

(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION
EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété
Intellectuelle
Bureau international



(43) Date de la publication internationale
6 août 2009 (06.08.2009)

PCT

(10) Numéro de publication internationale
WO 2009/095616 A1

(51) Classification internationale des brevets :
G06F 17/30 (2006.01)

(21) Numéro de la demande internationale :
PCT/FR2009/050129

(22) Date de dépôt international :
28 janvier 2009 (28.01.2009)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
0850580 30 janvier 2008 (30.01.2008) FR

(71) Déposant (pour tous les États désignés sauf US) :
FRANCE TELECOM [FR/FR]; 6, place d'Alleray,
F-75015 Paris (FR).

(72) Inventeurs; et

(75) Inventeurs/Déposants (pour US seulement) : **GENGEM-
BRE, Nicolas** [FR/FR]; Brehil, F-35470 Plechatel (FR).
LECHAT, Patrick [FR/FR]; 7, rue de la Fosse Gauchère,

F-35770 Vern sur Seiche (FR). **BERRANI, Sid Ahmed**
[DZ/FR]; 13, bd de Metz, F-35700 Rennes (FR).

(74) Mandataire : **LEDEY Michel**; FRANCE TELECOM
R & D/PIV/BREVETS, 38-40, rue du Général Leclerc,
F-92794 Issy Moulineaux Cedex 9 (FR).

(81) États désignés (sauf indication contraire, pour tout titre de
protection nationale disponible) : AE, AG, AL, AM, AO,
AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH,
CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG,
ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL,
IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK,
LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW,
MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT,
RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ,
TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM,
ZW.

(84) États désignés (sauf indication contraire, pour tout titre
de protection régionale disponible) : ARIPO (BW, GH,

[Suite sur la page suivante]

(54) Title: METHOD OF IDENTIFYING A MULTIMEDIA DOCUMENT IN A REFERENCE BASE, CORRESPONDING
COMPUTER PROGRAM AND IDENTIFICATION DEVICE

(54) Titre : PROCEDE D'IDENTIFICATION D'UN DOCUMENT MULTIMEDIA DANS UNE BASE DE REFERENCE, PRO-
GRAMME D'ORDINATEUR, ET DISPOSITIF D'IDENTIFICATION CORRESPONDANTS

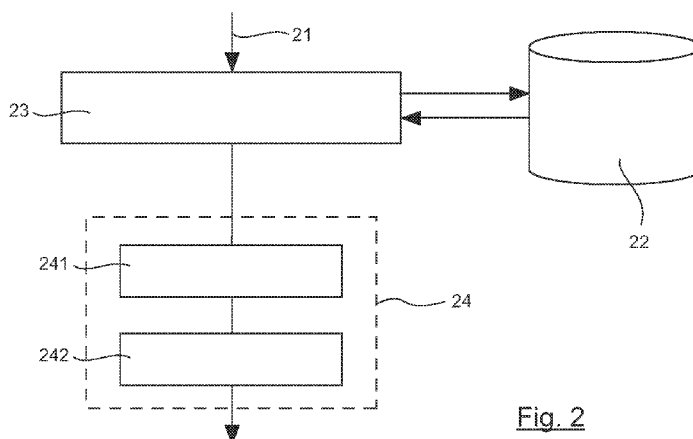


Fig. 2

(57) Abstract: The invention relates to a method of identifying a multimedia document, aimed at verifying whether the multimedia document to be identified (21) is similar or not to at least one multimedia document referenced in a base (22) of reference multimedia documents, comprising the following steps: assignment (23) of a number of votes to at least one reference multimedia document, selection (24) of multimedia documents similar to said multimedia document to be identified. According to the invention, the selection step comprises the following substeps, determination (241) of a probabilistic distribution of the number of votes assigned to a reference multimedia document, as a function of the total number of documents referenced in said base and of the total number of votes, under a random voting assumption; obtainment (242) of a threshold of selection of said similar multimedia documents from among the reference multimedia documents, on the basis of said probabilistic distribution.

[Suite sur la page suivante]

WO 2009/095616 A1



GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), européen (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Publiée :

- avec rapport de recherche internationale
- avant l'expiration du délai prévu pour la modification des revendications, sera republiée si des modifications sont reçues

Déclaration en vertu de la règle 4.17 :

- relative à la qualité d'inventeur (règle 4.17.iv))

(57) Abrégé : L'invention concerne un procédé d'identification d'un document multimédia, visant à vérifier si le document multimédia à identifier (21) est similaire ou non à au moins un document multimédia référencé dans une base (22) de documents multimédia de référence, comprenant les étapes suivantes : affectation (23) d'un nombre de votes à au moins un document multimédia de référence, sélection (24) de documents multimédia similaires audit document multimédia à identifier. Selon l'invention, l'étape de sélection comprend les sous-étapes suivantes : détermination (241) d'une distribution probabiliste du nombre de votes affectés à un document multimédia de référence, en fonction du nombre total de documents référencés dans ladite base et du nombre total de votes, sous une hypothèse de vote aléatoire; obtention (242) d'un seuil de sélection desdits documents multimédia similaires parmi les documents multimédia de référence, à partir de ladite distribution probabiliste.

Procédé d'identification d'un document multimédia dans une base de référence, programme d'ordinateur, et dispositif d'identification correspondants.

1. Domaine de l'invention

5 Le domaine de l'invention est celui de la transmission ou de l'échange de documents multimédia, par exemple une image, une vidéo, un contenu audio, textuel, etc.

Plus précisément, l'invention concerne l'identification de tels documents multimédia, en vue notamment de la détection de copies d'un document référencé
10 (par exemple de copies illicites d'un document protégé).

