを持っている場合には、$t_{preferred}$に非常に近い持続時間を選択する。
－　t_{min}とt_{max}の間の持続時間を常に選択する。
－　t_{min}とt_{max}の間の距離を変化させることにより、音楽的リズム（編集ヒント・パルス）および好適なセグメント持続時間の相対的な影響を制御することができる。t_{min}とt_{max}の間の距離が接近している場合には、好適なセグメント持続時間が優勢になる。t_{min}とt_{max}との間の距離が離れている場合には、音楽的リズムが優勢になる。これは、異なるスタイルでまたは１つのスタイルの異なるサブスタイルですら異なるように設定できる要因である。$t_{preferred}$に対してt_{min}とt_{max}との位置を変化させることにより、さらに制御を行うことができ、$t_{preferred}$の近くに強いパルスが存在しない場合には、持続時間をもっと長くすることもできるしもっと短くすることもできる。さらに、この機構を修正したものは、ライン・セグメントが曲線で置換されている非線形しきい値を使用することができ、行動をもっと細かく制御する。
【０１４５】
多くの場合、$t_{preferred}$の値を、例えば、１ビート、１／２ビート、２ビート等のような現在の音楽テンポのビート速度に関連する持続時間に設定すると効果的である。また、多くの場合、コンストラクタは、編集決定内に変化、すなわち、サブスタイル、ゆっくりした展開、統計的発生および値サイクリングを生成するために、すでに説明したような技術により出力制作物内を進んでいく間に、各セグメントに対する異なる値を$t_{preferred}$に割り当てることに留意されたい。
【０１４６】
この節で説明した一組の技術を結合することにより、本発明は、音楽のリズムに関連して知覚され、音楽が非常に変化の少ない場合でも十分変化し、かつ、選択した音楽が何であれ何時でも許容できる範囲内に位置する編集決定を生成することができる。
【０１４７】
音楽をベースとする制作物の生成を自動化するための他の機能
そうしたい場合には、本発明を、例えば、音楽をベースとする制作物の生成を自動化し容易にするための下記のいくつかの他の機能で強化することができる。
【０１４８】
そのような機能とは例えば下記の機能である。
【０１４９】
－　音楽をベースとする制作物の場合、入力サウンドトラックまたはその一部内でミキシングすることが望ましい場合がある。１つのオプションは、一定の状態の相対的なレベルで全入力サウンドトラックを入力音楽とミックスする方法である。もう１つのオプションは、一方または他方が何時でもハッキリと聞こえ他方により聞きにくくならないように、入力サウンドトラックまたは入力音楽または両方のレベルを変化する方法である。例えば、この方法は、オーディオのプロにとっては周知であり、かつ、アナウンサが喋る場合には何時でも音楽のレベルを下げるために生のラジオ放送のような用途に広く使用されている「ダッキング」と呼ばれる技術を使用することができる。さらにもう１つのオプションは、音楽記述内の記述子の値により追加のオーディオ要素を使用したりしなかったりする方法である。例えば、入力音楽が歌であり、入力サウンドトラックが声を含んでいる普通

の場合には、声を歌声と同時にミックスすると、一般に、混同したり混乱が起こったりする。それ故、音楽の楽器だけの演奏のような歌声がない場合だけ入力サウンドトラックからのオーディオ内でミックスするのが望ましい。音楽記述が（すでに説明したように）インポートされた要素を含んでいる場合には、このようなミキシングは、歌声が含まれているかいないかを示すマニュアル的に生成した記述子を使用して行うことができる。これを自動化するために音楽アナライザに内蔵させることができる音楽内の歌声の存在を検出するための周知の信号分析技術も存在する。今説明した技術と一緒に使用することができる、入力サウンドトラックからのオーディオのミキシング・インを制御するためのもう１つの方法は、そのオーディオ特性によりサウンドトラックの領域を選択する方法である。例えば、当業者にとって周知の音声検出アルゴリズムを、他の音響に対して声の方が優勢なサウンドトラックの領域だけを選択するために使用することができる。逆に、音楽検出アルゴリズムは、確実に、音楽を含んでいるサウンドトラックのセクションが選択されないようにするために使用することができる。このことは望ましいことである。何故なら、サウンドトラック内の音楽は、通常、入力音楽とミックスすると不快な効果を発生するからである。これらのプロセスを自動化するためのオーディオ分析技術は完全に信頼できるものではない。例えば、周知の技術はすべてのタイプの音楽で完全に正確に歌声の存在を検出することはできない。しかし、このような技術は、本発明には十分に役に立つ働きをし、特に、（すでに説明したように）ユーザの仕上げがサポートされている実施形態の場合には、十分に役に立つ働きをする。

－　音楽をベースとする制作物において、プロの編集者は、多くの場合、落下物が地面に衝突する瞬間のような有意な機能のタイミングが音楽の注目すべき機能のタイミングと同期するようにビデオ要素をどのようにして整合するのかを説明してきた。これは、ビデオ運動分析のための周知の技術を、すでに説明したビート検出技術のような音楽の特徴を検出するための技術と結合することにより自動化することができる。例えば、運動ベクトルをブロック照合のような標準技術によりビデオから抽出することができ、衝突のような急激な減速のタイミングをフレームの領域内の運動ベクトルのスカラまたはベクトルの合計の急激な変化が起こる時点を識別することにより確立することができる。これら減速モーメントの１つまたはそれ以上の時間が入力ビデオのショット内で確立され、各減速の大きさが確立されると、最善の一致が存在するビデオと音楽との間の相対的タイミングを発見することによりショットを音楽に最適な状態で整合させることができる。これは、出力制作物のセグメントの持続時間中に計算した、ビート強度による減速の数学的相互関係が最大になる相対的な時間として定義することができる。

【０１５０】

ビート強度および音楽アナライザが入手する他の記述子は、テキスト／グラフィック・オーバレイのアニメーションを制御するために使用することができる。例えば、その位置、向き、大きさ、スキューイング、色等のようなオーバレイのパラメータを音楽信号の振幅により直接決定することができる。または、もっと高性能の実施形態の場合には、（すでに説明したように）しきい値超えテストに基づく音楽的ビートの表現をオーバレイのパラメータの急激な変化をトリガするために使用することができ、次に、オーバレイはどちらかといえばもっとゆっくりとそのデフォールト位置に弛緩させることができる。すなわち、アニメーションを、音楽信号からのパルスにより励起され、音楽的ビートに関連付けられる弛緩モデルに基づいて行うことができる。さらに、すでに説明した音楽セクション記述子を、セクション境界と整合していて、各セクションの音楽的特徴に関連するアニメーションの行動内の変化を制御するために使用することができる。例えば、大きな音楽中に発生するオーバレイを大きくし明るくしギクシャクとした方法で運動させ、一方、静かな音楽中に起きるオーバレイを小さく暗く滑らかに運動するように、上記のように運動するオーバレイ・テキスト／グラフィックの色、大きさおよび弛緩速度を現在の音楽セクションの平均音量に比例したものにすることができる。

【０１５１】

制作の作業の流れに対する変更

この最後の節においては、図１６および図１７を参照しながら、メディア制作に従事しているユーザに対する作業の流れを、本発明の典型的な実施形態がどのようにして変更するのかを説明する。これら２つの図面においては、点線で示したステップは、通常、自動化により自動化されているか、自動化により容易に行うことができるステップである。

【０１５２】

図１６は、入力ビデオから音楽をベースとする出力制作物を生成するために非線形ビデオ・エディタ（ＮＬＥ）のようなツールを使用する従来の典型的な場合の作業の流れを示す。最初に、入力ビデオが捕捉および／またはインポートされる［１６０１］。このステップは、通常、コンピュータに取り付けられたカメラによりビデオを記録するステップ、または、ビデオ・カムコーダから前に記録したビデオ素材をコンピュータに転送するステップ、または、デジタル・ビデオ・ファイルの形でビデオを入手するステップを含む。アナログ・カムコーダのようなアナログ記録デバイスを使用する場合には、このステップは入力信号のデジタル化ステップも含む。これらの他のシナリオ中の任意なものの場合には、このステップが完了した場合、入力ビデオ素材はＮＬＥ内に導入済みである。

【０１５３】

この例は音楽をベースとする制作物に関連しているので、ユーザは、また、例えば、音楽を記録することにより、またはそれをオーディオＣＤのような音楽的媒体から転送することにより、または音楽をデジタル・オーディオ・ファイルとして入手することにより捕捉／インポートしなければならない［１６０２］。これらの別のシナリオのどれかの場合、このステップが終了した場合、入手音楽はＮＬＥ内に導入済みである。

【０１５４】

いくつかのＮＬＥは、次のステップ［１６０３］を自動的に実行することができ、カラー・ヒストグラム内の突然の変化の検出のような技術により入力ビデオをショットに分割する。ショットは、通常、「クリップ」のセット、すなわち、入力ビデオの小さなセグメントとしてユーザに表示される。ＮＬＥが自動ショット細分化を含んでいない場合には、ユーザは入力ビデオをマニュアル的にセグメントに分割する。

【０１５５】

次に、ユーザは自分を入力ビデオのショットに慣れさせなければならない。これは、通常、ショットを「ロギング」することにより［１６０４］、すなわち、ショットをグループ内で組織化するか、またはある順序に配列し、各ショットについてのノートをとり、いくつかのショットを拒否する等して行われる。多数の入力素材を含むプロの制作物の場合には、これは通常は時間の掛かる作業である。短い軽い制作物の場合には、ほとんどの場合この作業を行わなくてもよいが、そうすると、通常は、結果としての制作物の質が落ちることになる。

【０１５６】

次の３つのステップ［１６０５，１６０６，１６０７］はシーケンシャルに行うことができ、または、ユーザは（例えば、出力制作物の１つのセクションを完了し、次のセクションに移る前に）これらのステップの順序を変えることもできるし、または、ユーザはこれらのステップ間の境界をぼかすような方法で作業することもできる。どのアプローチをユーザが採用した場合でも、ユーザは出力制作物をセグメント毎にマニュアル的に制作しなければならないし、スタイリッシュな音楽をベースとする制作物が対象である場合には、ユーザは、セグメントが入力音楽のリズム、タイミングおよび「フィーリング」に適合するように、注意深くセグメントを操作しなければならない。このプロセスは、上記技術の多くのものを含み、ほとんどの場合、時間が掛かる作業で、多くの場合には１分間の出力制作物を生成するのに１時間または数時間も掛かる。自分が満足できる品質基準の出力制作物を生成するのは、多くのアマのユーザの技術レベルでは不可能であり、特に、音楽素材および視覚素材の理解を必要とする音楽をベースとする作品の場合には不可能である。

【０１５７】

自分が満足できる一組の編集決定ができたとユーザが考えた場合には、ユーザは、どの時点で出力制作物をビデオ・ファイルまたは他の出力として制作するのかをレンダリング［

１６０８］するのかをＮＬＥに命じる。ユーザはこれをチェックして、満足できない場合には［１６０９］、制作物を変更したりより洗練したものにするために前のステップ中のあるステップに戻る。

【０１５８】

最後に、ユーザは、出力制作物を、自分および他の人たちがそれを見ることができる形でエクスポートする［１６１０］。ほとんどの基本的な場合、ユーザは、自分で見るために自分のコンピュータでビデオ・ファイルを簡単に使用することができるが、もっと一般的には、ユーザは、ビデオ・カセット・レコーダでテープにコピーするかまたは書き込み可能なコンパクト・ディスク（ＣＤ－Ｒ）のような光ディスク・フォーマットにコピーする。例えば、それを電子メール・アタッチメントとして送信し、それを他の人がアクセスすることができるサーバにアップロードするか、またはそれをいわゆる「ピア・ツー・ピア」ファイル共有によりユーザのローカルマシーンから共有することにより、インターネットによりビデオ・ファイルを配布する方法が次第に普及してきている。

【０１５９】

図１７は、本発明のある実施形態に基づくシステムによる通常の音楽をベースとする制作物の場合の作業の流れである。これを図１６のところで説明した従来の作業の流れと比較されたい。

【０１６０】

捕捉ステップ／インポート・ステップ［１７０１および１７０２］は、従来のＮＬＥの場合の上記対応するステップ［１６０１および１６０２］と同じものである。ショット細分化ステップ［１７０３］も、本質的には、上記対応するステップ［１６０３］と同じものである。システムは、細分化を自動化するための１つまたはそれ以上の周知の技術を使用し、そうしたい場合には、ユーザが結果としての細分化を無視したり調整できるようにする。

【０１６１】

次に、ユーザは、コンテンツ（入力素材の要素）および／または素材の処理を予め選択する［１７０４］。本発明は、上記のこのプロセスを楽に行うことができるようにする技術を提供する。このステップはオプションであり、ある実施形態の場合には、このステップはスキップすることができる。

【０１６２】

次のステップ［１７０５］は、本明細書において詳しく説明してきた多くの種類の自動分析および制作（構成）を含む。このステップが終了すると、編集決定の完全なセットがすでに生成されていて、これら編集決定は出力制作物を完全に定義する。通常、このステップはシステムにより完全に自動的に行われるので、ユーザが介入する必要はない。

【０１６３】

ここで、システムは出力制作物をレンダリングする［１７０６］。ユーザはこの出力制作物をチェックし、満足しない場合には［１７０９］、すでに説明した技術に基づいてシステムの助けを借りて出力制作物を仕上げることもできるし［１７０７］、または前のステップ中のどれかに戻ることもできる。

【０１６４】

最後に、ユーザはその出力制作物をエクスポートする［１７１０］。このステップは従来のＮＬＥの場合の上記対応するステップ［１６１０］に類似している。

【０１６５】

図１６および図１７を見て上記説明を読めば、本発明の通常の実施形態の作業の流れは、より多くの自動化が行われていて、ユーザのマニュアル的な作業が少なくなっていることが分かるだろう。これにより、制作プロセスがスピードアップし、それに要するユーザの時間が短くなり、未経験のユーザに対するサポートが強化される。

【０１６６】

ハードウェア実施形態

当業者であれば、本発明を、汎用コンピュータ、携帯情報端末、専用ビデオ編集ボックス

、セットトップ・ボックス、デジタル・ビデオ・レコーダ、テレビジョン、コンピュータ・ゲーム・コンソール、デジタル・スチール・カメラ、デジタル・ビデオ・カメラ、およびメディア処理を行うことができるその他のデバイスを含む多くの種類のハードウェア・デバイスで実施することができることを理解することができるだろう。本発明は、その機能の異なる部分が２つ以上のハードウェア・デバイスに内蔵される複数のデバイスを備えるシステムとして実施することもできる。

【０１６７】

特定の実施形態を参照しながら本発明を説明してきたが、当業者なら理解できると思うが、本発明の範囲から逸脱することなしに本発明を種々に修正することができる。

【図面の簡単な説明】

【図１】

図１はこの発明の実施例の相関する機能のモジュールを示す。

【図２】

図２は図１に記載の実施例の動作の例を示す。

【図３】

図３は図１に記載の実施例の動作原理を模式的に示す。

【図４】

図４は好適なセグメントを導くために入力すべきビデオ素材を探索する図１の実施例を示す。

【図５】

図５は図４の例においてビデオセグメントを選択するための論理を示すフローチャートであって、出力制作物は入力された素材中に認められるセグメントの順序が維持されている。

【図６】

図６は図４の例においてビデオセグメントを選択するための論理を示すフローチャートであって、出力制作物は入力された素材中に認められるセグメントの順序が維持されておらず、しかしその代わりに、そのセグメントの記述子の値の相同性により選択される。

【図７】

図７は図１の実施例において用いられる論理を示すフローチャートであって、候補となるセグメントのセットと対象となる記述子の値のセットとの間の相同性の指標を計算する。

【図８】

図８はメディアシーングラフの構成を示し、このメディアシーングラフは図１の実施例により形成され、かつ出力制作物の様式の完全な代表若しくは出力制作物を作るための完全なインストラクションのセットである。