2. Art antérieur

2.1 Détection de copies illicites

L'avènement du haut débit offert par l'ADSL a induit l'émergence de nouveaux services permettant une consommation facilitée du contenu multimédia,
15 comme les services de vidéo à la demande.

Les fournisseurs historiques, comme France Télévision, TF1, Gaumont, etc (marques déposées) ainsi que d'autres acteurs issus du monde des télécoms, comme Orange, Neuf, Free, etc (marques déposées), des moteurs de recherche comme Google video, Yahoo video, etc (marques déposées) ou encore des
20 sociétés spécialisées comme vodo.fr, glowria, blinkx, TVEyes, skouk, etc (marques déposées), proposent ainsi en ligne une partie de leur catalogue vidéo. Les documents multimédia proposés par ces services sont protégés, et soumis par exemple au paiement d'une redevance pour pouvoir les télécharger.

Par ailleurs, le développement récent des sites d'échange de documents
25 multimédia tels que YouTube, DailyMotion, MySpace, etc (marques déposées) met en évidence l'existence d'une seconde source de documents multimédia. Ces documents proviennent des utilisateurs eux-mêmes. Malheureusement, bien qu'une partie des documents observés sur ces sites d'échanges provienne de documents réellement créés par les utilisateurs, une autre partie est constituée de
30 documents illégalement proposés au téléchargement.

Il est donc souhaitable de pouvoir détecter les copies illicites d'un document multimédia protégé.

Plus précisément, la détection de copies vidéo permet :

- 5 - d'identifier les contenus référencés en catalogue, c'est-à-dire référencés dans une base de référence, afin de détecter les copies illicites des contenus de référence ;
- de relever les contenus fortement copiés (dé-doublonnage) afin de déceler les contenus générateurs d'audience, ou de limiter les tailles de stockage ;
- 10 - de localiser un programme intégral à partir d'un court extrait.

Une telle détection doit pouvoir prendre en compte les altérations usuelles que peut subir un document multimédia dans ce contexte : forte compression, ré-échantillonnage, recadrage, mais aussi incrustation de texte, logos, projections filmées (en anglais « camcording »), etc. En effet, un document multimédia copié
15 subit généralement des transformations intentionnelles, afin de le rendre difficilement détectable, ainsi que des transformations non-intentionnelles, dues à l'enregistrement du document, à son transcodage, ou encore à des contraintes éditoriales lors de sa republication.

Classiquement, la détection de copies de documents multimédia (images,
20 sons, vidéos, etc) consiste à rechercher la présence ou non d'un document requête « suspect » dans une base de documents protégés. Une telle technique repose sur deux aspects essentiels :

- la description du contenu visuel du document multimédia, i.e. les descripteurs utilisés ;
- 25 - la technique d'indexation des descripteurs, i.e. la méthode utilisée pour structurer la base des descripteurs des documents protégés, qui permet d'exécuter efficacement des recherches.

2.2 Descripteurs de documents

Classiquement, le descripteur d'un document est un vecteur numérique qui
30 représente, en le résumant, le contenu du document ou d'une partie du document.

Dans le domaine de l'analyse de contenus vidéo, on utilise couramment une description basée sur les images clés. Cette technique consiste à sélectionner à partir d'un document de type vidéo un sous-ensemble d'images, appelées images clés, et à décrire ces images clés. Par exemple, ces images clés peuvent provenir
5 d'un algorithme sélectionnant de manière adaptative les images représentatives de la vidéo, ou d'un sous-échantillonnage temporel régulier sélectionnant par exemple une image par seconde. Ces images clés sont représentées par un ou plusieurs descripteurs calculés à partir du contenu visuel de l'image.

On distingue deux approches pour les descripteurs :

- 10 - les approches locales : à partir de chaque image clé, un ensemble de points d'intérêt est sélectionné dans l'image. Ces points d'intérêt correspondent à des points visuellement remarquables de l'image que l'on peut retrouver même après altération. Un descripteur est ensuite calculé au voisinage de chaque point d'intérêt ;
- 15 - les approches globales : chaque image de la vidéo, ou chaque image clé est décrite dans son ensemble en calculant un seul descripteur.

En particulier, les descripteurs doivent être robustes aux altérations des documents.

Ainsi, une large part des techniques de détection de copies de documents
20 multimédia utilise une description locale du document, considérant que les descripteurs locaux sont plus robustes que les descripteurs globaux. L'information décrivant le document multimédia est ainsi répartie en différentes régions du document. Par conséquent, l'altération de certaines de ces régions (par exemple lors de l'incrustation d'un logo dans une image, ou encore lors du recadrage ou du
25 rognage de l'image, en anglais « cropping ») n'affecte pas les autres régions qui permettent d'identifier le document.

2.3 Recherche par similarité

Comme déjà indiqué, la détection de copies de documents multimédia consiste à rechercher la présence ou non d'un document requête à identifier dans
30 une base de documents protégés.

Cette recherche repose sur deux phases distinctes :

- une phase dite « hors ligne » pour la construction de la base de documents multimédia de référence ;
- une phase dite « en ligne » pour la recherche de la présence ou non du document à identifier dans la base de référence.

5

Plus précisément, la phase de recherche associe une mesure de similarité (souvent une distance) à un document à identifier. Cette mesure de similarité permet de quantifier la ressemblance entre deux documents en mesurant la proximité entre leurs descripteurs respectifs.

10

Dans une application de détection de copies vidéo par exemple, on ne recherche pas exclusivement des documents identiques, mais également des documents de ressemblance modérée, pour tenir compte des éventuelles altérations subies par la vidéo.

A l'inverse, il ne suffit pas que deux documents aient quelques descripteurs en commun pour qu'ils soient copie l'un de l'autre (par exemple, deux documents textes peuvent avoir des mots en commun sans pour autant traiter du même sujet).