【図９】

図９は第１の、簡単なＧＵＩを示し、このＧＵＩは３人の主たるユーザのコントロールを伴う図１に示した実施例において好適に用いられる。

【図１０】

図９は第１の、簡単なＧＵＩを示し、このＧＵＩは５人の主たるユーザのコントロールを伴う図１に示した実施例において好適に用いられる。

【図１１】

図１０はこの発明の他の実施例をしめし、この実施例では通常の使用状態でユーザの干渉を必要としない。

【図１２】

図１２は図３に示された原理の詳細を示し、特に音楽による制作物の形成を示す。

【図１３】

図１３は図１の実施例の特徴を示し、その特徴は音楽の一片のマクロ構造とサブスタイルシーケンスが一対一の関係で適合する。

【図１４】

図１４は図１の実施例の一つの方法を示すフローチャートであり、ここにおいてサブスタ

イルシーケンスは入力音楽のマクロ構造と自動的に適合される。

【図１５】

図１５は閾値のメカニズムを示し、このメカニズムは編集決定を形成するためにスタイル情報からの編集優先度と音楽記述から導かれる編集ヒントとを合成する。

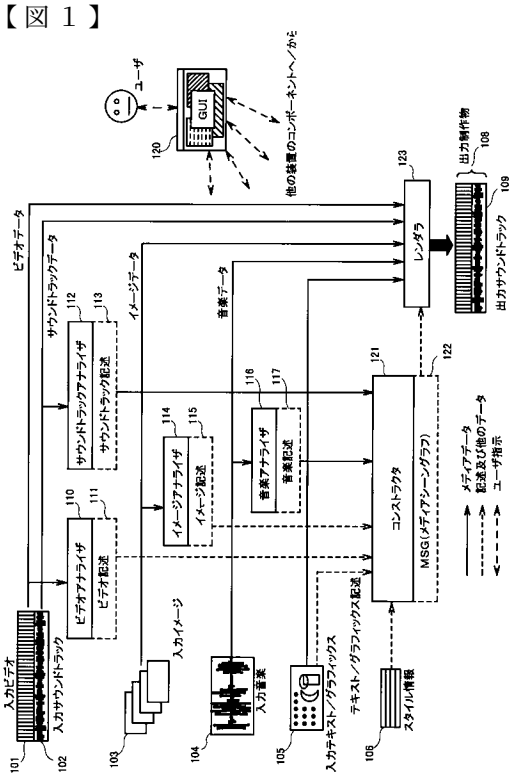
【図１６】

図１６はユーザが従来例の非線形ビデオエディタを用いてビデオ制作物を形成する典型的なワークフローを示す。

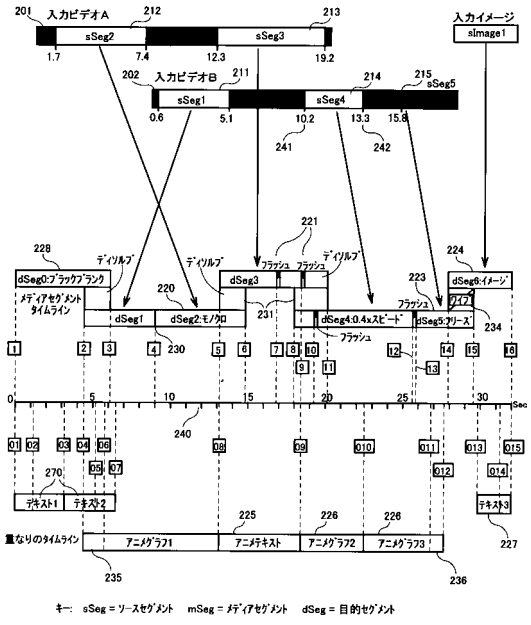
【図１７】

図１７はユーザが図１の実施例を用いてビデオ制作物を形成する典型的なワークフローを示す。

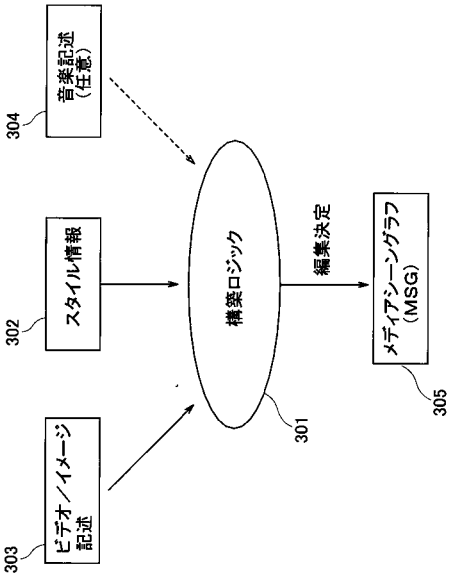
【図１】



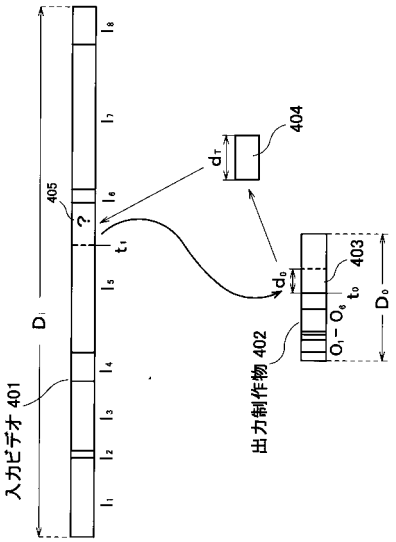
【図２】



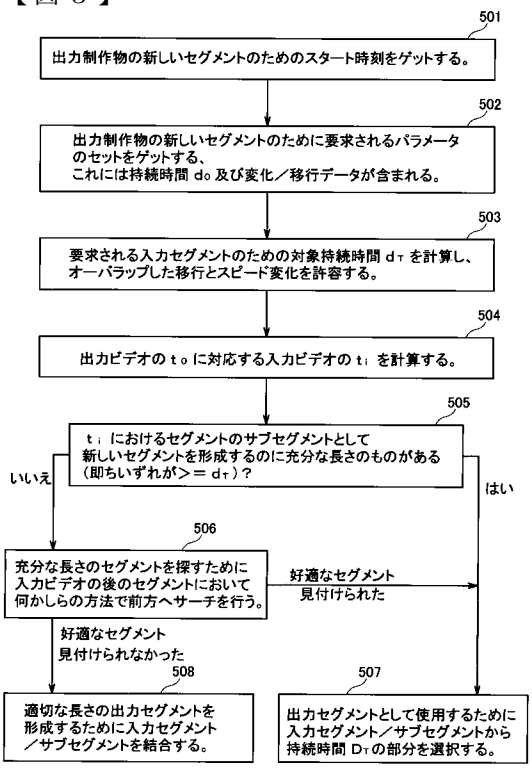
【図３】


音楽記述
(任意)
304
スタイル情報
302
ビデオ／イメージ
記述
303
構築ロジック
301
編集決定
メディアシーングラフ
(MSG)
305

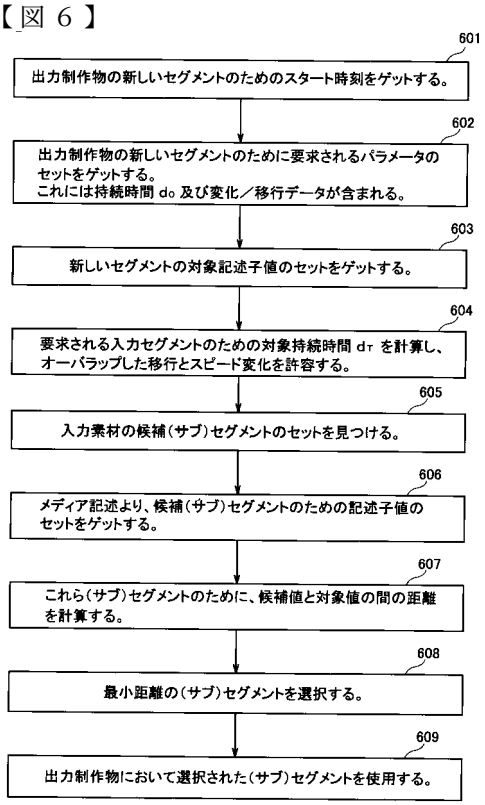
【図４】


入力ビデオ 401
出力制作物 402
403
404
405
?

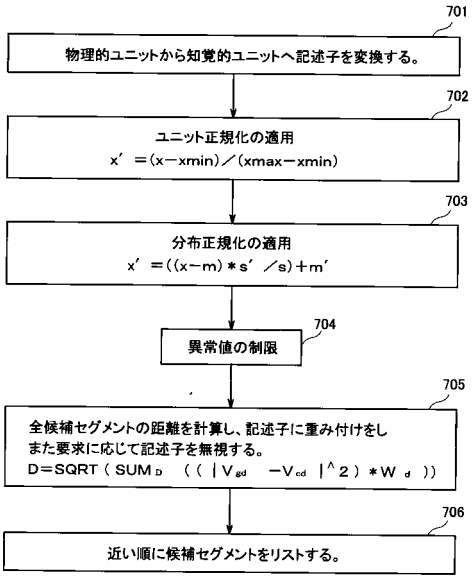
【図５】


501
出力制作物の新しいセグメントのためのスタート時刻をゲットする。
502
出力制作物の新しいセグメントのために要求されるパラメータのセットをゲットする、
これには持続時間 do 及び変化／移行データが含まれる。
503
要求される入力セグメントのための対象持続時間 dT を計算し、
オーバラップした移行とスピード変化を許容する。
504
出力ビデオの to に対応する入力ビデオの ti を計算する。
505
ti におけるセグメントのサブセグメントとして
新しいセグメントを形成するのに充分な長さのものがある
(即ちいずれが>＝ dT)？
いいえ
はい
506
充分な長さのセグメントを探すために
入力ビデオの後のセグメントにおいて
何かしらの方法で前方へサーチを行う。
好適なセグメント
見付けられた
好適なセグメント
見付けられなかった
508
適切な長さの出力セグメントを
形成するために入力セグメント
／サブセグメントを結合する。
507
出力セグメントとして使用するために
入力セグメント／サブセグメントから
持続時間 DT の部分を選択する。

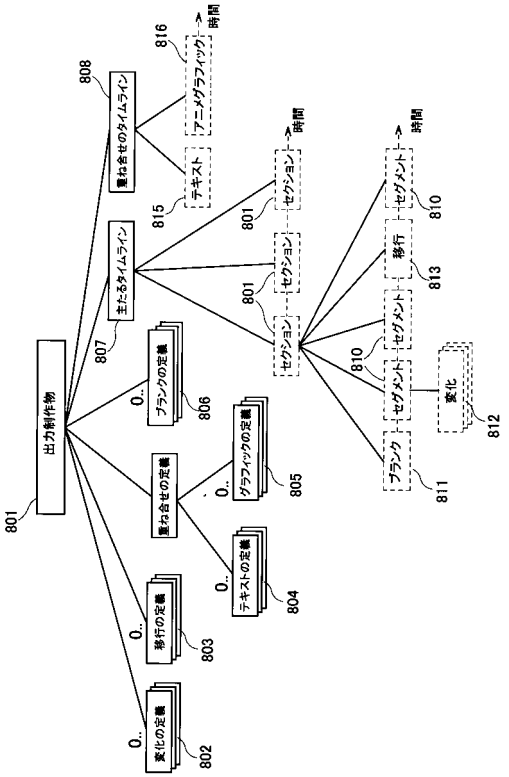
【図６】


601
出力制作物の新しいセグメントのためのスタート時刻をゲットする。
602
出力制作物の新しいセグメントのために要求されるパラメータのセットをゲットする。
これには持続時間 do 及び変化／移行データが含まれる。
603
新しいセグメントの対象記述子値のセットをゲットする。
604
要求される入力セグメントのための対象持続時間 dT を計算し、
オーバラップした移行とスピード変化を許容する。
605
入力素材の候補(サブ)セグメントのセットを見つける。
606
メディア記述より、候補(サブ)セグメントのための記述子値のセットをゲットする。
607
これら(サブ)セグメントのために、候補値と対象値の間の距離を計算する。
608
最小距離の(サブ)セグメントを選択する。
609
出力制作物において選択された(サブ)セグメントを使用する。

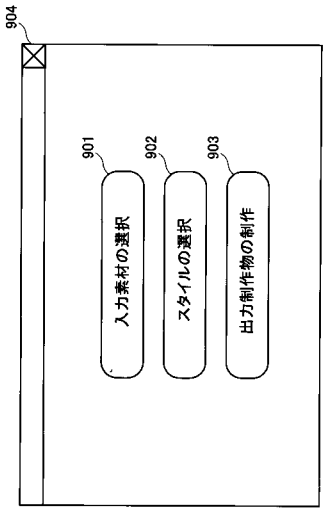
【図７】


701
物理的ユニットから知覚的ユニットへ記述子を変換する。
702
ユニット正規化の適用
x′ =(x−xmin)／(xmax−xmin)
703
分布正規化の適用
x′ =((x−m)＊s′ ／s)+m′
704
異常値の制限
705
全候補セグメントの距離を計算し、記述子に重み付けをし
また要求に応じて記述子を無視する。
D=SQRT (SUM D ((| V gd −V cd |^2) ＊W d))
706
近い順に候補セグメントをリストする。

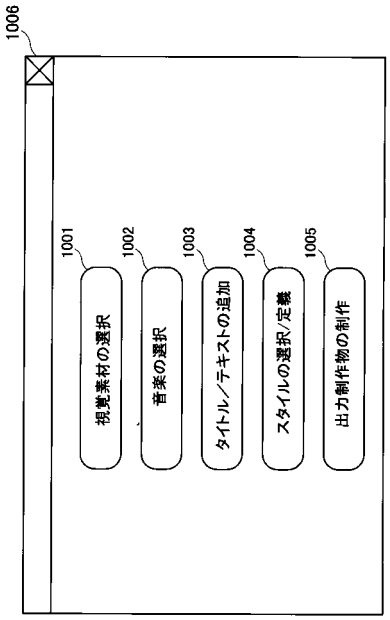
【図８】


出力制作物
801
808
重ね合せのタイムライン
816
アニメグラフィック
時間
テキスト
815
807
主たるタイムライン
801
セクション
セクション
セクション
時間
801
セグメント
移行
セグメント
セグメント
セグメント
時間
810
813
810
ブランク
811
変化
812
0..
ブランクの定義
806
重ね合せの定義
0..
グラフィックの定義
805
0..
テキストの定義
804
0..
移行の定義
803
0..
変化の定義
802

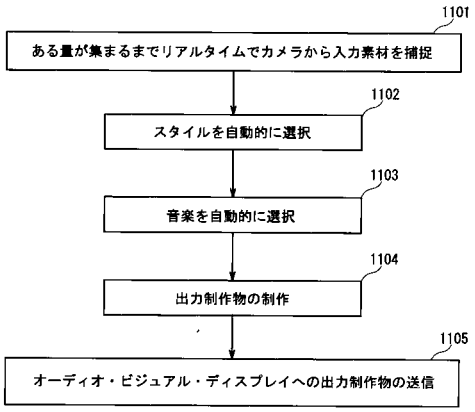
【図９】


904
901
入力素材の選択
902
スタイルの選択
903
出力制作物の制作

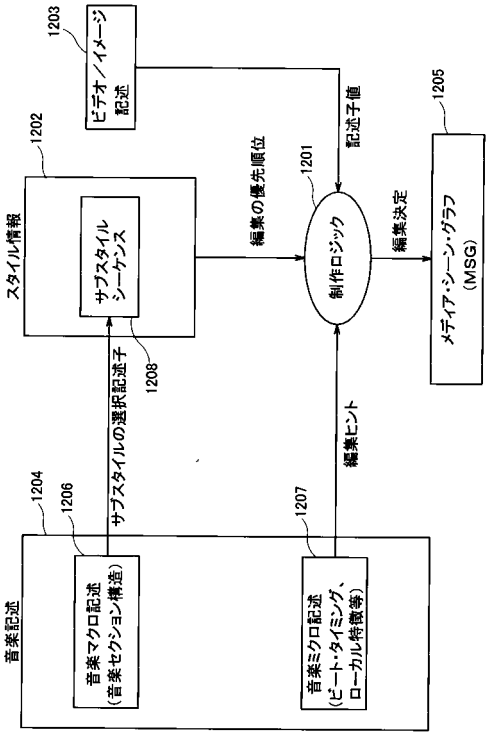
【図１０】


1006
1001
視覚素材の選択
1002
音楽の選択
1003
タイトル／テキストの追加
1004
スタイルの選択／定義
1005
出力制作物の制作

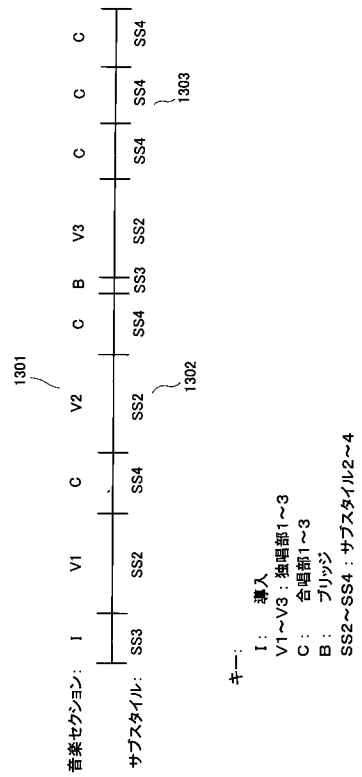
【図１１】



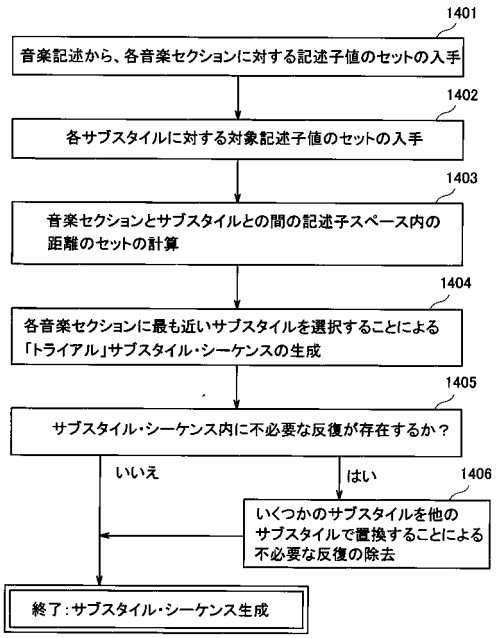
【図１２】



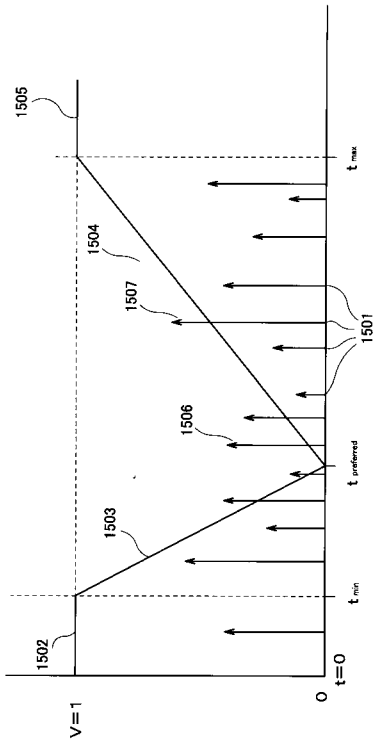
【図１３】



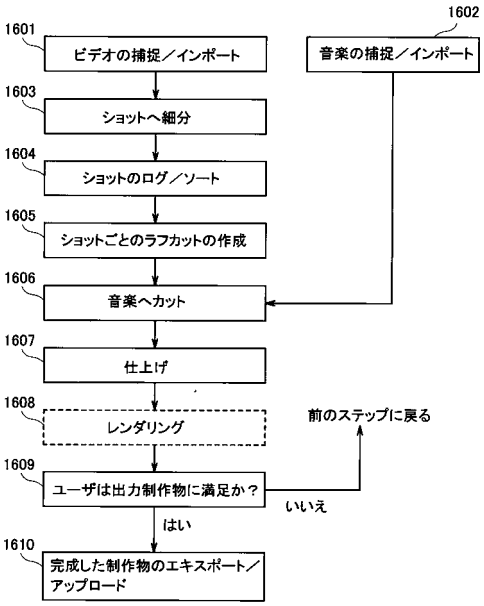
【図１４】



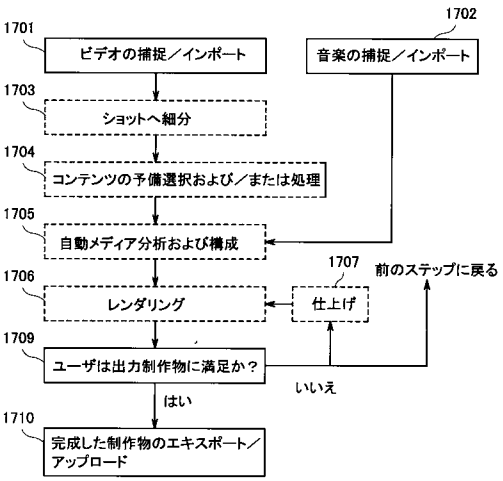
【図１５】


1505
1504
1507
1501
1506
1503
1502
t max
t preferred
t min
t=0
0
V=1

【図１６】


1601
ビデオの捕捉／インポート
1602
音楽の捕捉／インポート
1603
ショットへ細分
1604
ショットのログ／ソート
1605
ショットごとのラフカットの作成
1606
音楽へカット
1607
仕上げ
1608
レンダリング
前のステップに戻る
1609
ユーザは出力制作物に満足か？
いいえ
はい
1610
完成した制作物のエキスポート／アップロード

【図１７】


1701
ビデオの捕捉／インポート
1702
音楽の捕捉／インポート
1703
ショットへ細分
1704
コンテンツの予備選択および／または処理
1705
自動メディア分析および構成
1707
前のステップに戻る
1706
レンダリング
仕上げ
1709
ユーザは出力制作物に満足か？
いいえ
はい
1710
完成した制作物のエキスポート／アップロード

【国際公開パンフレット】

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 July 2002 (04.07.2002)

PCT

(10) International Publication Number
WO 02/052565 A1

(51) International Patent Classification⁷: G11B 27/031, H04N 5/91, G06F 3/14

(21) International Application Number: PCT/SG00/00197

(22) International Filing Date:
22 December 2000 (22.12.2000)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **KENT RIDGE DIGITAL LABS** [SG/SG]; 21 Heng Mui Keng Terrace, Singapore 119613 (SG).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KELLOCK, Peter, Rowan** [GB/SG]; 97a Upper Thomson Road, #08-02 Lakeview, Singapore 574327 (SG). **ALTMAN, Edward, James** [US/SG]; 41 Hume Avenue, #05-12 Symphony Heights, Singapore 598738 (SG).

(74) Agent: **GREENE-KELLY, James, Patrick**; Lloyd Wise, Tanjong Pagar, P.O. Box 636, Singapore 910816 (SG).

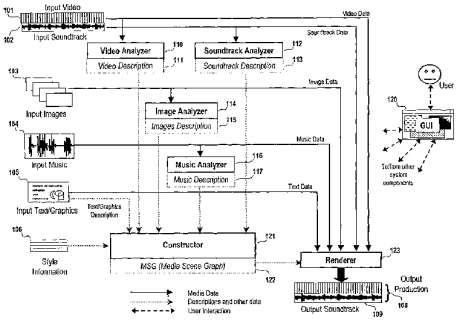
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
with international search 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 MEDIA PRODUCTION



(57) Abstract: An editing system is proposed for automatically, or semi-automatically, editing input data to generate output production. The input material is annotated by, or analyzed to derive, a set of media descriptors which describe the input material and which are derived from the input material. The style of editing is controlled using style data which is optionally derived from a user. The input material may include any or more of motion video, still images, music, speech, sound effects, animated graphics and text. The style data and the descriptors are used to generate a set of operations which, when carried out on the input data, generate an edited output production.

WO 02/052565 A1

System and Method for Media Production

Field of the invention

The invention relates generally to computer generation of media productions. In particular, the invention relates to automated or semi-automated editing of productions including any one or more of the following media: motion video, still images, music, speech, sound effects, animated graphics and text.

Background of the invention

Today, analogue media are gradually being replaced by digital media. In the case of audio this transition has already largely taken place, and it is well underway for image, video, graphics animation and other media. As these media become digital and the capability/cost ratio of computing resources continues to increase, new users and markets for digital media production are opening up. Of particular relevance to this invention are emerging markets for *casual* media production and especially casual *video* production, i.e. cases where the high cost of professional video production would preclude its use and where, until recently, the cost of the necessary equipment was too high. These include home video production (e.g. of holidays, weddings, etc), some informal corporate uses (e.g. internal communications and team-building), use by societies and other organizations, etc.

The concept of casual or "desktop" video production has existed for about a decade, but widespread adoption has been held back by a number of problems. These include:

1. Problems of technical infrastructure: inconvenience and loss of quality when digitizing video from a camera, limited hard disk space, insufficient processing power, etc.

2. The lack of convenient, low-cost distribution mechanisms: until recently the only widespread formats have been videotapes, but the cost and time involved in duplication and distribution preclude many potential applications.

3. The time and expertise required to make acceptable-quality productions, particularly at the stage of editing and "post-production".

The first and second of these problems are today disappearing thanks to technologies such as DV cameras, the IEEE 1394 ("Firewire") interface and video distribution on the world-wide web.

This invention attempts to address the third problem, allowing automated or semi-automated editing of digital media, particularly video.

Today, the main tool used for editing video is the "Non-Linear video Editor" or NLE. These are computer programs which adopt many paradigms from conventional editing methods such as film cutting and linear dub-editing using video tape machines. They employ manual methods of editing which are well suited to scenarios where the user is experienced and the desired result is a high-quality video production. There are many products of this type including Premiere from Adobe Inc., and iMovie from Apple Inc.

The NLE is a considerable advance on earlier technology, yet there remain many scenarios in which the user is not a media professional, in which professional quality is not essential, or in which it is necessary to edit material very quickly. Even NLEs which claim to be aimed at non-professionals have a significant learning curve and require substantial time to produce acceptable productions. It is generally accepted that in typical cases a user will have to spend one hour in order to create one minute of output video, in other words a ratio of 60:1 of production time to playback duration. It is one of the goals of the current invention to reduce this ratio dramatically through automation, to the point where in some cases acceptable results can be produced without any user intervention.

There also exist several tools which allow a user to create productions involving the real-time display of images and text synchronized to an audio track. These include animation tools (e.g. Flash from Macromedia Inc.), slideshow tools (e.g. PowerPoint

from Microsoft Inc.) and authoring tools for streaming media (e.g. RealPlayer from Real Networks Inc.). But once again, users often find that they need to spend hours in order to produce a simple production lasting a few minutes.

Summary of the Invention

This invention aims to provide new and useful apparatus and methods for generating media productions from input media material.

In general terms, the invention proposes that input material is edited to construct an output production. The process includes deriving a set of media descriptors which describe the input material, either by analysis or from an external source, or a combination of the two. This is followed by a computer-based construction process which includes making edit decisions based on (i) the set of media descriptors, and (ii) style data, such as user-generated style data, for defining the editing style.

The input material may include any one or more of motion video, still images, music, speech, sound effects, animated graphics and text.

The set of media descriptors may be supplemented by descriptors which are pre-generated (e.g. outside the apparatus of the invention) and imported, for example together with the input material.

The style data may be generated by a process which includes either or both of deterministic and stochastic (probabilistic) operations.

The editing may include any one or more of the following processes applied to the input material: segmentation (of video/audio), selective inclusion, sequencing, transformation and combination. These processes may optionally be supplemented with user intervention. This is supported at two stages: a pre-selection stage prior to the automatic construction process and a touch-up stage after construction.

A particularly preferred feature of the invention is the ability to produce music-based productions in which the input material consists of a) motion video material and/or a set of images, and b) recorded music. The system analyses both the video/images and the music to create media description data for both, then uses this information to create the output production, influenced or determined by the structure of the music.

Typical applications of the current invention include the production of video and other time-based media for home, corporate and hobbyist environments, production of slideshows synchronized to music, production of rich-media electronic greeting cards, production of media for world-wide-websites, production of rich-media online catalogues, production of rich-media online content related to consumer-to-consumer sales applications such as online auctions and classified advertisements, some professional video applications such as the production of karaoke videos, etc.

The invention, which includes both method and apparatus aspects (i.e. apparatus comprising respective means to perform the steps of the methods), may be embodied within various kinds of hardware including general-purpose computers, personal digital assistants, dedicated video-editing boxes, set-top boxes, digital video recorders, televisions, games consoles, digital still cameras, and digital video cameras.

Brief description of the drawings

Embodiments of the invention are described hereinafter, for the sake of example only, with reference to the drawings, in which:

Fig. 1 illustrates an embodiment of the invention comprising a set of interconnected functional modules;

Fig. 2 illustrates an example of the operation of the embodiment of fig. 1.

Fig. 3 illustrates schematically an operating principle of the embodiment of fig. 1.

Fig. 4 shows the embodiment of fig. 1 searching input video material to derive a suitable segment.

Fig. 5 is a flowchart illustrating logic for selecting a video segment in the example of fig. 4, such that the output production preserves the order of segments found in the input material.

Fig. 6 is a flowchart illustrating logic for selecting a video segment in the example of fig. 4, such that the output production does not preserve the order of segments found in the input material, but instead selects by similarity of the segment descriptor values.

Fig. 7 is a flowchart illustrating logic used by the embodiment of fig. 1 to calculate a similarity measure between a set of candidate segments and a target set of descriptor values.

Fig. 8 illustrates the structure of a media scene graph which is generated in the embodiment of fig. 1 and which is a complete representation of the form of the output production or a complete set of instructions for making the output production.

Fig. 9 illustrates a first, simple GUI suitable for use in the embodiment of fig. 1with three main user controls.

Fig. 10 illustrates a first, simple GUI suitable for use in the embodiment of fig. 1 with five main user controls.

Fig. 11 illustrates an embodiment of the invention which does not require user interaction in normal use.

Fig.12 illustrates an elaboration of the principle illustrated in fig. 3, specific to the creation of music-based productions.

Fig. 13 shows a feature of the embodiment of fig. 1 in which the sub-style sequence is matched in a one-one correspondence with the macro-structure of a piece of music.

Fig. 14 is a flowchart illustrating one way in the embodiment of fig. 1 in which a sub-style sequence can be matched automatically to the macro-structure of input music.

Fig. 15 illustrates a thresholding mechanism for combining edit preferences from style information with edit hints derived from a music description in order to generate edit decisions.

Fig. 16 shows a typical workflow for a user creating a video production using a conventional non-linear video editor.

Fig. 17 shows the typical workflow for a user creating a video production using the embodiment of Fig. 1.

Detailed description of the embodiments

Fig. 1 shows the overall structure of an embodiment of the invention.

Referring to Fig. 1, the material input to the system includes one or more of the following:

- "input video" [101], i.e. motion video such as a digital video stream or one or more digital video files. Typically this is unedited "raw footage" such as video captured from a camera or camcorder. Optionally it may include an input soundtrack [102].

- "input images" [103], i.e. still images such as digital image files. These may be used instead of motion video, or in addition to motion video.

- "input music" [104] in a form such as a digital audio stream or one or more digital audio files. In the embodiment music provides the timing and

framework for the output production: the input visual material is edited in ways which relate to the structure of the music in order to produce a music-based production.

- input text and/or graphics [105] typically used for titles, credits, subtitles, etc.

- "style information" [106], i.e. data or logic used by the system to control or influence aspects of the automatic construction process - in other words the "editing style". The user may select from a number of predefined styles, and/or have access to individual style parameters. Depending on the embodiment, styles may be external to the system or form part of the system.

In this document the term "input material" is used to mean one or more pieces of media which are presented as input to the system. Supported media types include motion video, still images, music, speech, sound effects, static or animated graphics and static or animated text. The term "input visual material" refers to input material of any visual type including video, images, animation, graphics or text.

Output

Referring to Fig. 1, the output production [108] created by the system is a piece of time-based media such as a video, animation, or timed sequence of images; this may include an associated soundtrack, the output soundtrack [109], consisting of music, speech and/or other sounds. The output production is formed from some or all of the input material which has been subjected to one or more of the following processes by the system:

- "Segmentation". That is, input video is segmented according to visual or sonic characteristics, for example into shots, parts of shots, segments that contain a particular voice or background sound, etc. A shot is a single contiguous piece of video which does not have breaks or cuts, such as a segment of video which was recorded without pausing or stopping a video camera.

- "Selective inclusion". That is, elements of the input material such as segments of video, music or soundtrack, selected images, or regions within images or video frames are included in the output production, while others are excluded. Typically - as in conventional media production - a large fraction is excluded.

- "Sequencing". Elements of the input material may be sequenced so that the time-ordering of the elements comprising the output production corresponds to the time ordering of those elements in the input material, or they may be sequenced according to some other criterion such as descriptor similarity.

- "Transformation". Elements of the input material may be transformed, e.g. by a process including any of the "special effects" well-known in the prior art, including transformations of color (e.g. monochrome and flash effects), speed (e.g. slow-motion), size (e.g. artificial zoom), position (e.g. artificial pan), shape (e.g. warping), etc.

- "Combination". Elements of the input material are combined both simultaneously and sequentially. For example, images and video segments from the input material may be presented simultaneously with input music, and input text/graphics may be overlaid onto the video. Images and segments of video may be concatenated with overlaps allowing the use of transitions such as dissolves and wipes well-known in the art. Segments of the input soundtrack may be mixed with segments of the input music. Multiple images and/or video segments can be presented simultaneously in different regions of the frame area of the output production or mixed over each other to create composite images ("mixage").

The output production may also include material generated by the system without reference to the input material, such as colors and textures used as backgrounds, static and animated graphical elements, etc.

Analysis and Description Components

Referring again to fig. 1, the embodiment has the following components concerned with analysis and description of the input material.

- The video analyzer [110]. This analyses the input video to produce a video description [111] containing one or more descriptors. The video analyzer applies signal analysis techniques or other kinds of processing to individual frames or multiple frames of the input video in order to create the descriptors. Typical descriptors are measures of brightness or color such as color histograms, measures of texture, measures of shape, measures of motion activity, descriptors identifying the times of shot and other segment boundaries in the input video, categorical likelihood measures (e.g. probability that a segment of the input video contains a human face, probability that it is a natural scene, etc), measures of the rate of change and statistical properties of other descriptors, descriptors derived by combining two or more other descriptors, etc. Many such descriptors and techniques are well known to those skilled in the art and new ones are constantly being defined.

- The soundtrack analyzer [112]. This analyses the input soundtrack of the input video to produce a soundtrack description [113] containing one or more descriptors. The soundtrack analyzer applies signal analysis techniques or other kinds of processing to the input soundtrack in order to create the descriptors. Typical descriptors are measures of audio intensity or loudness, measures of frequency content such as spectral centroid, brightness and sharpness, categorical likelihood measures (e.g. probability that a segment of the input soundtrack contains a human voice), measures of the rate of change and statistical properties of other descriptors, descriptors derived by combining two or more other descriptors, etc. Many such descriptors and techniques are well known to those skilled in the art and new ones are constantly being defined.

- The image analyzer [114]. This analyses the input images to produce an images description [115] containing one or more descriptors. The image analyzer applies

signal analysis techniques or other kinds of processing to individual images or groups of images in order to create the descriptors. Typical descriptors are measures of brightness or color such as color histograms, measures of texture, measures of shape, categorical likelihood measures (e.g. probability that an image contains a human face, probability that it is a natural scene, etc), measures of the statistical properties of other descriptors, descriptors derived by combining two or more other descriptors, etc. Many such descriptors and techniques are well known to those skilled in the art and new ones are constantly being defined.

- The music analyzer [116]. This analyses the input music to produce a music description [117] containing one or more descriptors. The music analyzer applies signal analysis techniques or other kinds of processing to segments of the music in order to create the descriptors. Typical descriptors are measures of intensity or loudness, measures of beat strength, musical rhythm and tempo, measures of frequency content such as spectral centroid, brightness and sharpness, measures of musical pitch content such as root note pitch, consonance, musical key membership and chordal content, measures of the rate of change and statistical properties of other descriptors, descriptors derived by combining two or more other descriptors, etc. Many such descriptors and techniques are well known to those skilled in the art and new ones are constantly being defined. The music analyzer may also provide a representation of the structure of the input music at various timescales, from the "macro" timescale of major sections such as introduction, verse, chorus, etc to the "micro" timescale of bars, beats and sub-beats. Means of representing musical structure are well-known to musicians, music theorists, and others, and many techniques for extracting this type of information by signal analysis are known in the field of computer music analysis.

In this document, the analyzer components described above [110, 112, 114 and 116] are collectively known as the "media analyzers", and the descriptions [111, 113, 115 and 117] are known as "media descriptions".

Media descriptions can also be stored for later use, for example by saving the description data to disk or non-volatile memory. (For simplicity, this is not shown in Fig. 1.) This allows the user to construct different output productions from the input material without the need to re-analyze material, thus reducing the processing time needed to view multiple alternative productions.

In addition to, or alternatively to, signal analysis, descriptors may be imported into the system and stored in the media descriptions. (For simplicity, this is not shown in fig. 1.). Such descriptors have been created at some earlier time and are typically embedded within, or in some way linked to, the input material. Such descriptors include video descriptors generated by camera instrumentation such as time-of-shooting, focal distance, geographical location generated by satellite positioning systems (e.g. GPS) attached to the camera, measures of ambient light level during shooting, etc. They may also include music descriptors generated during the music production process, such as elements extracted or derived from music sequencers or MIDI (Musical Instrument Digital Interface) data. Music sequencers and MIDI are widely used in music production and can be used to create descriptive information which is difficult to derive from the music audio signal after it is mixed down: for example information about musical pitch, instrumentation, music repetition structures, etc.

Imported descriptors can also originate from a manual or semi-automatic process, for example in which a user annotates the input music, video or images before importing the input material plus its descriptions into the system. Such descriptors may be closely related to descriptors created by signal analysis. For example it is sometimes desirable to create descriptors using the system, correct or refine these descriptions manually, and then use the refined description as the basis for processing by the other modules of the system.

Imported descriptors may be stored directly in the media descriptions, or they may require further analysis, conversion or interpretation after they are imported; this function is also provided by the media analyzers.

Other Components

Referring again to fig. 1, the system further includes the following components:

- The graphical user interface or GUI [120]. This acts as intermediary between the user and the system, communicating with several of the other modules of the system. User interaction typically includes the following capabilities:

    - Overall control, such as selection of files containing the input material and selection of a destination file for the output production. Other aspects of control include the initiation of analysis and construction tasks.

    - User interaction with the style information – for example the selection of predefined styles, or creation of new styles, or alteration of existing styles.

    - Manual intervention, both at a pre-selection stage and at a touch-up stage.

  Features and variants of the GUI are further described below.

- The constructor [121]. This contains much of the core logic of the system. It receives as input the one or more media descriptions and receives (or contains within it) the style information [105]. Its main function is to use these inputs to make all the edit decisions necessary to specify the form of the output production [108] and to store this specification of the output production in a structure called the "media scene graph" or MSG [122]. The MSG can be regarded as a complete representation of the form of the output production or as a complete set of instructions for making the output production; this includes the source and timing of all elements of the input material (such as segments of video, music or soundtrack, selected images, or regions within images or video frames) which are used in the output production, the types of transformations and special effects applied to these elements, the types of transition effect used in the output production, the source and presentation of all overlays such as text and graphics

used in the Output production, the timing of all of these elements, etc. The MSG controls the renderer (see just below) and also plays an important role during manual touch-up: it is the primary underlying data structure which the user interacts with at this stage, being a full representation of the current production at all times and being updated to reflect changes made by the user.

The MSG can optionally be saved and reloaded for later use, allowing progressive touch-up of the final production. Also, parts of the MSG (for example temporal regions or certain types of edit information) can be "locked" and others "unlocked". This allows an output production to be made by progressive refinement: the user instructs the system to run the constructor (and renderer), views the resulting output production, locks regions or features that he/she likes, runs the constructor (and renderer) again to replace the unlocked regions/features, views the altered output production, locks another set of regions/features, and so on.

The logic of the constructor and the structure of the MSG are described in detail below.

- The renderer [123]. This produces an output production according to the information in the MSG. In other words, it interprets the MSG data as instructions and, according to these instructions, selects elements of the input material, applies processes such as sequencing, transformation, combination and concatenation to the selections, and transfers or copies them to an output such as a file or an audiovisual monitor. The result is the output production. The kind of operations performed by the renderer are generally well-known in the art and do not require further explanation, being found in many non-linear video editors and generally supported by standard video architectures such as DirectShow from Microsoft Inc. and QuickTime from Apple Inc. The renderer may include a compression module, compressing the output production using techniques such as digital video compression and digital audio compression which are well-know in the art, for example as defined by the MPEG (Motion Picture Experts Group) standards body.

Distributed Production

In general in this document, the invention is described as a single system including the media analyzers, the constructor and the renderer. However it can also be a distributed system in which each of these modules is a separate program, potentially run at different times at different locations by different parties. It has already been mentioned that media descriptions can be stored and imported when needed by the constructor. Such media descriptions can be created by media analyzer modules invoked at any earlier time at any location by any party.

Likewise, because the MSG is a complete representation of the form of the output production or a complete set of instructions for making the output production, the renderer can be run separately from the constructor or analyzers. It can even be run in real-time while the output production is viewed, in other words creating the output production on the fly, in which case the renderer is in effect a sophisticated playback engine. All that is required to make this possible is that the MSG and the input material are available at the time of rendering.

For example, in an application where two parties share access to a common body of input material, or have two identical copies of the input material, one party can run the analyzers and constructor in order to create an MSG, then send this MSG to the second party whereupon the second party runs the renderer to create the output production "on the fly" as she/he views it. In another example, a community of people can first acquire copies of a common body of input material and associated pre-created media descriptions, then individually produce output productions which they share with each other simply by transmitting different MSG's. The advantage of this is that each MSG is a small amount of data compared to typical media data and can therefore be transmitted quickly and easily. The common body of media is suited to distribution on a medium such as CD-ROM or DVD; the community of people owning the CD-ROM/DVD can share their productions by, for example, forwarding MSG's as email attachments.

The process of automatic construction will now be described in detail with reference to figs. 2 to 8.

Video Editing Example

Fig. 2 shows a typical example in which an output production is created from input material by the application of the construction processes listed above: segmentation, selective inclusion, sequencing, transformation and combination. (This figure is a purely visual example, not showing audio.) In traditional linear and non-linear editing these processes are well-known and applied manually. The main purpose of the current invention is to automate them fully or partially. Before describing how the invention achieves such automation, it is useful to consider some of the examples illustrated in fig. 2:

- Segmentation . Two pieces of input video [201, 202] such as digital video files are segmented to produce five "source" segments, sSeg1 to sSeg5 [211, 212, 213, 214, 215]. One of these, sSeg5 [215] is a segment consisting of a single frame.

- Selective inclusion. The five source segments [211 - 215] are included in the output video production while the remaining material from the input video is not used. A single image, sImage1 [216] is also included.

- Sequencing. In this example, the order of the segments comprising the output production is not the same as their order in the input material. For example, in the output production, the first two segments from input video B [211, 214] are interspersed by two segments from input video A [212, 213].