Par conséquent, il est souhaitable de définir de manière efficace le degré de similarité (encore appelé seuil de sélection) à partir duquel on considère que les documents ont une ressemblance significative.

20

En effet, un seuil trop faible entraîne la présence de beaucoup de fausses alarmes, en considérant des documents multimédia non similaires comme similaires, tandis qu'un seuil trop élevé conduit à des non détections, en ne détectant pas certains documents similaires (documents similaires non retournés par le système).

25

La figure 1 illustre plus précisément les différentes étapes mises en œuvre pour la phase en ligne de recherche de la présence ou non d'un document à identifier dans la base de référence.

On considère par exemple un document à identifier Q 11, correspondant à une image.

30

Au cours d'une première étape de description 12, un ensemble de m descripteurs locaux est extrait du document à identifier. On considère que plus l'image est complexe, plus le nombre de descripteurs locaux augmente. A l'inverse, si l'image est simple (image représentant le ciel par exemple), le nombre de descripteurs est faible.

Au cours d'une étape suivante de recherche 13, une requête à la base de documents multimédia de référence 14 renvoie, pour chacun des m descripteurs, un ensemble de documents candidats (zéro, un ou plusieurs) issus de la base de référence et possédant un descripteur similaire. Autrement dit, on associe à chaque descripteur j (pour j allant de 1 à m), D_j documents candidats issus de la base 14.

En particulier, on note que parmi les documents candidats renvoyés, certains apparaissent plusieurs fois, c'est-à-dire qu'ils sont renvoyés par plusieurs des m requêtes, lors de l'étape 13 de recherche par similarité dans la base de référence.

Au cours d'une étape suivante de sélection de documents similaires 15, on décide, en fonction du nombre de leurs apparitions, quels documents peuvent être considérés comme similaires au document à identifier 11. L'étape 15 de sélection de documents similaires peut donc être assimilée à une phase de dépouillement de votes : on considère que chaque descripteur j du document à identifier 11 « vote » pour des documents candidats (zéro, un ou plusieurs), et que les documents candidats ayant reçu le plus de votes seront les plus proches du document à identifier. On obtient ainsi un jeu de documents similaires au document à identifier.

Différentes techniques sont présentées dans la littérature pour le dépouillement des votes dans un système de recherche de documents similaires dans une base de référence.

Ainsi, une première technique repose sur un système de seuillage absolu. Autrement dit, on ne conserve que les documents candidats ayant reçu un nombre de votes supérieur à un seuil prédéterminé.

Il est à noter qu'une telle technique est peu performante, car elle ne s'adapte pas au nombre total de votes émis ou à la taille de la base de référence. Elle génère donc un nombre accru de fausses alarmes et de non détections.

Une autre technique présentée par S.-A. Berrani, L. Amsaleg, et P. Gros.
5 (« *Robust Content-Based Image Searches for Copyright Protection* », Proceedings of the ACM International Workshop on Multimedia Databases, pages 70--77, Nouvelle Orleans, Louisiane, USA, Novembre 2003) repose sur une analyse de la liste ordonnée des documents candidats par ordre croissant du nombre de votes. Une méthode de recherche de saut (méthode dite de Page-Hinkley) permet de
10 séparer la liste des votes non significatifs de ceux qui le sont.

Malheureusement, cette technique requiert une phase d'ordonnancement des documents candidats par le nombre de votes reçus. Cette technique requiert également que les documents candidats dont la similarité est significative se distinguent nettement du bruit de fond (correspondant aux votes non significatifs).
15 Une telle technique est donc contraignante, et coûteuse en termes de ressource et en temps.

3. Exposé de l'invention

L'invention propose une solution nouvelle qui ne présente pas ces inconvénients de l'art antérieur, sous la forme d'un procédé d'identification d'un
20 document multimédia, visant à vérifier si le document multimédia à identifier est similaire ou non à au moins un document multimédia de référence référencé dans une base de documents multimédia de référence, comprenant les étapes suivantes :

- affectation d'un nombre de votes à au moins un document multimédia de référence, chacun desdits votes étant significatif d'une proximité entre un
25 descripteur dudit document multimédia de référence et un descripteur dudit document multimédia à identifier,
 - sélection, parmi lesdits au moins un document multimédia de référence, de documents multimédia similaires audit document multimédia à identifier.
- Selon l'invention, l'étape de sélection comprend les sous-étapes suivantes :
- 30 - détermination d'une distribution probabiliste du nombre de votes affectés à

un document multimédia de référence, en fonction du nombre total de documents référencés dans ladite base et du nombre total de votes, sous une hypothèse de vote aléatoire,

- obtention d'un seuil de sélection desdits documents multimédia similaires, parmi les documents multimédia de référence, à partir de ladite distribution probabiliste.

Ainsi, l'invention propose une solution nouvelle et inventive permettant de déterminer automatiquement un seuil de sélection des documents multimédia de référence similaires au document multimédia à identifier.

- 10 Pour ce faire, on considère un nombre de votes affecté à au moins un document multimédia de référence, et par exemple à tous les documents référencés dans la base. Ainsi, ce nombre de votes sera égal à zéro pour un document n'ayant pas reçu de vote.

- 15 Les documents multimédia (de référence et à identifier) peuvent être des images fixes, des vidéos, des contenus audio, des contenus textuels, etc. Ces documents multimédia sont chacun décrits par au moins un descripteur.

- Plus précisément, si les documents multimédia (à identifier et de référence) sont décrits par au moins deux descripteurs locaux, caractérisant un aspect et/ou une région desdits documents multimédia, on affecte un vote à un document multimédia de référence lorsqu'un des descripteurs du document multimédia à identifier est similaire à un des descripteurs du document multimédia de référence.

- 20 Si les documents multimédia (à identifier et de référence) sont décrits par un descripteur vectoriel global, comprenant au moins deux composantes, on affecte un vote à un document multimédia de référence lorsqu'une des composantes (ou sous-ensemble de composantes) du descripteur du document multimédia à identifier est similaire à une des composantes (ou sous-ensemble de composantes) du descripteur du document multimédia de référence.

- 25 On détermine ensuite une distribution probabiliste du nombre de votes affectés à un document multimédia de référence, en fonction du nombre total de
- 30

documents référencés dans la base et du nombre total de votes. Autrement dit, cette distribution probabiliste est valable pour l'ensemble des documents de référence. Elle permet de représenter le nombre de votes affectés à un document i , sous une hypothèse de vote au hasard. Cette distribution probabiliste est encore
5 appelée représentation probabiliste de la répartition du nombre de votes, ou modélisation probabiliste.

On obtient alors un seuil de sélection des documents multimédia similaires, parmi les documents multimédia de référence de la base, à partir de cette distribution probabiliste.

10 En particulier, le seuil de sélection est défini en tenant compte du nombre de fausses alarmes possibles, estimé à partir de ladite distribution probabiliste, de façon que le nombre de fausses alarmes pour le seuil de sélection soit inférieur à une valeur de décision ε prédéterminée.

Ce seuil de sélection tient donc compte de la distribution probabiliste
15 précédemment déterminée.

Plus précisément, une « fausse alarme » pour un document multimédia de référence revient à considérer ce document comme similaire au document à identifier, alors qu'il ne l'est pas. Le nombre de fausses alarmes peut s'exprimer par le produit du nombre total de documents multimédia référencés dans la base et
20 de la probabilité qu'un document multimédia de référence ait un nombre de votes supérieur ou égal au seuil de sélection S . De nouveau, cette probabilité se calcule sous une hypothèse de vote au hasard.

Par exemple, la valeur de décision est choisie égale à 1 ($\varepsilon=1$).

Le choix de cette valeur de décision permet notamment de s'affranchir
25 d'un paramètre.

En effet, en fixant cette valeur à 1, on sait que statistiquement, moins d'un document multimédia de référence sur l'ensemble des documents multimédia de référence recevra un nombre de votes supérieur au seuil S si les votes se produisent aléatoirement. Si un document multimédia de référence particulier
30 reçoit un nombre de votes dépassant ce seuil S , il constitue une fausse alarme

constatée, alors que la distribution probabiliste suivant le vote aléatoire en prédit moins.

Ainsi, on peut supposer qu'un tel nombre de votes n'est peut être pas dû au hasard mais plutôt à une certaine similarité avec le document multimédia à identifier.

Selon un aspect particulier de l'invention, où les votes aléatoires sont répartis uniformément, la distribution probabiliste met en œuvre une loi binomiale de paramètres V et $1/n$, notée $B\left(V_i; V, \frac{1}{n}\right)$, où :

- n est le nombre total de documents multimédia référencés dans la base ;
- V est le nombre total de votes ;
- V_i est le nombre de votes pour un document multimédia de référence i référencé dans la base.

Une telle loi correspond à l'expérience suivante : on renouvelle V fois de manière indépendante une épreuve de Bernoulli de paramètre $1/n$ (expérience aléatoire à deux issues possibles, généralement dénommées respectivement « succès » et « échec », avec une chance de succès de $1/n$). On compte alors le nombre de succès V_i obtenus à l'issue des V épreuves.

L'ensemble des valeurs prises par V_i suit alors une loi binomiale $B\left(V_i; V, \frac{1}{n}\right)$

En particulier, la loi binomiale peut être approximée par une loi de Poisson de paramètre $L = V/n$, selon l'équation suivante :

$$B(k; V, \frac{1}{n}) \approx \frac{L^k}{k!} \exp(-L).$$

Cette approximation permet notamment de simplifier la mise en œuvre numérique des calculs, et de minimiser les temps de calcul.

En particulier, l'étape d'obtention d'un seuil de sélection met en œuvre un algorithme itératif à partir d'une valeur d'initialisation du seuil de sélection valant zéro et tant que le nombre de fausses alarmes pour le seuil de sélection est supérieur à la valeur de décision ε .

Cet algorithme itératif peut notamment être mis en œuvre lorsque la loi binomiale est approximée par une loi de Poisson.

Selon une variante, le seuil de sélection S est déterminé préalablement à l'étape de sélection pour différentes valeurs du nombre total de documents multimédia référencés dans ladite base (n) et du nombre total de votes (V), et stocké dans une table. L'obtention du seuil de sélection met alors en œuvre une lecture de la table.

Un autre aspect de l'invention concerne un produit programme d'ordinateur téléchargeable depuis un réseau de communication et/ou enregistré sur un support lisible par ordinateur et/ou exécutable par un processeur comprenant des instructions de code de programme pour la mise en œuvre du procédé d'identification décrit précédemment.

Dans un autre mode de réalisation, l'invention concerne un dispositif d'identification d'un document multimédia, visant à vérifier si le document multimédia à identifier est similaire ou non à au moins un document multimédia de référence référencé dans une base de documents multimédia de référence, lesdits documents multimédia à identifier et de référence étant décrits par au moins un descripteur, comprenant :

- des moyens d'affectation d'un nombre de votes à au moins un document multimédia de référence, chacun desdits votes étant significatif d'une proximité entre un descripteur dudit document multimédia de référence et un descripteur dudit document multimédia à identifier,
- des moyens de sélection, parmi lesdits au moins un document multimédia de référence, de documents multimédia similaires audit document multimédia à identifier.