- Transformation. Several examples of transformation are shown in fig. 2. The segment sSeg2 is transformed to monochrome by removing its color information to preserve only its luminosity [220]. sSeg3 is transformed by adding flash effects, i.e. in which the luminosity of regions within one or more frames is increased [221]. sSeg4 is subjected to a time transformation, slowing it to 0.4x its

original speed by, for example, creating new interpolated frames between the original frames [222]. sSeg5 is subjected to a more extreme time transformation, in which its single frame is copied to several successive frames to create a freeze [223]. sImage1 is also copied to a number of successive frames so that it forms a segment of the output production [224]. Many other such video transformations are well-known in the art. In addition, text and graphic elements used as overlays may be transformed in various ways: for example animated so that they change position, size, shape, color, etc as time progresses, possibly in response to parameters of music as described below. (These are indicated on fig. 2 as "AniText" [225] and "AniGraphic" [226].) Text and graphic elements may also be faded in [235] and out [236].

- Combination. Fig. 2 also illustrates several ways of combining the input material. The transformed segments dSeg1 and dSeg2 are concatenated to form a cut or "butt-edit" [230]. Other segments are concatenated with partial overlaps, allowing the use of dissolves [231], wipes [234] and other transition effects well-known in the art. Text and graphic elements, both static [227] and animated [225, 226] are overlaid on the video to form the final production.

Fig. 2 also contains a simple example of material generated by the system without using the input material: a black background [228] on top of which text [227] is overlaid.

All the above involve timing references relative to the output production; these are shown as vertical dotted lines projected onto the timeline [240] of the output production. Segments of input video involve an additional set of timing references relative to their input video source file, for example the start time [241] and end time [242] of sSeg4.

In conventional NLEs, the user makes all decisions about which of these processes to apply and where to apply them. The current invention creates an output production automatically by making the decisions itself and invoking processes such as those

above accordingly. The constructor [121] is the heart of the system and decides which processes to apply and where to apply them, while the renderer [123] performs the actual processing.

The Construction Process

Fig. 3 shows a central construction principle of the invention. Construction logic [301] in the constructor takes style information [302] and media descriptions (descriptions of video and/or images [303] and optionally a music description [304]) as input, using information from both to make a set of edit decisions which are stored in the MSG [305] and which specify the output production. The style information may be considered a set of preferences, suggestions or requests to the construction logic. The way in which the construction logic acts upon these preferences depends on the values of data in the media descriptions, so that the specific set of edit decisions is dependent both upon the style information and upon the nature of the input material.

Some examples of this process will now be presented in more detail, starting from the nature of styles.

Styles

Styles may be defined by data or logic or some mix of the two. For example, the style information [302] of fig. 3 could be a set of manually-defined parameters which are imported by the construction logic, or they could be a set of parameters generated by programmed style logic such as a style class in an object-oriented programming implementation. This distinction is not very important here and the following discussion refers to both interchangeably.

Style information is created by a style designer, for example by a process of manually defining a set of values for parameters, and the aim of the style designer is to create styles which will cause the system to generate high-quality output productions. The information comprising styles may be categorized according to which part of the

construction process they affect, using a similar categorization to that used above. For example, the style information of one embodiment has the following:

- "Segmentation parameters". A number of these affect the way in which the input video or input soundtrack will be segmented. Many techniques for segmentation of video are well known in the art, such as segmentation into shots using color histogram techniques, segmentation based upon the sonic characteristics of the associated soundtrack, etc. The segmentation may be linear, specifying a set of segments of equal weight in a list from start to end of the input material, or it may be hierarchical, in which the input material is divided into segments which contain other segments in a hierarchy of segment durations. Each style specifies which techniques to use, and specifies parameters controlling the segmentation including threshold values (such as degree of change of color histogram which is to be interpreted as a shot transition), minimum and maximum segment lengths, minimum number of segments to be specified, etc. In addition to these parameters controlling the segmentation of the input video or input soundtrack there is a parameter controlling the preferred segment duration - i.e. the preferred duration of the segments which are to comprise the output production. This controls the "cutting speed", an important characteristic of the output production.

- "Selective inclusion parameters". These are a set of parameters which control the selection of elements of the input material (such as segments of video, music or soundtrack, selected images, or regions within images or video frames) to be used at different points in the output production. In particular, in this embodiment they are a set of target values for media descriptors including brightness (average luminosity of video or image) and preferred activity level (average total motion of video). In other embodiments, any of the kinds of descriptors mentioned above (under "Analysis and Description Components") can be used.

- "Sequencing rules". Each style specifies the way in which sequencing is to be handled. For example, a parameter can specify whether the elements of the input material comprising the output production are to be chosen sequentially (in same order as they occur in the input material), non-sequentially (without regard to their sequence in the input material) or partly-sequentially (for example, within a certain distance of a time location which moves sequentially through the material, thus preserving the original sequence at a macroscopic scale but allowing non-sequential selection at smaller scales).

- "Transformation parameters". These specify a set of transformations to be used in each style, and specify rules for which kinds of transformation are to be applied at different points in the output production. For example a set of parameters may specify a particular type of flash effect to be used in terms of its brightness, radius, duration, etc, and a set of rules may specify when this flash is to be applied, such as "in every fourth segment of the output production, but only if the time since the last flash effect exceeds 10s and the brightness of the current segment is below a given value". Transformation parameters also specify the ways in which text and graphic elements are to be presented and animated, including static and dynamic values for position, size, shape, color, etc.

- "Combination parameters". These specify the way in which elements of the input material (and material generated by the system) are to be combined: for example the types of transition (cut/dissolve/wipe) to use, how often and in what sequence to use each type, the duration of transitions, when and for how long to generate blank backgrounds, when to overlay text/graphics elements and what type of material they may be overlaid on top of (for example, to avoid overlaying white text on video material of brightness above a certain value), etc.

The precise choice of parameters and their values is both highly dependent on context and partially subjective. The range of possibilities is enormous and influenced by

factors such as the type and range of input material which must be handled successfully, the demographics and preferences of target users of the system, and other such factors.

Generating Variety in Edit Decisions

In order to create an interesting production, it is usually necessary to introduce some variation in the edit decisions through the course of a production. For example, in most cases it is desirable to vary the preferred segment duration introduced above. A production consisting of segments of identical length would quickly become tedious, so the duration of segments must be varied to create a satisfying "edit rhythm".

In one embodiment, this need to introduce variety is addressed in several ways which may be used singly or in combination:

- "Sections" and " sub-styles". The output production is structured as a series of sections, each of which is assigned a different sub-style. These sub-styles are used in a certain order, the sub-style sequence, in which sub-styles may optionally be repeated. Each sub-style contains values for some or all of the style parameters (and/or logic for generating style information). For example this scheme makes it possible to specify a style which defines three sections in the output production, in which the first section comprises long segments, of low brightness, with few special effects, concatenated with slow dissolve transitions, the second section comprises short segments, of high brightness, with many special effects, butt-edited together with sharp cuts, and the third section has the same characteristics as the first.

- "Gradual evolution". It is also possible to specify gradual changes for some subset of the style parameters. For example, instead of the two contrasting sections of the previous example, there can be a slow evolution from the characteristics of the first sub-style to the second sub-style. In this example it is also possible to have two clearly-defined sections with most parameters

changing abruptly at the sub-style transition, yet allow a small number of parameters to vary gradually during the course of the output production.

- "Stochastic generation". Limited random variations are introduced at the level of each segment of the output video, providing the constructor with some variation in parameter values for each segment. For example a sub-style may specify that preferred segment duration is to be assigned a random value between 1S and 2S using a normal distribution with standard deviation of 0.25S. In this case, each time the constructor requests a value from the sub-style, the supplied value will be different, but will always lie between the 1S and 2S limits.

- "Value cycles". These also operate at the level of each segment of the output video. Each parameter is assigned a series of values and these values are used in a repeating sequence. For example in a particular sub-style, preferred segment duration might have a sequence of 3 values: 4, 2, 2 (seconds). Wherever this sub-style used, the durations of the segments in the output production will cycle 4, 2, 2, 4, 2, 2, 4, 2, 2, etc. Cycle lengths for different parameters may be the same or different. For example, in table 1 below, segment target brightness alternates between dark and bright (cycle length of 2), segment duration and transition type have a cycle length of 3, every 4th segment is transformed to monochrome and every 8th segment includes a flash effect. The overall pattern will only repeat every 24th segment. This creates variety, yet introduces a cyclic quality into the edit rhythm of the output production. Many viewers will not notice this explicitly - it may be subliminal – but it creates a different effect to stochastic variation and will be perceived as improving the quality of the production in some cases. This is particularly true when the output production is a music-based production as described below.

	Cycle Length	1	2	3	4	5	6	7	8	9	10	11	12
Target Brightness	2	Dark	Bright	Dark	Bright	Dark	Bright	Dark	Bright	Dark	Bright	Dark	Bright
Duration (s)	3	4	2	2	4	2	2	4	2	2	4	2	2
Transition type	3	Cut	Cut	Diss.	Cut	Cut	Diss.	Cut	Cut	Diss.	Cut	Cut	Diss.
Color or Monochrome?	4	M	C	C	C	M	C	C	C	M	C	C	C
Flash effect?	8	No	No	No	No	No	No	No	Yes	No	No	No	No
Time		→		→		→		→			→		

Table 1

Selection of Elements of the Input Material to Construct the Output production

A central function of the constructor is to select and sequence the elements of the input material (such as segments of video, music or soundtrack, selected images, or regions within images or video frames) which will form the output production. This will now be described for cases where the input material is video and the elements of the input material in question are video segments. The process for other media such as a set of images is related and generally simpler.

As described above (see "sequencing rules"), styles specify whether the elements of the input material comprising the output production are to be chosen sequentially, non-sequentially or partly-sequentially from the input material. The process of selecting elements of the input material involves a number of complexities which will now be explained with reference to a sequential case and a non-sequential case. Variants of these cases, such as partly-sequential cases, can be achieved using a mix of the techniques described in the following.

Segment Selection: A Sequential Case

Fig. 4 shows a common sequential case in which there is a single contiguous piece of input video [401]. The input video has duration D_i, significantly longer than the output production [402], which is of duration D_o. The ratio of input to output durations is $R_{io} = D_i / D_o$. The input video has been divided into segments such as shots, labeled I_1 - I_8 in this figure.

The constructor builds the output production segment by segment. In this example, it has already built 6 segments, O_1 to O_6 and is about to build the next segment. To select a new segment the constructor follows the flowchart in fig. 5. The process will now be described with reference to both figs. 4 and 5.

The constructor first gets the start time in the output production [501] for the new segment [403], labeled t_o in Fig. 4. It then gets a set of parameters required for the new segment, for example from the style information, including the required segment duration d_o and data about effects and transitions [502]. The duration then has to be adjusted [503] to produce a target segment duration d_T [404] for the segment which is to be taken from the input video, allowing for two things:

- If there are overlapping transitions such as dissolves before and/or after the segment, the duration of these must be included in the target segment duration d_T.

- If the effects to be applied involve any speed change, the duration has to be scaled. For example, if the output segment is to be played at double speed, the target segment duration d_T has to be twice the duration of the output segment d_o.

The constructor then calculates a time t_i in the input video at which it will start looking for a suitable segment [504]. In sequential cases it is generally desirable that the ouput production be approximately linear with respect to the input video, and to achieve this the input video segment should ideally be taken from a time location calculated as follows:

$$t_i = R_{io} * t_o$$

In other words the relative position in the input and output videos should be the same.

The constructor checks whether there is a subsegment of the segment at t_i which is long enough to form the new segment – i.e. which is at least d_T in duration [505]. In addition to having duration >= d_T, the choice of subsegment is subject to two constraints:

- It should not cross a segment boundary in the input video. For example, if the input video has been segmented into shots, it is undesirable to cross a segment boundary because doing so will introduce an unintended cut into the output production. Also shot boundaries in raw video material are often not clean cuts; for example there may be a few bad frames as a camcorder re-synchronizes after being re-started, making it undesirable to use material that crosses a shot boundary. Referring to fig. 4, the question is whether the subsegment of video between t_i and the end of input segment I_5 [405] is at least d_T in duration.

- Since this is a strictly sequential case, output material is always presented in the same time order as it appears in the input video and may not repeat. Thus in order for a subsegment to be selected it must start from a location in the input video which is later than previously-selected material. The search logic may optionally search backwards from t_i but it must only go back as far as the end of the previously-used material. (This is not shown explicitly in fig. 5.)

If such a piece cannot be found within the input segment at t_i , the constructor searches forward [506]into later segments looking for a segment which is long enough (duration >= d_T). However there is no point searching too far forward: selecting a segment far ahead of the current location in the input video would not allow later segments to be sequential . A suitable location in the input video at which to stop searching is given by the formula $t_{i\text{-}stop} = R_{io} * (t_o + d_o)$.

If the constructor finds a segment or subsegment from the above, it then chooses a piece of duration d_T from within it [507] to use as the output segment. The choice of this piece can be simple: it may, for example, simply choose the first part of the subsegment. Or it be sophisticated, attempting to find a piece of length d_T which meets other criteria, for example by matching descriptor target values (using similar principles to those described below for a non-sequential case) or by selecting pieces which are estimated to be more interesting or superior in quality to the surrounding material (also see below); this is most useful when the segments of input video are significantly longer than the segments of the output video, a common situation.

If the constructor is unable to find a suitable segment from either of the above approaches, it relaxes the constraint that an output segment should not contain a segment boundary in the input video and builds an output segment of duration d_T from two or more segments/subsegments of the input video [508].

Segment Selection: A Non-Sequential Case

In this non-sequential case (fig. 6) some of the steps are the same as in the sequential case just described.

As in the above, the Constructor first gets the start time in the output production for the new segment [601] and then gets a set of parameters required for the new segment, for example from the Style Information, including the required segment duration d_o and data about effects and transitions [602]. In this non-sequential case, it also gets a set of target descriptor values from the style information [603]; it will select segments which match this set of values.

The duration then has to be adjusted to produce a target segment duration d_T for the segment which is to be taken from the input video [604], allowing for transitions and speed changes in the same way as described above for a sequential case.

The next step [605] is to find a set of candidate segments or subsegments of the input video. These are segments which are at least d_T in duration. They may also have to satisfy other criteria. For example, although some re-use of material may be permitted in certain non-sequential cases (unlike a strictly sequential case) it is generally desirable to limit the number of times the same material appears in the output production. This can be achieved by keeping a count of how often each part of the input video has been used, in which case a candidate (sub)segment is any contiguous part of a single segment of the input material that has been used less than the maximum permitted number of times and is at least d_T in duration.

If no such (sub)segment can be found, the constructor relaxes a constraint - for example, as in the sequential case above, it may build an output segment of duration d_T from two or more segments/subsegments of the input video. (Not shown in figure.)

The Constructor then gets descriptor values from the media description for these "candidate" (sub)segments [606] and evaluates a distance in descriptor space between each of these candidate points and the target descriptor values [607]. (This process is further described below, and shown in expanded form in Fig. 7). Finally the constructor selects the candidate segment corresponding to the candidate point that has the smallest distance from the target point [608] and uses it in the output production [609].

Selecting Segments by Proximity in Descriptor Space

As mentioned above, there is a need to select a best-match (sub)segment from a set of candidate (sub)segments. The best match is the (sub)segment which lies closest to the set of target values in the "descriptor space" (an n-dimensional space in which each of n descriptors is represented) – i.e. for which a distance measure between the given point (coordinates defined by the target values from the style information) and the candidate point (coordinates defined by the set of values in the media description) is smallest.

Although simple in principle, there are several issues to consider in this matching process. They will be described with reference to Fig. 7. This description concerns cases where the input material is video, but the principles apply to other media.

1. To ensure that the distance calculation gives results which correspond well to human expectations, it is important that all descriptors use a perceptual scale [701]. This is a scale in which a given difference in the descriptor value according to the scale is experienced by a user as a given difference in the perceived value, regardless of the position within the overall descriptor range. In most cases this can be approximated by the logarithm of some physical property.

2. In general, descriptors may be in different units with very different ranges. For example segment duration may be in seconds ranging from 0 to 30 while another descriptor uses a scale from 1 to 1000. To prevent this from affecting the distance calculation, we must normalize the units to a common scale such as 0 to 1. Such "unit normalization" [702] can be done using a straightforward linear transformation such as:

```
x' = (x - xmin) / (xmax - xmin)
```

where:

- x is a value in native (not unit-normalized) units,
- xmin is the minimum value in native units
- xmax is the maximum value in native units
- x' is the value in unit-normalized units

3. It is desirable that the system should produce satisfactory output for any kind of input video material provided by a user, without any constraints. Thus the system has no control over the spread of the descriptor values in the video description. For example, consider a case in which a set of descriptors extracted

by analysis have similar values for all but one of the segments of the input video material. In other words, all but one of the points representing the segments are clustered together in a small part of the descriptor space, and the remaining point is far away. In this case it is possible that the single isolated point is the closest point to *all* the target values provided by the style information. If a simple distance measure were used, it would lead to this segment being chosen every time, potentially resulting in an output production which consists of the same segment repeated a great many times – clearly not an acceptable result.

One approach to solving this problem is to exploit variation in the extracted descriptors in order to create variety in the output video, even when the variation is small. This can be achieved by "distribution normalization" [703]: i.e. linearly scaling and shifting the values of the descriptors for each point so that such clustering is eliminated or reduced. To normalize the distribution, we apply a formula such as the following to each descriptor in turn:

$$x' = ((x - m) * s' / s) + m'$$

where:

- x is a value before distribution normalization
- m is the mean of the input values
- s is the deviation* of the input values
- m' is the mean of the desired (output) distribution
- s' is the deviation* of the desired (output) distribution
- x' is the value in distribution-normalized units

* For example this can be the standard deviation or the average deviation (in their usual statistical definitions). The standard deviation is generally considered to be more accurate in most cases, while the average deviation can be calculated more quickly.

Distribution normalization can be applied in either of two ways:

a) Normalize both the set of descriptor values from the video description and the set of target values from the style information so that they conform to a common, standardized distribution – i.e. fixed values for m' and s'. (Another way to do this, identical in end-result, is to adjust one set of values so that it has the same distribution as the other.)

b) Normalize just one set of values – for example just the values from the video description – to a common, standardized distribution. In this case the distribution of each set will not necessarily be the same.

These two approaches may be used in different cases. Each has advantages and disadvantages and may be supported in different styles. The advantage of a) is that it tends to give unique matches because the distributions "lie on top of each other". Its disadvantage is that it discards any deliberate global bias of the mean of the values in a style; in fact it becomes impossible to bias the mean of a style towards either extreme. For example, if we create a style in which all target values of brightness are high, then option a) will discard that preference for bright segments, giving the same bright/dark mix as a non-biased set of values. Conversely the advantage of b) is that it can preserve such biases, and its disadvantage is that it may not give unique matches so successfully since the two distributions may not "lie on top of each other". (On the other hand, since the distribution of values from the Style Information is under control of the system designer, they can probably be made similar manually. But this may not be easy in all cases.)

4. After applying distribution normalization, outliers in the data may fall outside a given range. To facilitate the distance calculation it is usually necessary to discard such outliers or to bring them back within given limits [704].

5. Some descriptors may be more significant than others in determining perceived similarity. It is desirable to weight descriptors accordingly.

It is also desirable to allow certain descriptors to be ignored in some, but not all cases. For example a particular sub-style may specify target brightness and segment duration, but ignore another descriptor such as activity level. This sub-style may have to be used alongside other sub-styles which *do* specify activity level, and the distance values produced in each case must be comparable with each other. This can be achieved in the distance calculation by adding distance only for descriptors which are "significant", i.e. which are not to be ignored. This is equivalent to saying that, for a descriptor which is not significant, any value is a perfect match to the target value of that descriptor.

A distance calculation which takes into account weighting and allows descriptors to be ignored is as follows [705]:

$$D = SQRT(SUM_D((|v_{gd} - v_{cd}| \ ^ \ 2) * w_d))$$

where:

- D is the distance for a pair of points (one given, one candidate)
- SQRT is a square root operation
- SUM_D is the sum over the set of significant descriptors (i.e. excluding the ones to be ignored)
- v_{gd} is the value of the d'th descriptor for a given point
- v_{cd} is the value of the d'th descriptor for a candidate point
- ^ 2 is a squaring operation
- w_d is the weight for descriptor d

6. The candidate segments, or a subset consisting of the those which lie closest to the target point, are listed in the order of their proximity to the target point [706]. Note that in the example described above with reference to Fig. 6 it is only necessary to supply the single closest match. However, to support manual touch-up as described elsewhere in this document, it is desirable to have a list of alternative candidate segments ordered by proximity.

The above describes direct matching of descriptors in which the types of descriptor in the Style Information are identical to the types of descriptor in the media description: for example a brightness descriptor in the Style Information is matched to a brightness descriptor in the media description. It is also possible to use indirect matching, in which the set of descriptors used in the Style Information are mapped to a different set of descriptors in the media description via a mathematical or algorithmic relationship. For example the style information might have a "PeopleEnergy" descriptor defined as:

PeopleEnergy = 0.6 * Log(Activity) + 0.4 * PersonProbability

where "Activity" is a measure of the total average movement in a video segment and "PersonProbability" is a measure of the likelihood that the segment contains images of at least one person (for example using a skin-color detection algorithm well-known in the art). Such descriptors, defined by a mathematical or algorithmic manipulation applied to one or more other descriptors, may be termed "derived descriptors".

It is sometimes effective to define target values in the style information in terms of such derived descriptors, because this allows the use of "higher-level" descriptors which are closer to the kind of semantically-meaningful descriptors understood easily by human beings. In the above example, the style information would contain target values for PeopleEnergy while "Activity" and "PersonProbability" would be extracted by signal analysis of the input video.

If derived descriptors are used, the constructor logic can calculate values for the derived descriptors by applying mathematical or algorithmic manipulation to the lower level

descriptors of the media description and then perform descriptor matching on the values of the derived descriptors.

Optimizing the Selection of Material

As mentioned above, the amount of input material is often much greater than the length of the output production and it is therefore desirable in some cases to select parts of the input material which are estimated to be more interesting or superior in quality to the rest of the material This is related to segment selection as described above, and it may use some of the same techniques, but its purpose is somewhat different: segment selection as described above is primarily concerned with *where* parts of the input material should be placed in the output production, whereas optimizing the selection of material is primarily concerned with *which* parts of the input material should be used in the output production.

Present technology does not provide techniques to determine the semantic content of video or images reliably across all kinds of material using signal analysis techniques. It is therefore impossible for an automatic system to select material exactly as a human video editor would do it. Furthermore, this is a highly subjective subject: different human editors would select different material. Nevertheless it is possible to bias the selection of material in such a way that the majority of users will judge it to be more interesting or superior in quality to the average of the input material. To put it another way, the goal is automatically to select material which is generally "better", at least within certain types of material, than an unbiased set of samples taken from the input material.

Here are two examples of how this may be achieved:

1. Across many kinds of content, material containing images of people will generally be considered more interesting than material not containing images of people. Image processing techniques for detecting the presence of human beings are well known in the art - for example using skin color, face shape, or body shape. Using such techniques, it is possible to calculate a descriptor which represents the

probability that an image or a segment of video contains one or more human beings. Material with a high value of this descriptor can then be selected in preference to material with a low value of the descriptor.

2. In "handheld" video footage (i.e. video recorded by a camera held in the hands rather than attached to a fixed mount such as a tripod) there is tendency for users, especially non-professional users, to move the camera around until they see something of special interest in the viewfinder. In other words, for handheld material, segments of the resulting video with low camera movement tend to be more interesting than segments with high camera movement. Techniques for estimating camera movement are well known in the art - for example techniques based upon extracting motion vectors. Thus it is possible first to identify that input video material is handheld (this can be determined by analyzing the pattern of movement in time, or it can simply be information provided by the user in response to a prompt) and then, if it is handheld material, to select portions with low camera movement.

These techniques may be provided as options which a user of the system can invoke when desired. Alternatively they may be invoked when the user selects certain styles and not invoked for other styles: for example the low-motion handheld techniques just described may be desirable in a style intended to produce output productions which are soothing or "laid back", but not suitable in a style intended to create high-energy, fast-paced productions.

The Media Scene Graph (MSG)

As explained above, the MSG is a data structure produced by the constructor which completely represents the form of the output production. In this sense it is related to the concept of an edit decision list (EDL) well known in the prior art. However the MSG is also the primary underlying data structure which the user interacts with during touch-up, while a typical EDL is a linear structure which does not lend itself well to this kind of

manipulation. An MSG structure which is better-suited to this kind of manipulation will now be described with reference to Fig. 8.

The structure is essentially a tree in which the output production is the root [801]. Some of the branches of the tree are concerned with definitions; these specify the properties of certain entities which are used in the output production. They include a definition for every kind of transformation [802] used in the Output Production (e.g. specifying that a certain flash effect has a certain brightness, color, radius, duration, etc). They also include definitions for transitions [803] such as dissolves, definitions for text [804] including animated text, definitions for graphic elements [805] including animated graphics, definitions for blank backgrounds [806], etc.

The MSG also has one or more branches for timelines. Fig. 8 shows one main timeline [807] and one overlay timeline [808] with purposes similar to the two timelines of Fig. 2. The main timeline contains an entry for each of the segments forming the output production including segments derived from elements of the input material [810] and blanks constructed by the system [811]. Transformations [812] of those segments and the transitions [813] between them are also specified; these are in the form of references to the transformation and transition definitions mentioned above. The main timeline may also support an additional level of structure in the form of sections [814], each corresponding to the use of a single sub-style (see " Creating Variety in Edit Parameters " above); this facilitates user-selected touch-up operations which are to be applied to entire sections of the output production. Finally, the overlay timeline [808] specifies a sequence of overlays by referencing the textual [815] and graphical [816] definitions, optionally including animation information.

The entries for segments, overlays, etc which comprise the timelines contain time data relating both to the output production, and in some cases to the input material. For example, for video segments it is necessary to specify a location (such as a start-time) and a duration in the output production; it is also necessary to specify the source such as a start-time and duration in the input material.

Graphical User Interface (GUI)

Due to the high degree of automation in the production process, the invention can in some cases produce an output production of acceptable quality without human intervention. Thus, in certain embodiments of the invention, the GUI can be very simple, or indeed non-existent.

An example of a very simple, yet viable GUI is shown in Fig. 9. There are three main user controls, such as buttons, for performing the following functions:

1. A control allowing the user to select the input material [901]. For example, this can prompt the user to give the name of one or more video or image files containing the input material

2. A control allowing the user to select a style [902]. For example, when this is invoked, the user can be shown a list of available styles and prompted to select one.

3. A control which causes the output production to be created [903]. For example, this can prompt the user to give the name of a file which will store the output production. Once the user supplies this filename, the main processing modules of the system - the media analyzers, the constructor and the renderer - are invoked in order to create the output production.

There is also a standard control for closing the program [904].

A variant of this is shown in Fig. 10. This has five main user controls, such as buttons, for performing the following functions:

1. A control allowing the user to select the input visual material [1001]. For example, this can prompt the user to give the name of one or more video or

image files containing the input material. It can also prompt for the names of one or more files containing graphical overlays such as logos.

2. A control allowing the user to select input music [1002]. For example, this can prompt the user to give the name of one or more sound files containing recorded music.

3. A control allowing the user to add text [1003]. For example, this can prompt the user to enter textual information into a form. The text will be overlaid on the output production. Uses of overlaid text include titles, credits (for people and organizations involved in the production), subtitles, messages such as explanatory or advertising messages, etc.

4. A control allowing the user to select or define a style [1004]. To select a style the user can be shown a list of available styles and prompted to select one as described in the previous example. To define a style, the user can for example be shown a form containing the values of all the parameters of the style Information and prompted to enter or alter the values.

5. A control which causes the output production to be created [1005]. This can prompt the user to give the name of a file which will store the output production as described in the previous example. Once the user supplies this filename, the main processing modules of the system - the media analyzers, the constructor and the renderer - are invoked. In this example, the visual material is edited to music to in order to create a music-based production as described below, and the music replaces or is mixed with the input soundtrack. The text and graphical elements are then overlaid to produce the output production. The text and graphics may be animated to the music as described below.

There is also a standard control for closing the program [1006].

In either of the above examples the output production can be viewed from an external program such a media player. Alternatively, the GUI elements described above can be supplemented with a viewing window and "transport controls" well-known in the art, so that the user can view the output production from within the system.

In other embodiments, the GUI may include additional features for manual interaction. The motivation for these is that, although the primary purpose of the invention is to automate the editing process, it cannot always do this completely in every case. Depending on the nature of the input material and the application in question, an output production created fully automatically may not match the user's preferences in every detail. Thus it may be desirable to support features for manual interaction such as:

- Pre-selection of content. This provides an option for the user to select or de-select elements of the input material (such as segments of video, music or soundtrack, selected images, or regions within images or video frames) prior to automatic construction. The user identifies elements of the input material and specifies whether, where, or in what sequence they are to be used during the construction process. For example, the user may specify that a particular segment A *must* be included in the output production and must be the final shot, that another segment B is *optional* with a certain probability of inclusion depending on other factors that arise during construction, that a third segment C should be included only if B is included and must occur later than B in the output production, and that a fourth segment D *must not* be included. This process of pre-selection may be assisted by the media descriptions: for example, segmentation information in a video description can be used to present input video to the user as a series of shots; this is generally more convenient for the user than a single contiguous piece of video. Information from the media descriptions can also be used to categorize or cluster input material in ways which help the user: for example a set of input images or input video segments can be presented to the user in a set of "bins" each containing a set of images which are similar in some respect. The user can, if required, refine this categorization manually by adding or removing items from the bins; she/he can

then apply instructions such as those above ("include", "do not include", etc) to entire bins of images.

- Pre-selection of treatment. This provides an option for the user to select or specify, prior to automatic construction, aspects of the processing which will be applied to elements of the input material. For example the user might specify that all transition effects of the output production taking place during a certain section of the input music must be of a certain type, such as dissolves. Or she/he may manually select a subset of the input images and specify that those images are to be rendered in monochrome in the output production. Once again, automatic processes such as segmentation and clustering based upon information from the media descriptions can be used to assist the user. For example the system can categorize segments of the input video input by brightness, present the user with the set of segments which fall below a certain brightness threshold, allow the user to add/remove segments from this set, and then let the user specify that the brightness of these segments is to be increased by a certain percentage in order to enhance their visual quality.

- Touch-up of the output production. This allows the user to edit the output production *after* automatic construction, for example by replacing video segments of the output production with alternative segments from the input material while preserving the duration and effects applied to the segment, or by changing some of the transition effects, by adding or removing special effects, by overlaying additional text or graphics, etc. Yet again, information from the media descriptions can be used to assist the user in these operations. For example, when the user wishes to replace a segment of video in the output production, the system can present her/him with a representation of a set of alternative segments from which to choose. These segments can be listed in order of their similarity with the original segment according to a similarity measure derived from the video description. In a variant of this example, the user can be presented with two options such as "Replace with Similar Segment"

/ "Replace with Contrasting Segment"; once the user has selected one of these options, the system will supply a suitable alternative segment.

A quite different example of how information in a media description can be used to assist the manual touch-up process concerns the case where the output production is a music-based production. When video is "edited to music" by experienced video editors, the usual practice is to match certain visual elements to certain timing characteristics of music such as beats. In this case, timing information derived from the music description can be used to influence touch-up operations which the user is performing manually on the visual material of the output production so that time-critical visual events such as cuts and flashes are automatically aligned with beats, sub-beats and other significant times in the music. For example, as the user alters a cut point between two segments of the output production using a standard GUI operation such as dragging, information from the music description can be used to cause the cut point to jump between times in the music at which the amplitude of the music signal is high or there is other indication that a strong beat is present. A related option is to use quantization, a technique well-known in the field of music sequencers, in which event boundaries are aligned to a timing grid which is itself aligned to the beat of the music.

The GUI for supporting these manual operations can be constructed using standard elements including lists, hierarchical representations (such as those used in file managers), visual thumbnails, audio waveform displays, timelines, clip windows with transport controls, etc. These elements are well known in the art, being common in tools such as Non-Linear Video Editors (NLE's), image editors, audio editors and other media-processing software.

The invention can also be embodied in a non-interactive system which simply presents output productions and does not include any GUI elements for normal use (although such a system does require a GUI for configuring and managing it). Logic for an example of such an embodiment is illustrated in Fig. 11. This is suited to creating

output productions from input material which is arriving continuously, for example video or images from a "web cam" (a camera connected to the Internet). Material is captured from the camera until a certain quantity or duration has been collected [1101]. At this point, a style, and optionally a piece of input music, are selected automatically [1102, 1103]. These can simply be random selections from a number of options, or the style and music can be matched to characteristics of the video description / images description by a process of descriptor matching as described elsewhere in this document. The system now has the information it needs to make an output production and it does so [1104]. Finally it sends the output production to an audiovisual display device such as a multimedia computer or a television set [1105]. During the creation and delivery of this output production, this system can continue capturing material ready for another production. One use for this embodiment of the invention would be to provide automatically-constructed audiovisual productions at regular intervals to people in a public space, where the input material is being captured from a live camera.

Music-Based Productions

The embodiment is particularly suited to creating output productions in which the processing and timing of visual elements is governed by the characteristics and timing of an underlying music track. This is sometimes called "cutting to music" and is common in music videos, animated productions, promotion and marketing videos, television commercials and many other forms. Such productions are referred to as "music-based productions" in this document.

The general principle of music-based productions is that the music acts as the time reference. The visual elements are manipulated to conform to the music, but the music itself is not altered. Visual elements to which this may apply include motion video, images, animation, graphics and text. In addition, some non-musical audio elements such as speech and sound effects may be manipulated or positioned in time in ways which are influenced by the music. In general terms, the music is "master" and the other elements are "slaved" to it.

Music-based productions are constructed using a number of techniques. These techniques, today achieved through the skill of professional editors, include the following:

- The editing "pace" of the visual material is usually governed or influenced by some general characteristics of the music such as its tempo (i.e. beat speed), loudness, and overall level of perceived "energy". For example, when the music is faster or louder, the output production will be constructed from shots of shorter average duration and the transitions between shots will be faster, using more abrupt cuts and fewer slow dissolves. The musical characteristics controlling this not only vary from one piece of music to another but also from section to section within a single piece of music: for example the "energy" level in many pop songs is higher in the choruses than in the verses. A professional video editor will sense this and use a faster editing pace in the choruses than in the verses.

- The selection of visual material may also be influenced by the general characteristics of the music. For example, video with brighter colors or faster motion may be selected to accompany music with greater energy, and darker or more static visual material selected to accompany music which is slower or quieter.

- The timing of cuts and other transitions in the video will generally be synchronized with the beat of the music or with the timing of significant features of the music. This is sometimes known as "cutting to the beat" and is used extensively when video material is edited over a musical foundation.

- To varying degrees, the timing of events within shots of motion video may also be synchronized with the beat of the music or with the timing of significant features of the music. This is particularly true of motion events involving an abrupt deceleration, such as collisions between objects. For example, if a professional editor is incorporating a shot in which a falling object hits a floor,

she/he is likely to align this moment with a strong beat or other prominent event in the music.

- Furthermore, the selection and timing of special effects applied to the video is often influenced by characteristics of the music. For example, flashes may be included in time with strong beats or other prominent musical events, or a brief freeze-frame effect may be applied at a static moment in the music. At a larger time-scale, some visual effects may be applied to entire sections of the music: for example in a music video accompanying a pop song, the visual material of the verses may be presented in monochrome, while the visual material of the choruses is presented in full color.

- Overlays such as text and graphics may be influenced by characteristics of the music. For example, the times at which these elements appear or disappear may be linked to strong beats or other prominent musical events. They may even be animated to the music so that their appearance and motion is dependent on the music: for example they may be animated to jump between different locations on each musical beat, or change size or color at certain times related to the musical structure.

In summary, when visual material is to be edited to match music, the professional editor has available a repertoire of techniques across a range of timescales, from the "micro-structure" of musical beats or even subdivisions of beats, all the way up to the "macro-structure" of the main sections comprising the piece of music. When this is done successfully, the effect on the viewer/listener is enhanced: music and video are more likely to be perceived as a unified production and the emotional or dramatic impact is enhanced.

The embodiment automates the creation of music-based productions in several ways which will now be described.

Automation for Music-Based Productions

The nature of the music analyzer [116] and music description [117] have been presented above and we have already introduced several ways in which the creation of music-based productions can be automated or facilitated. This aspect of the invention will now be further described.

One simple way to match editing style to music structure is to control the editing parameters defining the visual character of the output production directly from parameters of the music description. For example, the tempo of the music can be used to control the cutting speed (the inverse of the average segment duration), beat-strength used to control the ratio of cuts to dissolves, and loudness used to control the brightness of segments selected from the input video. In a straightforward mapping of this kind, a fast-cut output production will result if the user selects a piece of music with a fast tempo. Or, to take another example, if the user selects a piece of music with contrasting loud and quiet sections, the output production may have corresponding bright and dark sections.

This approach is effective in some cases, and the invention allows for it to be supported: for example, it can be implemented in certain styles, so that the user can select this mode of operation by selecting those styles. However, this approach has limitations because it relinquishes nearly all control to the music. For example, if the music is very uniform, the output production may be monotonous, because the mechanisms described above for introducing variety are not active. Conversely, if the music has many rapid contrasts, the output production may lack coherency. So this approach tends to lack robustness to different pieces of music: it may produce acceptable output productions for some pieces of music, but is not guaranteed to work for a wide range of musical pieces.

A more sophisticated alternative is to select styles and/or sub-styles according to the characteristics of the music, but then to allow the style information to control or influence the individual edit decisions. This produces results which are more predictable and coherent for any input music, because all edit decisions may be placed

within bounds allowed by the style information. It also allows the style information to create variety even when the music is very uniform, for example using the techniques of stochastic generation and value cycling described above.

This approach conforms more closely to the central construction principle of the invention described above with reference to Fig. 3. It will now be elaborated for the case of music-based productions, with reference to Fig. 12.

As in the previous case discussed with reference to Fig. 3, the construction logic [1201] receives information from the style information [1202], the video/images description [1203], and the music description [1204]. In response to these inputs it generates edit decisions which are stored in the media scene graph [1205]. This diagram shows how the music description may be composed of two parts, a macro-description [1206] and a micro-description [1207], each performing substantially different functions.

The music macro-description [1206] contains a description of the input music at the timescale of major sections of the music, such as introduction, verse, chorus, etc. The characteristics of these sections are represented by a set of music section descriptors which are used to produce a sub-style sequence [1208]. As mentioned above, the sub-style sequence defines the order in which the sub-styles are to be used to generate the output production. Once the sub-style sequence has been established, there exists, for any time in the output production, a corresponding sub-style. Thus, when edit information is required for a particular time in the output production, that information will be supplied by the correct sub-style.

The role of the music micro-description [1207] will now be described. Referring back to the case, described earlier, where there is no input music, the information passed from styles/sub-styles to the construction logic [1201] is effectively a set of edit ***commands***, and the construction logic attempts to obey these commands if at all possible. (It may not always be possible, as some decisions depend upon the video/images description – see the discussion above about video segment selection – but generally it is possible and where it is, the construction logic will obey the command.)

However, in the case of music-based productions the information which the sub-style passes to the construction logic is a set of ***preferences:*** these preferences are to be followed only after the local features of the music, derived from the music micro-description [1207], are considered. The micro-description contains a description of the input music at the timescale of bars, beats and sub-beat. This description can include, or be used to generate, a series of "edit hints". For example, one kind of edit hint, which can be derived directly from a music amplitude descriptor, indicates that it is desirable to produce a segment transition in the output production at a certain time such as on a strong beat of the music.

Once the sub-style sequence has been created, the construction logic [1201] is able to build the MSG as follows, starting from the beginning of the output production and traversing to the end of the output production:

- Acquire edit preferences relevant to the current time in the output production from the sub-style corresponding to this time.