Selon ce mode de réalisation, les moyens de sélection comprennent :

- des moyens de détermination d'une distribution probabiliste du nombre de votes affectés à un document multimédia de référence, en fonction du nombre total de documents référencés dans ladite base et du nombre total de votes, sous une hypothèse de vote aléatoire,

- des moyens d'obtention d'un seuil de sélection desdits documents multimédia similaires parmi les documents multimédia de référence, à partir de ladite distribution probabiliste.

Un tel dispositif d'identification est notamment adapté à mettre en œuvre
5 le procédé d'identification décrit précédemment. Il est par exemple inclus dans un serveur d'analyse, permettant l'échange ou le téléchargement de documents multimédia, et notamment la détection de copies de documents multimédia.

4. Liste des figures

D'autres caractéristiques et avantages de l'invention apparaîtront plus
10 clairement à la lecture de la description suivante d'un mode de réalisation particulier, donné à titre de simple exemple illustratif et non limitatif, et des dessins annexés, parmi lesquels :

- la figure 1 présente les différentes étapes mises en œuvre pour la recherche de documents similaires selon l'art antérieur ;
- 15 - la figure 2 illustre les principales étapes du procédé d'identification selon l'invention ;
- la figure 3 représente un exemple de distribution de probabilité du nombre de votes sous hypothèse de vote aléatoire ;
- la figure 4 présente la structure d'un dispositif d'identification selon un
20 mode de réalisation particulier de l'invention.

5. Description d'un mode de réalisation de l'invention

5.1 Principe général

Le principe général de l'invention repose sur l'utilisation d'une approche probabiliste pour identifier un document multimédia, c'est-à-dire pour vérifier si
25 un ou plusieurs documents multimédia référencés dans une base de documents multimédia de référence sont similaires (ou non) avec le document multimédia à identifier. Un tel document multimédia peut être une image (éventuellement extraite d'une vidéo), une vidéo, un contenu audio, un contenu textuel, etc.

Plus précisément, l'invention permet de décider quels documents
30 multimédia de référence peuvent être considérés comme similaires au document à

identifier, en tenant compte d'un seuil de sélection déterminé automatiquement.

Par seuil de sélection déterminé automatiquement, on entend un seuil qui n'est pas pré-établi (comme dans les techniques mettant en œuvre un seuillage absolu), mais qui est calculé automatiquement par l'algorithme de l'invention.

5 La figure 2 illustre plus précisément le principe général de l'identification d'un document multimédia selon l'invention, visant à vérifier si un document multimédia à identifier 21 est similaire ou non à au moins un document multimédia référencé dans une base 22 de documents multimédia de référence décrits chacun par au moins un descripteur.

10 Pour ce faire, au cours d'une première étape 23, on affecte un nombre de votes à au moins un des documents multimédia référencé dans la base 22. Chacun de ces votes est significatif d'une proximité entre un descripteur du document multimédia de référence et un descripteur du document multimédia à identifier. Par exemple, on affecte un nombre de votes à chacun des documents référencés
15 dans la base 22. Les documents de référence ne recevant pas de vote se voient attribuer un nombre de votes égal à zéro.

Par exemple, dans le cas d'un document multimédia décrit à partir de descripteurs locaux, on associe zéro, un ou plusieurs documents multimédia de référence à chaque descripteur local j , en recherchant dans la base 22 les
20 documents multimédia de référence comprenant ce descripteur ou un descripteur proche de celui-ci (en termes de distance par exemple). Autrement dit, on considère que chaque descripteur j du document à identifier « vote » pour des documents multimédia de référence (zéro, un ou plusieurs).

Dans le cas d'un document multimédia décrit à partir d'un descripteur
25 global, on associe zéro, un ou plusieurs documents multimédia de référence à chaque composante du descripteur global. Autrement dit, on considère que chaque composante du descripteur global du document à identifier « vote » pour des documents multimédia de référence (zéro, un ou plusieurs).

Par exemple, si la base 22 comprend quatre documents multimédia de
30 référence notés D1 à D4, et que le document multimédia à identifier est décrit par

trois descripteurs locaux, le premier descripteur local peut voter pour les documents multimédia de référence D1 et D3, le deuxième descripteur local peut voter pour le document multimédia de référence D3, et le troisième descripteur local peut ne voter pour aucun document multimédia de référence. Alors le
5 nombre de votes affectés au document D1 sera égal à 1, le nombre de votes affectés aux documents D2 et D4 sera égal à 0, et le nombre de votes affectés au document D3 sera égal à 2. Le nombre de votes total sera alors égal à 3.

On sélectionne ensuite (24), dans la base 22, les documents multimédia similaires au document multimédia à identifier 21.

10 Pour ce faire, on détermine tout d'abord (241) une distribution probabiliste du nombre de votes affectés à un document multimédia de référence, en fonction du nombre total de documents présents dans la base et du nombre total de votes, sous une hypothèse de vote aléatoire. Une telle modélisation vaut pour tous les documents multimédia de référence.

15 On obtient ensuite (242) un seuil de sélection des documents multimédia similaires parmi les documents multimédia de référence de la base, à partir de la distribution probabiliste, les documents multimédia similaires présentant un nombre de votes supérieur au seuil de sélection. Pour ce faire, on peut notamment tenir compte du nombre de fausses alarmes possibles, estimé à partir de la
20 distribution probabiliste.

Autrement dit, seuls les documents multimédia de référence présentant un nombre de votes supérieur au seuil de sélection sont considérés comme des documents similaires au document multimédia à identifier.

En particulier, le procédé selon l'invention peut être mis en œuvre de
25 diverses manières, notamment sous forme câblée ou sous forme logicielle.