- Acquire edit hints relevant to the current time in the input music (which is directly related to the current time in the output production) from the music micro-description [1207].

- Where required – when making a decision relating to segment selection - acquire descriptor values from the video/images description [1203].

- Make edit decisions by combining these inputs and store the edit decisions in the MSG [1205].

The two major aspects of the above will now be described in greater detail by example: first, how a sub-style sequence matched to music macro-structure can be created, and second, a way in which the constructor can combine edit preferences with edit hints to produce edit decisions.

Creating a Sub-Style Sequence Matched to Music Macro-Structure

The general principle used to create a sub-style sequence matched to the music macro-structure is to use descriptor matching, a similar technique to that described in detail above for selecting input video segments by descriptor matching.

The goal of this process is to produce a sub-style sequence linked to the music structure such as the example shown in Fig. 13. This shows a sequence of music sections [1301] following a structure found in many popular songs: Introduction, Verse 1, Chorus, etc. These are matched in a one-to-one relationship with a set of sub-styles [1302]. The sequence of these sub-styles - SS3, SS2, SS4, etc in this example - is the sub-style sequence.

Before proceeding it is worth noting two features of this example. First, each time the same music or similar music occurs, it is linked to the same sub-style: for example the chorus is always linked to SS4 in this case. This is normally desirable whenever the music sections are very similar, and the procedure about to be described will cause this result in many such cases. Secondly, there is no requirement for all the sub-styles of a particular style to be used: there is no "SS1" in this figure, implying that sub-style 1 has not been selected for this particular piece of music.

Fig. 14 shows one way in which such a sub-style sequence may be derived automatically from the structure of the music. First, a set of descriptor values, one set for each music section, is acquired from the music description [1401]. Suitable descriptors for a music section include the duration of the music section, its average tempo, loudness, and beat-strength. Many other kinds of descriptors can be used, such as those listed earlier, and as mentioned, they may be generated by signal analysis, produced as a by-product of the music production, entered manually or generated by any other means. The only fixed requirement is that the set of descriptors for each music section characterizes some perceptually-significant qualities of the music section.

The next step, [1402] is to retrieve from the style information a set of target descriptor values, one set for each sub-style. The set of target values in a sub-style constitutes a description of the characteristics of music which this sub-style would be particularly well matched to. Typically these are created by the style designer by a manual process of entering a set of target values for each sub-style. For example, when the style designer creates a fast-cut sub-style (i.e. one which contains or generates small values for the preferred segment duration, introduced above), she/he might define that this sub-style is best suited to a music section which exhibits high values for the tempo and beat-strength descriptors, but is not dependent on loudness.

The next step, [1403] is to calculate a set of distances in descriptor space between music sections and sub-styles. This is similar to the process described above for selecting input video segments in a non-sequential case, and the techniques introduced for optimizing the calculation of proximity (see Fig. 7) may also be applied in this case. From the set of distances, a "trial" version of the sub-style sequence can now be created [1404] by assigning the closest sub-style to each music section.

The next step [1405] is to check the sub-style sequence for undesirable repeats. This is necessary because, even if techniques such as descriptor distribution normalization (described above in connection with Fig. 7) are applied, it may happen that the same sub-style gets mapped to too many of the music sections. This is particularly undesirable if the same sub-style gets mapped to two music sections which are consecutive yet different. Note that in the example presented above with reference to Fig. 13, the only consecutive occurrences of the same sub-style are the three occurrences of SS4 [1303] which occur because the Chorus repeats 3 times. This is a desirable case of repetition, but any other repeats in this example would probably be undesirable. Such undesirable repeats can often be detected, for example by checking whether the total number of occurrences of one sub-style exceeds a certain value or the total duration of consecutive repeats exceeds a certain time value.

If such undesirable repeats are found, they are eliminated [1406] by replacing some of the sub-styles in the sub-style sequence with alternatives such as the next-nearest sub-style for each music section found in step [1403] above.

Because this technique is similar to the techniques for selecting input video segments described in above with reference to Figs. 6 and 7, many of the details and alternatives presented above may also be applied here.

Combining Edit Preferences with Edit Hints to Produce Edit Decisions

Fig. 15 shows a graphical representation of one technique for combining edit preferences from the style/sub-style information with edit hints from the music micro-description in order to produce edit decisions. This technique operates at the timescale of musical beats. It will be described as a technique for making cut decisions (i.e. identifying time locations in the output production at which there should be a change of segment) but the technique, or variants of it, can be used to make other kinds of edit decision, such as identifying time locations at which to insert flashes or other special effects.

In this example, the horizontal axis is time, and the vertical arrows [1501] are edit hint pulses received or derived from the music micro-description. The height of these arrows is related to a perceptually-significant characteristic of the music and their horizontal location indicates the time at which they occur relative to a start-time $t = 0$. Typically the characteristic in question is one which is closely related to the musical beat, such as a signal derived from the amplitude variations in the audio signal. Many techniques are known in the art for automatically extracting such representations of musical beat: for example, the overall amplitude, or the amplitude of a frequency band within the signal, can be subjected to a threshold-crossing test. Further refinements, such as the use of a phase-locked loop, can synchronize the detection mechanism with the periodicities in amplitude variation which occur when the beat is regular, as it is in most popular music. Whatever technique is used, it is desirable that the edit hint pulses have the following tendencies:

- That the majority fall on beats, or on simple fractions of beats such as 1/2, 1/4, 1/3, etc.

- That pulses occurring on strong beats, such as the first beat of each bar, have higher values.

- That the value of off-beat pulses (those occurring between the main beats) have high values wherever there is a strong off-beat musical event; this is common in much music for example in styles of music known as "syncopated".

- That, in general the pulses correspond to the rhythm as it would be perceived by a human listener.

In this case the construction logic will interpret each edit hint pulse as a request to perform a cut at the corresponding time, and the height of each pulse as the strength of the request. The pulse height can be limited to a range such as 0 to 1; this is the case in Fig. 15.

However, the construction logic also has to take account of the style/sub-style information. One parameter specified by styles is the "cutting speed" as introduced earlier. What is relevant to this example is that the style information specifies, for any moment in the output production, a preferred segment duration for the next shot of the output production. This preferred duration is marked $t_{preferred}$ in Fig. 15 and is more generally represented by the four line segments [1502, 1503, 1504 and 1505]. These four segments form a threshold which will be applied to the edit hint pulses. The threshold reaches a minimum at $t_{preferred}$. It also takes the maximum permissible pulse value of 1 for $t < t_{min}$ and for $t > t_{max}$; This means that only pulses lying between t_{min} and t_{max} can cross the threshold.

Two more facts are required to fully understand the operation of this mechanism:

- The zero time, t = 0, corresponds to the previous cut: i.e. it is the start-time of the current video segment. As the construction logic creates the output production segment by segment, this is reset for every segment.

- The selected segment duration is the time, relative to t = 0, of the pulse for which the value $v_x = v_p - v_{th}$ is greatest, where v_p is the value of the pulse and v_{th} is the value of the threshold at the time of the pulse. In other words, it is the time of the pulse which exceeds the threshold by the greatest value, or if no pulse crosses the threshold, the pulse which comes closest to it. In Fig. 15 this is pulse [1506]. Note that pulse [1507] has a higher value, but is not used because the value v_x is greater for pulse [1506].

Taking into account all the above factors, it can be seen that this thresholding mechanism exhibits the following behaviors:

- It will favor durations which correspond to strong edit hint pulses, in other words it will tend to cause cuts related to the beats and other features of the music as described above.

- It will favor pulses which fall near to the preferred segment duration. In particular, if the music is very quiet so that the edit hint pulses are very weak, or the music is relatively featureless so that all the edit hint pulses are of similar strength, it will select a duration very close to $t_{preferred}$.

- It will always select durations which lie between t_{min} and t_{max}.

- By varying the distance between t_{min} and t_{max} it is possible to control the relative influence of the musical rhythm (the edit hint pulses) and the preferred segment duration. If t_{min} and t_{max} are close together, the preferred segment duration will dominate; if they are far apart, the musical rhythm will dominate. This is a factor which can be set differently in different styles, or even in different sub-styles of a single style. Changing the position of t_{min} and t_{max} relative to $t_{preferred}$

allows further control, biasing towards longer or shorter durations when there is no strong pulse close to $t_{preferred}$. Furthermore, variants of the mechanism can use non-linear thresholds, in which the line-segments are replaced by curves, providing even finer control over the behavior.

It is often effective to set the value of $t_{preferred}$ to a duration which is related to the beat speed at the current music tempo, for example, 1 beat, 1/2 beat, 2 beats, etc. Note also that the constructor often assigns $t_{preferred}$ a different value for each segment as it progresses through the output production, using techniques such as those described earlier for creating variety in edit decisions: the use of sub-styles, gradual evolution, stochastic generation and value cycling.

By combining the set of techniques described in this section, the invention is able to generate edit decisions which are perceived as relating to the rhythm of the music, which are sufficiently varied even if the music is very regular, and which always lie within acceptable limits, regardless of the selected music.

Other Features for Automating the Creation of Music-Based Productions

The invention may optionally be enhanced with several other features for automating or facilitating the creation of music-based productions, for example:

- In music-based productions, it is sometimes desirable to mix in the input soundtrack or parts of it. One option is to mix the entire input soundtrack with the input music at relative levels which remain constant. Another option is to vary the level of the input soundtrack or the input music or both, so that one or other is always clearly audible and not obscured by the other; for example this can use a technique known as "ducking" which is well-known to audio professionals and widely used in applications such as live radio to lower the level of music whenever an announcer speaks. Yet another option is to control the presence or absence of additional audio elements according to the value of descriptors in the music description. For example, in a common case where the input music is a song and the input soundtrack contains spoken voices, it will

generally create a confusing or muddled effect if the spoken voices are mixed simultaneously with the singing voice, so it is desirable to mix in audio from the input soundtrack only when there is no singing voice, such as in purely-instrumental sections of the music. In cases where the music description includes imported elements (as described above), this can be achieved by the use of manually-created descriptors which indicate the presence or absence of a singing voice. There are also known signal analysis techniques for detecting the presence of a singing voice in music which could incorporated into the music analyzer in order to automate this. A further possibility for controlling the mixing-in of audio from the input soundtrack, which can be used in conjunction with the techniques just described, is to select portions of the soundtrack according to their audio characteristics. For example, speech detection algorithms, which are well-known in the art, can be used to select only portions of the soundtrack in which speaking predominates over other sounds. Conversely, a music-detection algorithm can be used to ensure that sections of the soundtrack which contain music are not selected; this is desirable because music in the soundtrack would generally create an unpleasant effect if mixed with the input music. Although the audio analysis techniques for automating these processes are not completely reliable - for example, no known technique can detect the presence of a singing voice with complete accuracy across all types of music - they nevertheless work well enough to be useful in this invention, especially in embodiments where user touch-up (as described above) is supported.

- It has already been described how, in music-based productions, a professional editor will often align video elements so that the timing of significant features, such as the moment a falling object hits the ground is synchronized with the timing of notable features of the music. This can be automated by combining known techniques for video motion analysis with techniques for detecting musical features, such as the beat detection technique introduced above. For example, motion vectors can be extracted from video using standard techniques such as block-matching, and the timing of abrupt decelerations such as collisions

can then be established by identifying times when there is an abrupt change in the scalar or vector sum of the motion vectors within a region of the frame. Once the times of one or more of these deceleration moments has been established in a shot of the input video, and the strength of each deceleration established, the shot can be optimally aligned with the music by finding the relative timing between video and music for which there is the best match. This can be defined as the relative time for which the mathematical correlation of deceleration with beat strength, calculated over the duration of a segment of the output production, is at a maximum.

Beat strength and other descriptors derived by the music analyzer can be used to control the animation of text/graphic overlays. For example, parameters of an overlay such as its location, orientation, size, skewing, color, etc can be determined directly by the amplitude of the music signal. Or, in a more sophisticated implementation, representations of musical beat based on a threshold-crossing test (as introduced above) can be used to trigger sudden changes in parameters of an overlay, and the overlay then allowed to relax to its default position rather more slowly. In other words the animation can be based upon a relaxation model which is excited by pulses derived from the music signal and related to the musical beat. Furthermore, the music section descriptors introduced above can be used to control changes in the animation behavior which is aligned with the section boundaries and is related to the musical characteristics of each section; for example the color, size and relaxation speed of an overlaid text/graphic animated as above could be made proportional to the average loudness of the current music section, so that overlays occurring during loud music will be large, bright and move in a jerky manner, while overlays occurring during quiet music will be small, dark and move more flowingly.

Changes to the Production Workflow

This final section describes how a typical embodiment of the invention changes the workflow for a user engaged in creating a media production, with reference to Figs. 16 and 17. In these two figures, steps shown with dashed borders are ones which are typically automated or assisted by automation.

Fig. 16 shows the workflow in a typical conventional case, using a tool such as a Non-Linear Video Editor (NLE) to create a music-based output production from input video. First, the input video is captured and/or imported [1601]. This typically involves recording video using a camera attached to a computer, or transferring video material recorded earlier from a video camcorder to a computer, or acquiring video in the form of a digital video file. If an analogue recording device, such as an analogue camcorder, is used this step also involves digitizing the input signal. In any of these alternative scenarios, when this step is complete, the input video material has been introduced into the NLE.

As this example concerns a music-based production, the user also has to capture/import music [1602], for example by recording it, or transferring it from a musical medium such as an audio CD, or acquiring music as a digital audio file. In any of these alternative scenarios, when this step is complete, the input music has been introduced into the NLE.

Some NLE's are able to perform the next step [1603] automatically, segmenting the input video into shots using techniques such as detecting sudden changes in color histogram. The shots are presented to the user, typically as a set of "clips", i.e. small segments of input video. If the NLE does not include automatic shot segmentation, the user segments the input video manually.

Next the user needs to familiarize herself/himself with the shots of the input video. This is typically done by "logging" the shots [1604] - i.e. organizing them in groups or in certain orders, making notes about each shot, rejecting some shots, etc. For professional productions involving a lot of input material this is usually a lengthy task. For small casual productions it may largely be bypassed, although doing so is usually detrimental to the quality of the resulting production.

The next three steps [1605, 1606, 1607] may be performed sequentially, or the user may alternate between them (for example finishing one section of the output production

before moving on to the next section) or the user may work in a way which blurs the distinction between them. Whichever approach he/she adopts, the user must build the output production manually segment by segment, and - if a stylish music-based production is the goal - must carefully manipulate the segments so that they conform to the rhythm, timing and "feel" of the input music. This involves many of the techniques described above and is very time-taking in most cases, often requiring an hour, or several hours, to create each minute of the output production. It is also beyond the skill of many non-professional users to create a output production to a quality standard that they are happy with, particularly in the case of music-based productions, which require an understanding of music as well as visual material.

When the user believes that he/she has arrived at a satisfactory set of edit decisions, he/she instructs the NLE to render [1608], at which point it produces an output production as a video file or other output. The user views this and, if not satisfied [1609], returns to one of the earlier steps to alter or refine the production.

Finally the user exports their output production in a form which allows them, and others to view it [1610]. In the most basic case they may simply use the video file on their computer for local viewing, but more commonly they will transfer it to tape using a video cassette recorder, or to an optical disk format such as writeable compact disc (CD-R). It is also becoming increasingly common to distribute the video file using the Internet, for example by sending it as an email attachment, uploading it to a server which others can access, or sharing it from the user's local machine using so-called "peer-to-peer" file sharing.

Fig. 17 shows the workflow in a typical music-based production case using a system based upon an embodiment of the current invention, and should be contrasted with the conventional workflow just described with reference to Fig. 16.

The capture/import steps [1701 and 1702] are the same as the corresponding steps [1601 and 1602] described above for the conventional NLE case. The shot segmentation step [1703] is also essentially the same as the correspond step in the above [1603]. The

system uses one or more known techniques to automate the segmentation, and may optionally allow the user to override or adjust the resulting segmentation.

Next the user pre-selects content (elements of the input material) and/or treatment of the material [1704]. The invention provides techniques for assisting this process as described earlier. This step is optional and may be bypassed in some embodiments.

The next step [1705] comprises the many kinds of automatic analysis and construction which have been extensively described in this document. Once this step is finished, a complete set of edit decisions has been generated – these fully define an output production. Typically this step is performed fully automatically by the system and requires no help from the user.

The system now renders the output production [1706]. The user views this and, if not satisfied [1709], may either touch up the production with assistance from the system based upon the techniques described earlier [1707], or may return to any of the earlier steps.

Finally the user exports their output production [1710]. This step is similar to the corresponding step [1610] described above for the conventional NLE case.

It can be seen from Figs. 16 and 17 plus the above description that the workflow for a typical embodiment of the current invention involves more automation and less manual work by the user. This speeds up the production process, reduces the amount of the user's time involved in it, and provides greater support for inexperienced users.

Hardware Embodiments

It will be clear to those skilled in the art that the invention can be embodied in many kinds of hardware device, including general-purpose computers, personal digital assistants, dedicated video-editing boxes, set-top boxes, digital video recorders, televisions, computer games consoles, digital still cameras, digital video cameras and other devices capable of media processing. It can also be embodied as a system

comprising multiple devices, in which different parts of its functionality are embedded within more than one hardware device.

Although the invention has been described above with reference to particular embodiments, various modifications are possible within the scope of the invention as will be clear to a skilled persion.

Claims

1. A method for editing input data to form output data, said input data and output data both including at least one of visual and audio data, the method including the steps of:

analyzing said input data to generate one or more descriptors characterizing each of a plurality of portions of the input data;

defining style information for controlling the editing of the input data;

using (i) said input data, (ii) said descriptors, and (iii) said style information, to generate a set of edit decisions, the set of edit decisions specifying a set of editing operations to be performed on said input data; and

generating said output data by performing said set of operations upon said input material.

2. A method according to claim 1 including a step of supplementing said descriptors with additional pre-generated descriptors received from an external source, said additional descriptors being used in said step of generating said set of decisions.

3. A method according to claim 2 wherein said additional descriptors include descriptors generated by instrumentation at a time of recording the input data.

4. A method according to claim 2 or claim 3 wherein said additional descriptors include descriptors generated manually.

5. A method according to claim 2, claim 3 or claim 4 wherein said additional descriptors include music descriptors generated during music production.

6. A method for editing input data to form output data, said input data and output data both including at least one of visual and audio data, the method including the steps of:

receiving from an external source one or more pre-generated descriptors characterizing each of a plurality of portions of the input data;

defining style information for controlling the editing of the input data;

using (i) said input data, (ii) said descriptors, and (iii) said style information, to generate a set of edit decisions, the set of edit decisions specifying a set of editing operations to be performed on said input data; and

generating said output data by performing said set of operations upon said input material.

7. A method according to any preceding claim in which said output data comprises motion video data plus an associated soundtrack.

8. A method according to any preceding claim in which said output data comprises a sequence of images plus an associated soundtrack.

9. A method according to any preceding claim in which said set of operations include operations of at least one of the following types: segmentation, selective inclusion, sequencing, transformation or combination.

10. A method according to claim 9 in which said input data includes visual data, and said transformation operations include modification of the color of one or more parts of an image defined by said input data.

11. A method according to claim 9 or claim 10 in which said transformation operations include modification of the playback speed of one or more parts of said input material.

12. A method according to any of claims 9 to 11 in which said combination operations include video transitions.

13. A method according to any preceding claim in which the step of defining the style information is performed by selecting one of a plurality of predefined sets of style information based on said descriptors of the input data.

14. A method according to any preceding claim in which said style information includes a preferred segment duration parameter which influences the duration of segments of the input data incorporated into the output data.

15. A method according to any preceding claim in which said style information includes one or more target values for respective descriptors, and said step of generating the set of operations comprises selecting, for inclusion in the output data, one of more of the plurality of portions of said input data according to a calculation of the proximity of a) said target value or values and b) the descriptors for each said portion.

16. A method according to claim 15 in which said calculation includes a normalization of the descriptor values of each said portion of the input data.

17. A method according to claim 16 in which said calculation employs a weighting of the descriptors, whereby some descriptors are more significant in the calculation than others.

18. A method according to any preceding claim in which the order of portions of the output data is equal to, or at least correlated with, the order within the input data of corresponding portions of the input data.

19. A method according to any preceding claim in which said style information contains location data associated with locations in the output data, the location data being employed to generate the set of operations which produce the output data at the associated locations.

20. A method according to claim 19 in which said location data includes a plurality of data sections, each data section being associated with one or more sections of the output data and being used to generate the set of operations which produces the respective section or sections of the output data.

21. A method according to claim 20 in which said location data includes at least one parameter which varies as a function of location within the output data, whereby said edit decisions are influenced by the location within the output data of the section influenced by the decisions.

22. A method according to claim 21 in which the location data varies periodically with location in the output data.

23. A method according to any preceding claim in which said style information includes data generated from a probability distribution.

24. A method according to any preceding claim, further including receiving from a user a manual input identifying one or more elements of said input data and specifying, for each of said elements, one or more aspects of the way said element is to be edited into said output data.

25. A method according to any preceding claim, further including receiving from a user a manual input specifying that segments of said output data should be replaced, and modifying the set of operations to generate a set of modified operations for generating modified output data in which this replacement is effected.

26. A method according to claim 25 further comprising using said descriptors to suggest to the user segments of the input data resembling said segments of the output data to be replaced, whereby the user may decide to replace those segment of the output data with those segments of the input data.

27. A method according to any preceding claim further comprising receiving from a user an input indicating time-critical visual events to be aligned with particular times in the music of said output production, and performing said alignment using said descriptors.

28. A method according to any preceding claim, further including generating a data structure representing said set of operations, the data structure having substantially the structure of a tree.

29. A method according to claim 28 further comprising displaying the data structure to a user, and receiving inputs from the user indicating portions of the data structure to modify the corresponding set of operations.

30. A method according to claim 29 in which the user may indicate portions of the data structure which are provisionally prevented from being modified.

31. A method according to any preceding claim in which said descriptors include a human-probability descriptor for each of a plurality of elements of the input data, the human-probability descriptor representing a probability that a human being is present in each element of said input material, and said step of generating a set of operations generates operations for which the elements of the input data for which the value of said human-probability descriptor is high are more frequently incorporated into the output data than elements for which the human-probability descriptor is low.

32. A method according to any preceding claim in which said descriptors include at least one camera motion descriptor for each of a plurality of moving image elements of the input data which represent moving image data, the camera-motion descriptor

representing for each respective element a degree to which the camera which collected that element was moving when that element was collected, and said step of generating a set of operations generates operations for which the elements of the input data for which the value of said camera motion descriptor is low are more frequently incorporated into the output data than elements for which the camera motion descriptor is high.

33. A method according to any preceding claim further including a preliminary step of receiving, from a user, signals to determine said input data, to perform said step of defining style information, and to initiate said step of generating the set of decisions and said step of generating the output data.

34. A method according to any preceding claim in which said output data includes at least one overlay, said overlay comprising at least one of text and graphics.

35. A method according to claim 34 in which an overlay is animated.

36. A method according to claim 35 in which said input data includes music and at least one parameter of the animation of said overlay is determined by a music descriptor representing a characteristic of said music.

37. A method according any preceding claim in which at least two of said steps of defining said style information, generating said set of operations, and generating said output data, are initiated by different, spatially separated users.

38. A method according to any preceding claim in which said steps of defining said style information and generating said set of decisions are performed by a first user, and said sets of decisions are transmitted to a second user operating an apparatus with access to the input data, or a copy thereof, the second user initiating said step of generating said output data using said set, whereby the second user may inspect output data created by said first user without the need to transmit media data from said first user to said second user.

39. A method according to any preceding claim in which said descriptors include micro-descriptors associated with short sections of at least part of the input data, said micro-descriptors being used to derive editing hints which are used, in the step of generating the editing operations relating to the corresponding sections of the input data, in combination with, or to counteract, said style information.

40. A method according to claim 39 in which input data includes music data and said micro-descriptors are associated with sections of said music on a timescale of music bars or shorter.

41. A method according to any of claims 1 to 38 in which said input data includes music data, said descriptors including macro-descriptors describing a complete piece of music, said set of operations to be performed on said music data being generated using portions of said style information selected using said macro-descriptors, and micro-descriptors describing sections of the piece of music.

42. A method according to claim 40 or 41 in which one or more of said set of operations are determined by applying a time-dependent threshold governed by said style information to a time-variant set of values derived from said micro-descriptors.

43. A method according to any preceding claim in which said operations include operations to be performed on data in said input data relating to a first media type, and are derived depending on data in said input data relating to a second media type.

44. A method according to claim 43 in which the first media type is motion video and the second media type is music.

45. A method according to any preceding claim in which the input data includes a soundtrack associated with a motion video, and music, and said set of operations mixes portions of said soundtrack with said music so as to perform at least one of:

selecting said portions of soundtrack according to their audio characteristics,

determining when to mix in said portions of soundtrack according to the value of music descriptors, and

lowering the volume of said music when said portions of soundtrack are mixed in.

46. A computer program product, such as a recording medium, carrying program instructions which are readable by a computer apparatus and which cause the computer apparatus to perform a method according to any preceding claim.

47. An editing system for editing input data to form output data, said input data and output data both including at least one of visual and audio data, the system including:

analysis means for analyzing said input data to generate one or more descriptors characterizing each of a plurality of portions of the input data;

style definition means for defining style information for controlling the editing of the input data;

construction means for using (i) said input data, (ii) said descriptors, and (iii) said style information, to generate a set of one or more edit decisions specifying editing operations to be performed on said input data; and

rendering means for generating said output data by performing said set of operations on said input material.

48. An editing system for editing input data to form output data, said input data and output data both including at least one of visual and audio data, the system including:

means for receiving one or more descriptors characterizing each of a plurality of portions of the input data;

style definition means for defining style information for controlling the editing of the input data;

construction means for using (i) said input data, (ii) said descriptors, and (iii) said style information, to generate a set of one or more edit decisions specifying editing operations to be performed on said input data; and

rendering means for generating said output data by performing said set of operations on said input material.

Fig. 1

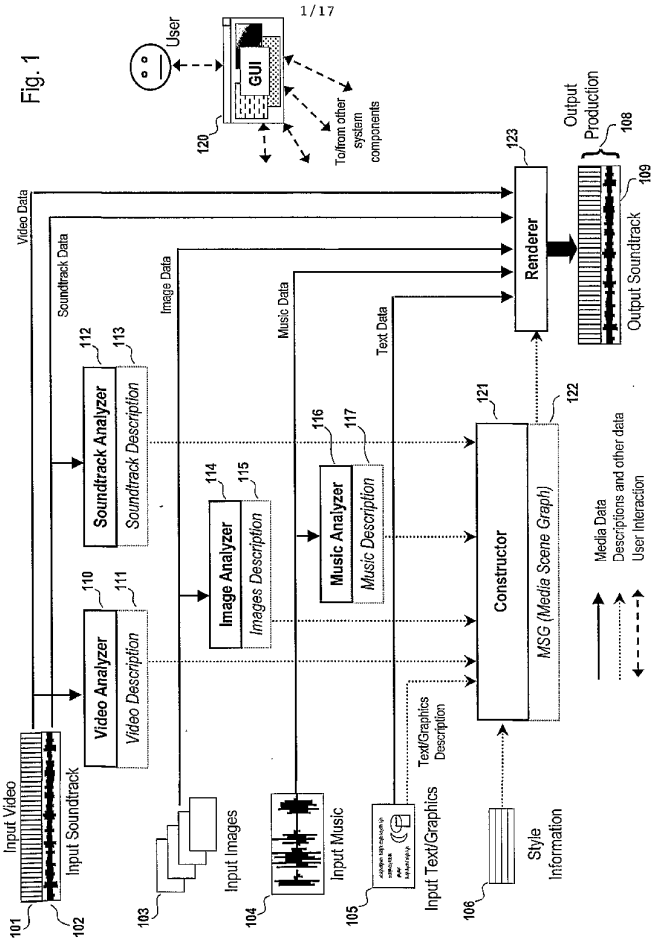
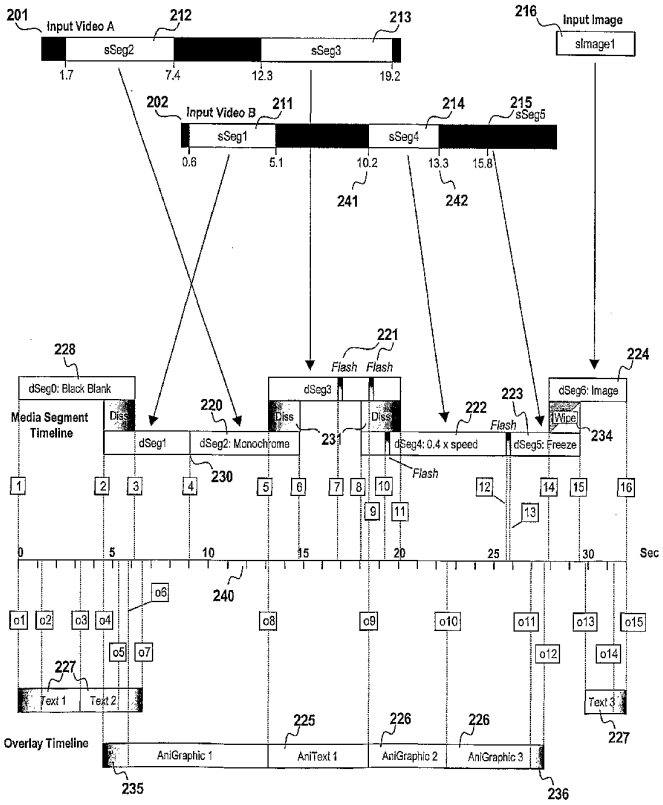

User
GUI
120
To/from other system components
Output Production
108
123
Renderer
109
Output Soundtrack
Video Data
Soundtrack Data
Image Data
Music Data
Text Data
112
113
Soundtrack Analyzer
Soundtrack Description
110
111
Video Analyzer
Video Description
114
115
Image Analyzer
Images Description
116
117
Music Analyzer
Music Description
121
122
Constructor
MSG (Media Scene Graph)
Media Data
Descriptions and other data
User Interaction
Input Video
Input Soundtrack
101
102
Input Images
103
Input Music
104
Input Text/Graphics
105
Text/Graphics Description
Style Information
106

Fig. 2



Key: sSeg = Source Segment. mSeg = Media Segment. dSeg = Destination Segment.
Diss = Dissolve. Text = StaticText (not animated).
AniGraphic = Animated graphic. AniText = Animated text.

Fig. 3

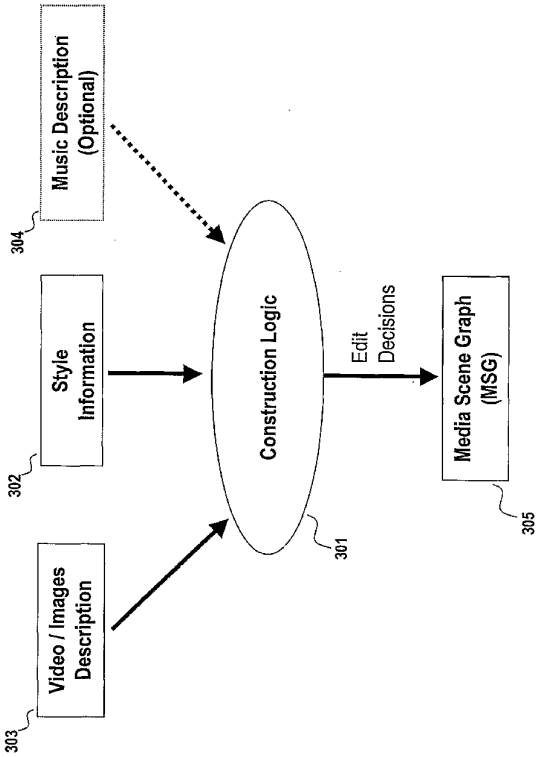

Music Description (Optional)
304
Style Information
302
Video / Images Description
303
Construction Logic
301
Edit Decisions
Media Scene Graph (MSG)
305

Fig. 4

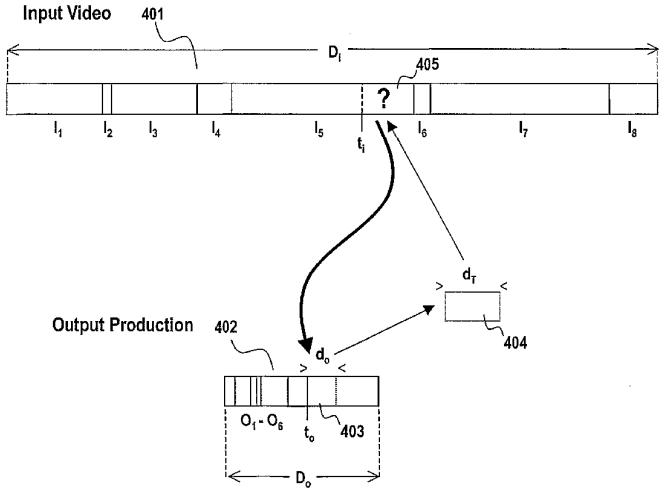

Input Video
401
D_I
405
?
I_1
I_2
I_3
I_4
I_5
t_i
I_6
I_7
I_8
d_T
404
Output Production
402
d_o
O_1 - O_6
t_o
403
D_o

Fig. 5

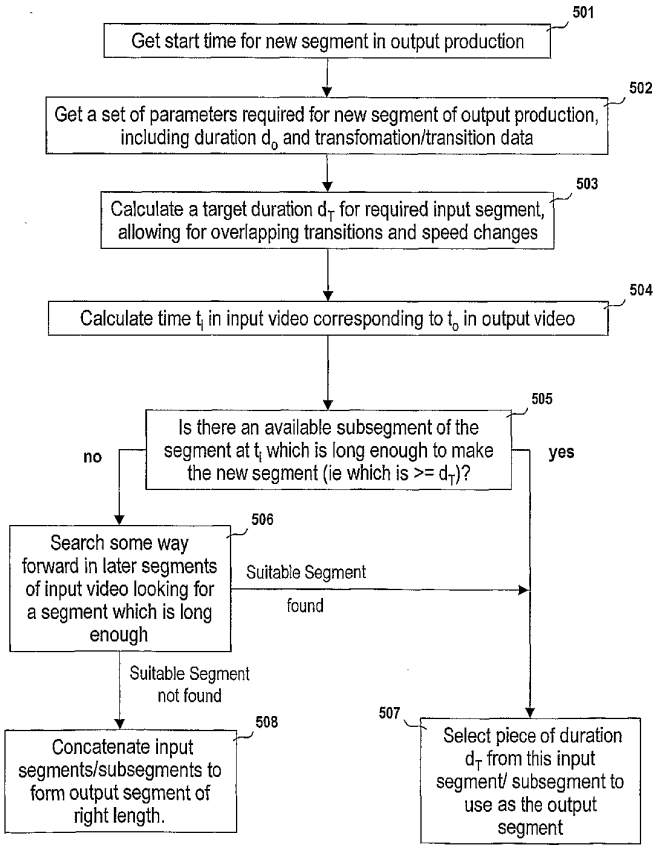

501
Get start time for new segment in output production
502
Get a set of parameters required for new segment of output production, including duration d_o and transfomation/transition data
503
Calculate a target duration d_T for required input segment, allowing for overlapping transitions and speed changes
504
Calculate time t_i in input video corresponding to t_o in output video
505
Is there an available subsegment of the segment at t_i which is long enough to make the new segment (ie which is >= d_T)?
no
yes
506
Search some way forward in later segments of input video looking for a segment which is long enough
Suitable Segment found
Suitable Segment not found
507
Select piece of duration d_T from this input segment/ subsegment to use as the output segment
508
Concatenate input segments/subsegments to form output segment of right length.

Fig. 6

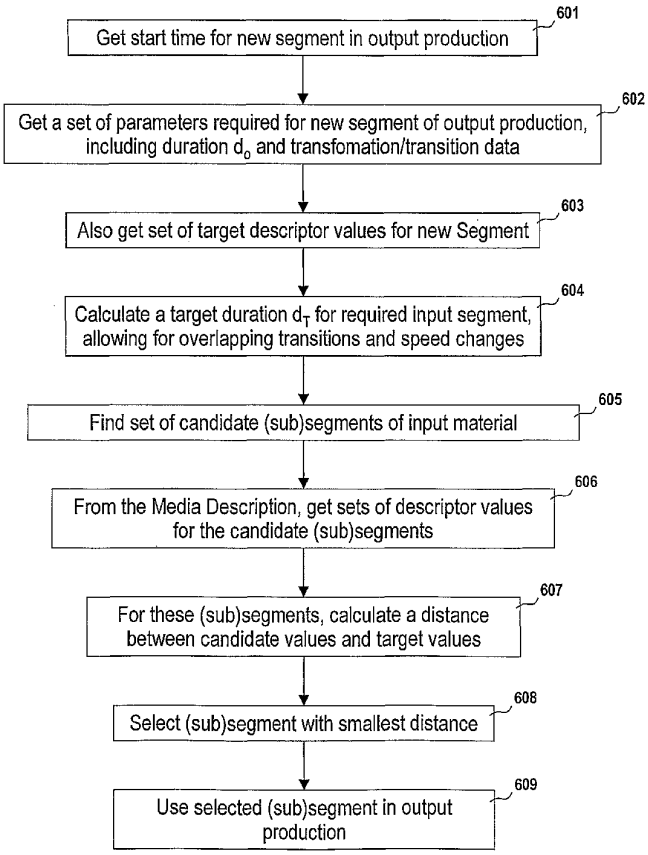

601
Get start time for new segment in output production
602
Get a set of parameters required for new segment of output production, including duration d_0 and transfomation/transition data
603
Also get set of target descriptor values for new Segment
604
Calculate a target duration d_T for required input segment, allowing for overlapping transitions and speed changes
605
Find set of candidate (sub)segments of input material
606
From the Media Description, get sets of descriptor values for the candidate (sub)segments
607
For these (sub)segments, calculate a distance between candidate values and target values
608
Select (sub)segment with smallest distance
609
Use selected (sub)segment in output production