5.2 *Cas des descripteurs locaux*

On décrit ci-après un exemple de mise en œuvre de l'invention, dans lequel la distribution probabiliste du nombre de votes affectés aux documents multimédia de référence est une loi binomiale. On considère également que le
30 document multimédia à identifier est décrit par une pluralité de descripteurs

locaux.

Plus précisément, on note n le nombre de documents multimédia référencés dans la base de documents multimédia de référence, et i l'un de ces documents multimédia de référence $i \in [1, n]$.

5 On note V_i le nombre de votes reçus par le document i (V_i pouvant être égal à 0), et V le nombre total de votes, reçus par l'ensemble des documents multimédia de référence. Ces votes sont issus de la recherche par similarité d'un ensemble de descripteurs d'un document à identifier Q dans la base de référence, comme décrit en relation avec l'art antérieur.

10 On cherche selon l'invention à déterminer le seuil de sélection S correspondant au nombre de votes minimal pour lequel on peut supposer que document multimédia de référence i est similaire au document multimédia à identifier Q .

Afin de déterminer ce seuil de sélection S , on se place donc dans une
15 hypothèse *a contrario*, en considérant que chacun des V votes a été effectué en choisissant au hasard, et de façon uniforme, un document multimédia de référence parmi les n documents multimédia référencés dans la base (hypothèse de vote aléatoire). Pour chaque vote, la probabilité de voter pour le document multimédia de référence i est alors de $1/n$.

20 En effet, le raisonnement *a contrario* dans ce contexte permet de s'interroger si le hasard suffit à expliquer les points communs observés entre le document à identifier et les documents de référence. Si ce n'est pas le cas, alors il y a effectivement ressemblance entre les documents.

Le fait de voter pour le document multimédia de référence i est un
25 phénomène aléatoire à deux issues possibles (généralement dénommées « succès » et « échec ») dont la distribution de probabilité suit la loi appelée distribution de Bernoulli de paramètre $1/n$. Autrement dit, si on choisit au hasard et de façon uniforme un document multimédia de référence de la base, il y a une chance sur n de choisir le document i . Ainsi, si l'on choisit le document i , le
30 résultat est un succès, et si on choisit un autre document de la base, le résultat est

un échec.

Quand on reproduit cette expérience V fois, avec V correspondant au nombre total de votes, la probabilité que l'on choisisse le document i plusieurs fois (V_i fois) suit quant à elle une loi binomiale à deux paramètres: V et $1/n$.

- 5 Ainsi, la probabilité que le document multimédia de référence i reçoive exactement V_i votes suit la loi binomiale de paramètres V et $1/n$. On note $B\left(V_i; V, \frac{1}{n}\right)$ cette probabilité.

On détermine ainsi une représentation probabiliste de la répartition des nombres de votes affectés à un document multimédia de référence (i), en fonction
10 du nombre total de documents présents dans ladite base (n), et du nombre total de votes (V).

On cherche ensuite à déterminer un seuil de sélection S des documents multimédia similaires (avec S un entier).

- On peut écrire la probabilité que le nombre de votes affectés au document
15 i , noté V_i , soit supérieur ou égal au seuil de sélection S sous la forme suivante :

$$p(V_i \geq S) = 1 - \sum_{k=0}^{S-1} B(k; V, \frac{1}{n})$$

- La figure 3 représente un exemple de distribution de probabilité du nombre de votes sous hypothèse de vote aléatoire. Plus précisément, la partie hachurée représente la probabilité que le nombre de votes pour un document multimédia de
20 référence i soit supérieur ou égal au seuil S .

- Selon cet exemple de mise en œuvre de l'invention, la décision sur la similarité ou non du document multimédia de référence i avec le document multimédia à identifier Q s'effectue en calculant, pour différentes valeurs de S croissantes, le seuil de sélection à partir duquel le nombre estimé de fausses
25 alarmes observées est inférieur à une valeur de décision, par exemple égale à 1. Cela signifie qu'un vote « au hasard » ne suffit pas à expliquer un tel nombre de votes, mais qu'une certaine similarité en est responsable. Ce nombre de fausses alarmes peut être estimé à partir de la distribution probabiliste illustrée en figure 3. Dans cet exemple, le nombre de fausses alarmes, noté $NFA(S)$, correspond au

nombre de documents multimédia de référence ayant reçu au moins S votes quand ceux-ci sont effectués au hasard.

Le nombre de fausses alarmes s'exprime par le produit de la probabilité qu'un document multimédia de référence ait un nombre de votes supérieur ou égal
5 au seuil de sélection S , par le nombre total de documents multimédia dans la base :

$$NFA(S) = n.p(V_i \geq S)$$

On peut également noter que la loi binomiale $B\left(V_i; V, \frac{1}{n}\right)$ qui intervient s'exprime à l'aide de combinaisons, elles-mêmes exprimées par des factoriels
10 (factoriel V notamment).

Pour des soucis de facilité de mise en œuvre numérique des calculs, il est possible d'approcher, de façon très fiable, la loi binomiale par une loi de Poisson dont le paramètre L vaut V/n .

On peut noter qu'une telle approximation est valable quand $1/n$ est petit et
15 V grand, ce qui est le cas en règle générale pour ce contexte (en pratique, cette approximation est utilisée quand $V > 30$ et $L < 5$).

Ainsi, on peut approcher la loi binomiale par l'expression suivante :

$$B(k; V, \frac{1}{n}) \approx \frac{L^k}{k!} \exp(-L)$$

Bien que la loi de Poisson fasse également intervenir une factorielle, cette
20 factorielle ne concerne cette fois, dans la mise en œuvre proposée, que des petites valeurs, et est facilement calculable.

Il est également possible de déduire une formulation récursive de la loi binomiale ainsi approchée :

- pour $k = 0$: $B(0; V, \frac{1}{n}) \approx \exp(-L)$;
- 25 - pour $k > 0$: $B(k; V, \frac{1}{n}) = \frac{L}{k} B(k-1; V, \frac{1}{n})$.

On peut ensuite se servir de cette formulation pour déterminer la valeur du seuil de sélection S .

On introduit les notations suivantes:

- $L = V/n$, où L est le paramètre de la loi de Poisson ;
- s correspond à différentes valeurs de seuil testées ;
- les grandeurs p et b , associés à la variable s , sont définies comme suit:

- 5
 - b est la probabilité qu'un document multimédia de référence ait reçu exactement s votes sous l'hypothèse de vote aléatoire précédemment décrite ;
 - p est la probabilité qu'un document multimédia de référence ait reçu au moins s votes sous l'hypothèse de
- 10 votes aléatoires précédemment décrite.

On initialise tout d'abord les variables :

- $s = 0$, correspondant à la première valeur de seuil de sélection testée ;
- $b = \exp(-L)$, correspondant à la probabilité qu'un document multimédia de référence ait reçu exactement zéro vote sous l'hypothèse de votes aléatoires précédemment décrite ;
- 15 - $p = 1$, correspondant à la probabilité qu'un document multimédia de référence ait reçu au moins zéro vote sous l'hypothèse de votes aléatoires précédemment décrite.

20 On réitère ensuite les étapes suivantes, tant que la probabilité de fausses alarmes NFA est supérieure à une valeur de décision ε prédéterminée, égale à 1 par exemple.

Ainsi, tant que $n.p > \varepsilon$ (soit $NFA(s) > \varepsilon$) :

- on incrémente la variable s de 1 ($s:=s+1$) et on met à jour les
- 25 variables qui en dépendent:
- on affecte la probabilité $p - b$ à la variable p ($p:=p - b$), qui devient ainsi la probabilité qu'un document multimédia de référence i ait reçu au moins s votes sous l'hypothèse de votes aléatoires précédemment décrite ;
- 30 - on affecte la probabilité $b \times L / s$ à la variable b ($b:=b * L/s$), qui

devient ainsi la probabilité qu'un document multimédia de référence i ait reçu exactement s votes sous l'hypothèse de votes aléatoires précédemment décrite ;

5 Finalement, quand la probabilité de fausses alarmes $NFA(s)$ est inférieure ou égale à la valeur de décision ε prédéterminée, avec $\varepsilon = 1$ par exemple, on affecte au seuil de sélection S la valeur finale de s . Les documents multimédia de référence ayant reçu un nombre de votes supérieur ou égal à S sont supposés similaires et sont retournés par la procédure.

 Selon une autre variante, on considère que le nombre de fausses alarmes
10 peut se déduire directement d'une valeur de seuil de sélection, c'est-à-dire que la valeur $NFA(s)$ peut être calculée sans utiliser la valeur $NFA(s-1)$. Comme la fonction $NFA(s)$ est monotone et décroissante en fonction de s , la détermination du seuil de sélection peut alors être mise en œuvre par dichotomie : la probabilité de fausses alarmes $NFA(s)$ est calculée pour différentes valeurs de s dans un
15 intervalle de valeurs possibles (généralement avec une borne inférieure de 0 et une borne supérieure liée au nombre de descripteurs utilisés). Les valeurs de s sont choisies de façon à diviser l'intervalle en deux sous-intervalles. L'estimation des probabilités de fausses alarmes $NFA(s)$ aux bornes de ces sous-intervalles et la propriété de monotonie permettent de localiser le sous-intervalle dans lequel la
20 fonction $NFA(s)$ passe par la valeur ε . On ne conserve que ce sous-intervalle et on répète les mêmes opérations, jusqu'à obtenir un intervalle dont les bornes sont deux entiers consécutifs. La valeur du seuil de sélection S recherchée est alors déterminée par la borne supérieure de cet intervalle.

 Selon une autre variante de mise en œuvre, le seuil de sélection S peut être
25 calculé à partir d'une des méthodes évoquées ci-dessus de façon préalable pour différentes valeurs possibles de V et de n , et stocké dans une table (si l'on utilise une base de données ayant un nombre fixé de documents de référence, on peut aussi effectuer cette tabulation uniquement pour différentes valeurs de V). Ainsi, durant une phase d'analyse, il n'est plus nécessaire de calculer la valeur de seuil S ,
30 mais il suffit de la lire dans ladite table, économisant ainsi encore du temps de

calcul.

5.3 Cas des descripteurs globaux

Selon l'invention, le document multimédia à identifier peut être décrit par un descripteur global, au lieu d'une pluralité de descripteurs locaux.

5 Un tel descripteur global prend généralement la forme d'un vecteur à m dimensions.

Dans ce cas, on applique la même technique que décrite précédemment, en assimilant chaque composante (ou sous-ensemble de composantes) du descripteur global à un descripteur local. Autrement dit, on considère que chaque composante
10 (ou sous-ensemble de composantes) du descripteur global du document à identifier « vote » pour un ensemble de documents multimédia de référence (zéro, un ou plusieurs).