Fig. 7

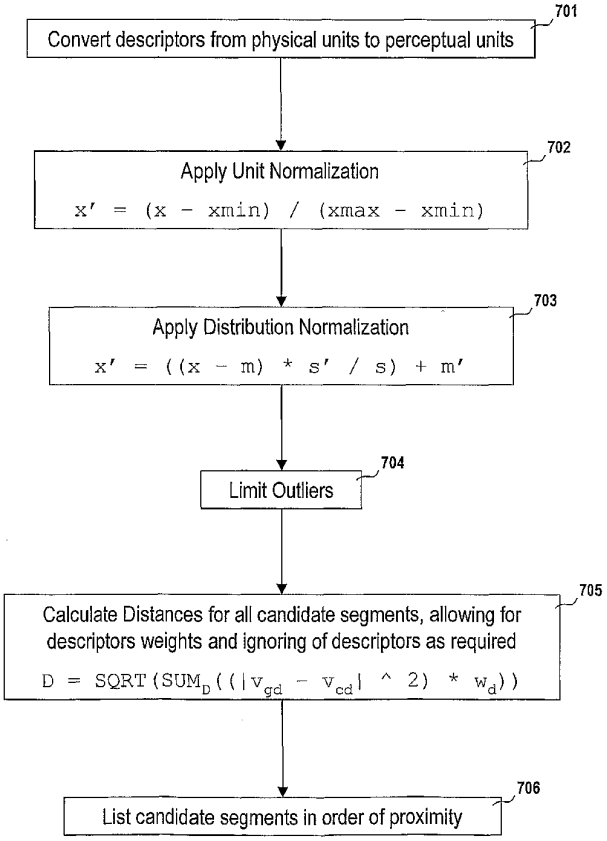
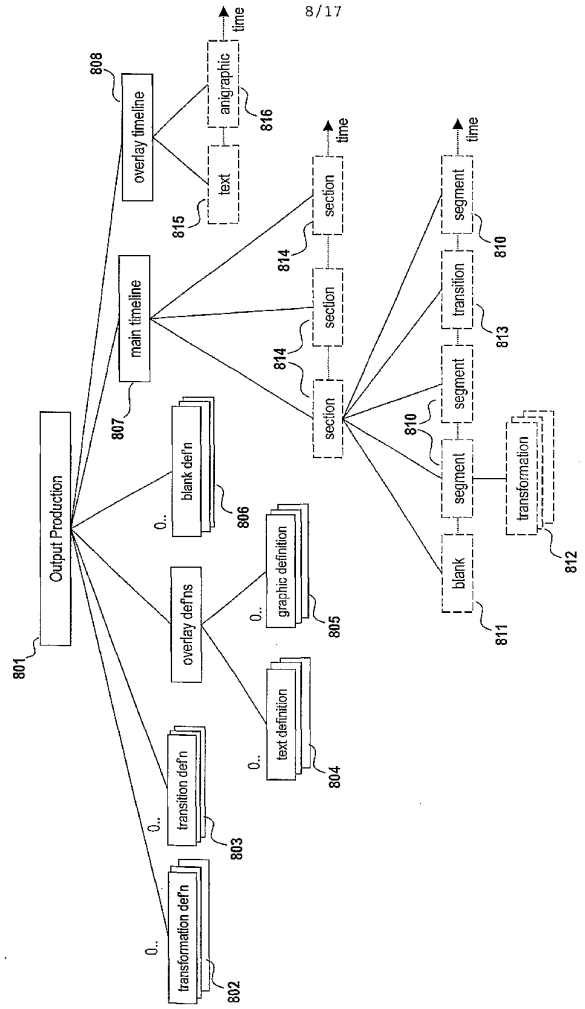

701
Convert descriptors from physical units to perceptual units
702
Apply Unit Normalization
x' = (x - xmin) / (xmax - xmin)
703
Apply Distribution Normalization
x' = ((x - m) * s' / s) + m'
704
Limit Outliers
705
Calculate Distances for all candidate segments, allowing for descriptors weights and ignoring of descriptors as required
D = SQRT(SUM_D((|v_gd - v_cd| ^ 2) * w_d))
706
List candidate segments in order of proximity

Fig. 8


Output Production
801
transformation def'n
802
transition def'n
803
text definition
804
graphic definition
805
blank def'n
806
main timeline
807
overlay timeline
808
overlay def'ns
0..
time
section
814
segment
810
transition
813
blank
811
transformation
812
text
815
anigraphic
816

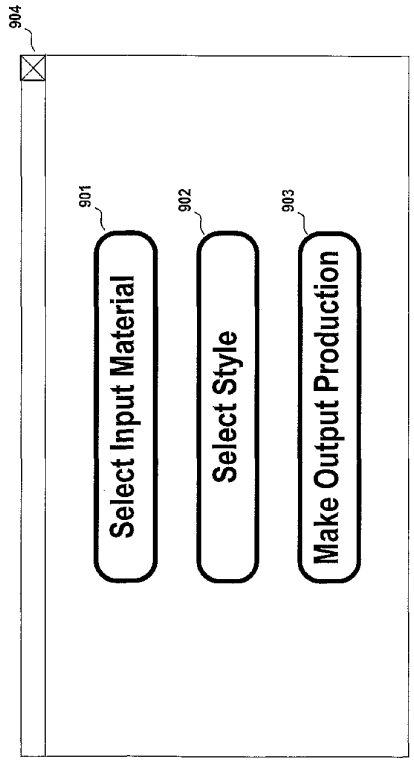

904
901
902
903
Select Input Material
Select Style
Make Output Production

Fig. 9

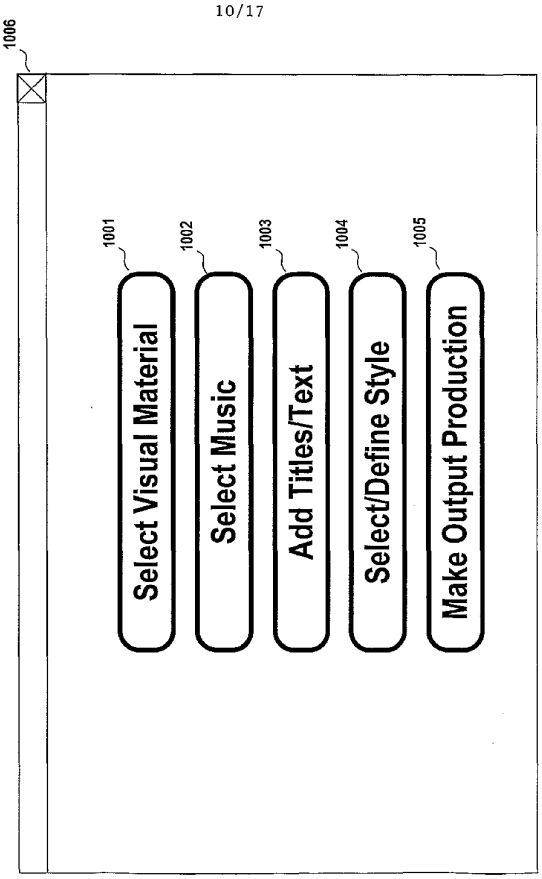

1006
1001
1002
1003
1004
1005
Select Visual Material
Select Music
Add Titles/Text
Select/Define Style
Make Output Production

Fig. 10

Fig. 11

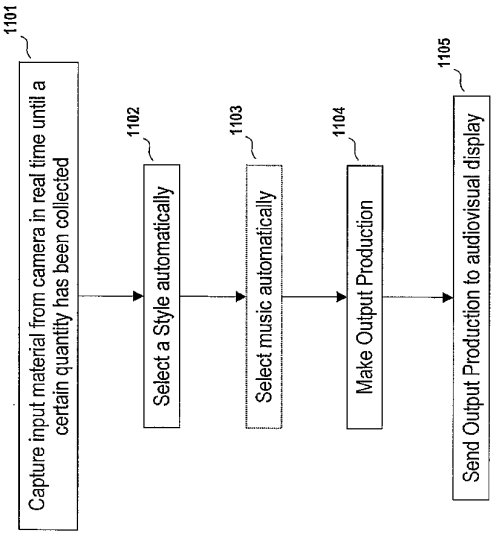

1101
Capture input material from camera in real time until a certain quantity has been collected
1102
Select a Style automatically
1103
Select music automatically
1104
Make Output Production
1105
Send Output Production to audiovisual display

Fig. 12

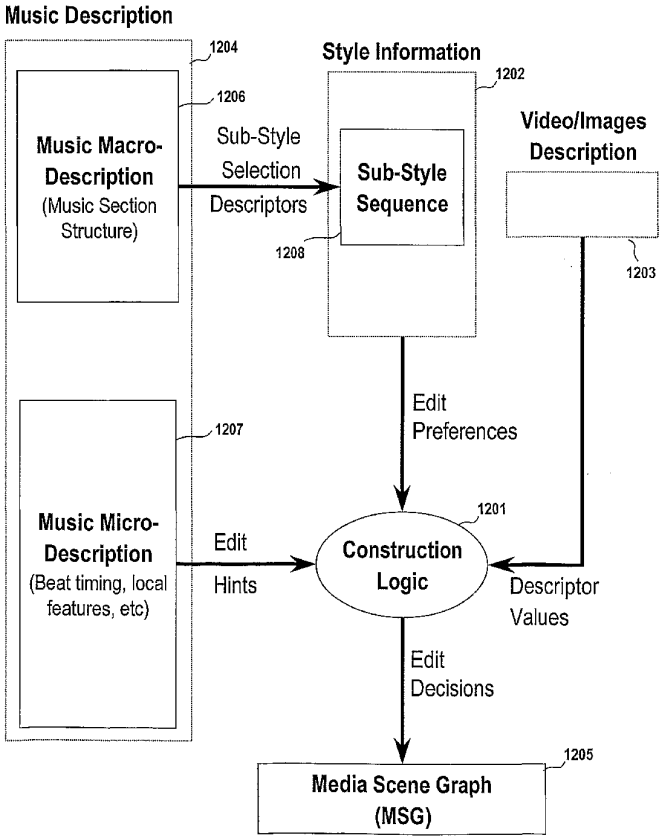

Music Description
1204
Style Information
1202
1206
Music Macro-Description
(Music Section Structure)
Sub-Style Selection Descriptors
Sub-Style Sequence
1208
Video/Images Description
1203
Edit Preferences
1207
Music Micro-Description
(Beat timing, local features, etc)
Edit Hints
1201
Construction Logic
Descriptor Values
Edit Decisions
1205
Media Scene Graph (MSG)

Fig. 13

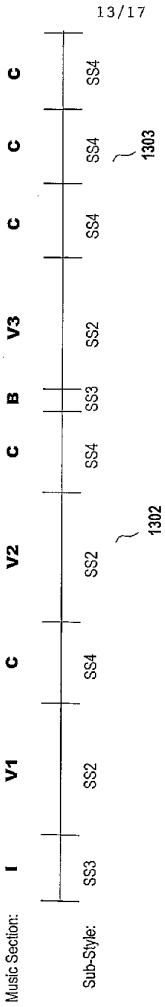

1301
1302
1303
Music Section: I V1 C V2 C B V3 C C C
Sub-Style: SS3 SS2 SS4 SS2 SS4 SS3 SS2 SS4 SS4 SS4
Key:
I: Introduction
V1 – V3: Verses 1 to 3
C: Chorus
B: Bridge
SS2 – SS4: Sub-Styles 2 to 4

Fig. 14

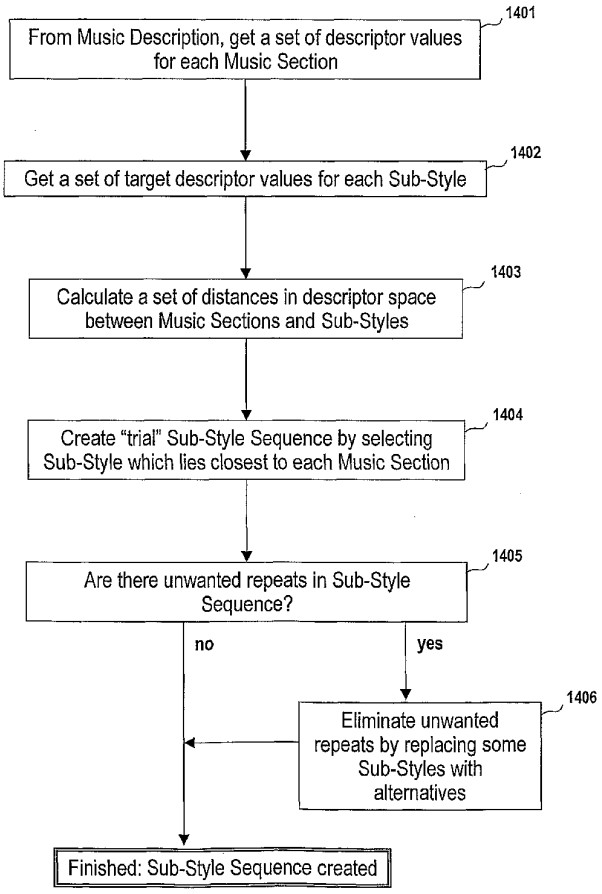

From Music Description, get a set of descriptor values for each Music Section
1401
Get a set of target descriptor values for each Sub-Style
1402
Calculate a set of distances in descriptor space between Music Sections and Sub-Styles
1403
Create "trial" Sub-Style Sequence by selecting Sub-Style which lies closest to each Music Section
1404
Are there unwanted repeats in Sub-Style Sequence?
1405
no
yes
Eliminate unwanted repeats by replacing some Sub-Styles with alternatives
1406
Finished: Sub-Style Sequence created

Fig. 15

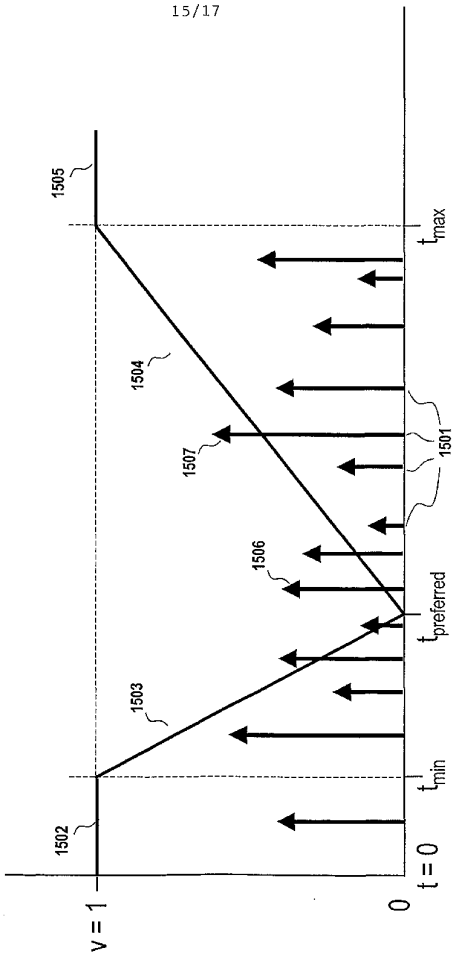

1505
1504
1507
1501
1506
1503
1502
v = 1
0
t = 0
t_{max}
$t_{preferred}$
t_{min}

Fig. 16

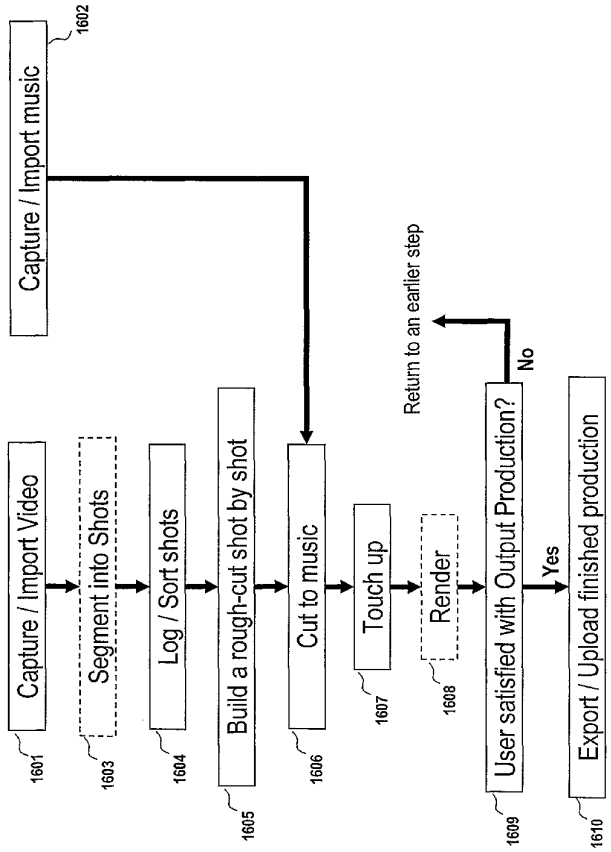
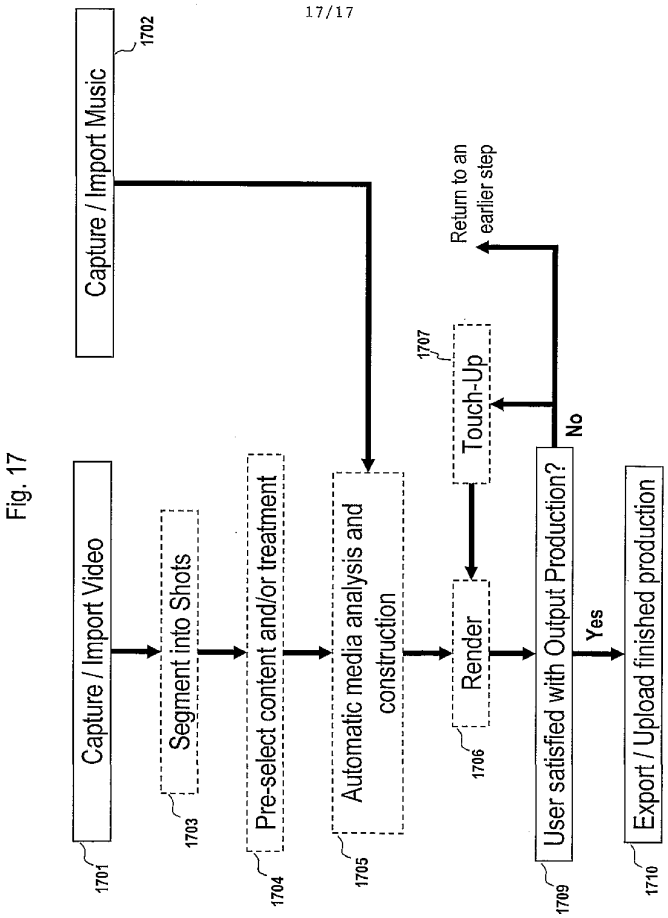

Capture / Import music
1602
Capture / Import Video
1601
Segment into Shots
1603
Log / Sort shots
1604
Build a rough-cut shot by shot
1605
Cut to music
1606
Touch up
1607
Render
1608
User satisfied with Output Production?
1609
Return to an earlier step
No
Yes
Export / Upload finished production
1610

Fig. 17


Capture / Import Video
1701
Capture / Import Music
1702
Segment into Shots
1703
Pre-select content and/or treatment
1704
Automatic media analysis and construction
1705
Render
1706
Touch-Up
1707
Return to an earlier step
User satisfied with Output Production?
1709
No
Yes
Export / Upload finished production
1710

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SG 00/00197

A. CLASSIFICATION OF SUBJECT MATTER

IPC[7]: G11B 27/031, H04N 5/91, G06F 3/14

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC[7]: G06F, G06T, G11B, H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

XPESP

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI EPODOC PAJ XPESP

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 00/39997 A (Dekel et al.), 6 July 2000 (06.07.00) *abstract, figures 1-3, 6-9.*	1-48
A	JP 2000-268540 A (Asukanet KK, Ricoh Co ltd), 29 September 2000 (29.09.00) &Patent Abstracts of Japan *abstract, figure 1.*	1,6,47,48

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
„A" document defining the general state of the art which is not considered to be of particular relevance
„E" earlier application or patent but published on or after the international filing date
„L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
„O" document referring to an oral disclosure, use, exhibition or other means
„P" document published prior to the international filing date but later than the priority date claimed
„T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
„X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
„Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
„&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 September 2001 (28.09.2001)	21 November 2001 (21.11.2001)
Name and mailing adress of the ISA/AT Austrian Patent Office Kohlmarkt 8-10; A-1014 Vienna Facsimile No. 1/53424/535	Authorized officer WERNER Telephone No. 1/53424/357

Form PCT/ISA/210 (second sheet) (July 1998)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/SG 00/00197

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
JP A2 00268540	29-09-2000	none	
WO A 0038897		none	

PCT/ISA/210 (patent family annex) (July 1998)

フロントページの続き

(74)代理人 100100424
弁理士 中村 知公
(74)代理人 100114362
弁理士 萩野 幹治
(72)発明者 ケロック，ピーター，ローワン
シンガポール国 シンガポール ５７４３２７ レイクビュー ＃０８－０２ アッパー トムソン ロード ９７エー
(72)発明者 アルトマン，エドワード，ジェームス
シンガポール国 シンガポール ５９８７３８ シンフォニー ハイツ ＃０５－１２ ヒューム アベニュー ４１

Ｆターム(参考) 5C053 FA14 GB06 GB21 JA01 JA21
5D110 AA12 AA27 AA29 BB01 BB20 CA16 CD02 CD22