5.4 Avantages liés à l'invention

La technique selon l'invention présente de nombreux avantages, selon au
15 moins un de ses modes de réalisation, et notamment :

- elle ne nécessite aucun paramètre à régler, si la valeur de décision ε prédéterminée est fixée à $\varepsilon = 1$;
- le seuil de sélection est évalué automatiquement, et ne nécessite pas de manipulation coûteuse des listes de valeurs prises par les nombres de
20 votes. En particulier, la décision de similarité ou d'absence de similarité par rapport au seuil de sélection ne requiert aucun ordonnancement de documents multimédia en fonction de leur nombre de votes. De même, le nombre de votes affectés à un « bon » document multimédia de référence (c'est-à-dire à un document multimédia de référence similaire à un
25 document multimédia à identifier) n'a pas besoin de se distinguer nettement de ceux affectés à des documents multimédia de référence non significatifs pour être détecté ;
- elle s'appuie sur un formalisme probabiliste rigoureux ;
- elle permet de contrôler le nombre de fausses alarmes. Indirectement, on
30 peut déduire la probabilité qu'un document multimédia de référence

sélectionné soit une fausse alarme, du nombre de votes qu'il a reçu. Cette caractéristique peut être utile notamment dans le cas d'un système de détection de copies vidéo pour lequel un filtrage séquentiel permet d'agréger temporellement les résultats obtenus à chaque image ;

- 5 - elle implique très peu de calculs et son exécution est donc rapide : selon un mode de réalisation particulier, elle permet d'écourter le délai de prise de décision avant d'avoir analysé tous les descripteurs locaux (ou toutes les composantes d'un descripteur global) du document multimédia à identifier. On peut décider, quand V' votes ont été collectés (avec $V' < V$,
10 où V est le nombre de votes total attribués en prenant en compte tous les descripteurs), d'évaluer ou de lire dans une table le seuil de sélection S associées aux valeurs V' et n , et de l'utiliser pour sélectionner d'éventuels documents multimédia de référence similaires au document multimédia à identifier. On peut alors choisir d'arrêter l'analyse dès lors qu'un
15 document multimédia de référence au moins a été identifié comme similaire.

5.5 Application de l'invention

L'invention peut notamment être mise en œuvre dans un système de détection de copies d'un document multimédia référencé (par exemple de copies
20 illicites d'un document protégé).

Par exemple, elle permet de détecter de manière efficace la présence de copies d'un contenu vidéo protégé au sein d'un flux vidéo suspect. En particulier, l'utilisation de descripteurs locaux selon un mode de réalisation de l'invention permet à cette détection d'être robuste à des altérations, volontaires ou non, du
25 document original.

L'invention peut ainsi s'intégrer dans un système automatique de protection des droits d'auteurs. Elle permet par exemple à une plateforme d'échange de contenus, tels que Youtube, MaZoneVidéo, Dailymotion, etc (marques déposées) d'intervenir très en amont du processus de dépôt de
30 documents multimédia (texte, image, audio ou vidéo) en filtrant les documents

illicites déposés, et de se mettre ainsi en conformité avec les règles de protection des droits d'auteurs.

Par ailleurs, et toujours dans le contexte des plateformes d'échanges de contenus, un tel système peut être utilisé pour détecter les multiples copies d'un même document référencé dans une base d'un serveur. En effet, un même document est en général chargé par plusieurs utilisateurs avec des noms et des descriptions textuelles différentes. Un tel système de détection de copies peut ainsi être appliqué à un moteur de recherche de documents multimédia pour supprimer les doublons dans la base et fournir des résultats de requête dédoublonnés. On présente ainsi à l'utilisateur une occurrence unique de chaque document multimédia (éventuellement avec un lien vers les autres copies).

Un tel outil peut également être utilisé à des fins d'analyse pour du contenu dont la diffusion est autorisée mais dont on souhaite connaître l'audience. Une autre application possible encore est la localisation et la restitution d'un programme (émission TV, vidéo, ...) à partir d'un extrait du document.

Plus généralement, la technique d'obtention d'un seuil de sélection et de dépouillement de votes selon l'invention peut s'appliquer à n'importe quel type de document multimédia (son, texte, images fixes, vidéo), ainsi qu'à tout système mettant en jeu une stratégie de vote avec un grand nombre (non infini) de candidats potentiels.

5.6 Structure du dispositif d'identification

On présente finalement, en relation avec la figure 4, la structure simplifiée d'un dispositif d'identification mettant en œuvre une technique d'identification selon le mode de réalisation particulier décrit ci-dessus.

Un tel dispositif comprend une mémoire 41 constituée d'une mémoire tampon, une unité de traitement 42, équipée par exemple d'un microprocesseur μ P, et pilotée par le programme d'ordinateur 43, mettant en œuvre le procédé d'identification selon l'invention.

A l'initialisation, les instructions de code du programme d'ordinateur 43 sont par exemple chargées dans une mémoire RAM avant d'être exécutées par le

processeur de l'unité de traitement 42. L'unité de traitement 42 reçoit en entrée un document multimédia à identifier 21.

Le microprocesseur de l'unité de traitement 42 met en œuvre les étapes du procédé d'identification décrit précédemment, selon les instructions du programme d'ordinateur 43, pour vérifier si le document multimédia à identifier est similaire ou non à au moins un document multimédia référencé dans une base de documents multimédia de référence. Pour cela, le dispositif d'identification comprend, outre la mémoire tampon 41, des moyens d'affectation d'un nombre de votes à au moins un document multimédia de référence et des moyens de sélection, parmi lesdits au moins un document multimédia de référence, de documents multimédia similaires au document multimédia à identifier. Plus précisément, les moyens de sélection comprennent :

- des moyens de détermination d'une distribution probabiliste du nombre de votes affectés à un document multimédia de référence, en fonction du nombre total de documents référencés dans la base et du nombre total de votes, sous une hypothèse de vote aléatoire,
- des moyens d'obtention d'un seuil de sélection des documents multimédia similaires parmi les documents multimédia de référence, à partir de ladite distribution, les documents multimédia similaires présentant un nombre de votes supérieur au seuil de sélection.

Ces différents moyens sont pilotés par le microprocesseur de l'unité de traitement 42.

Le dispositif d'identification délivre en sortie zéro, un ou plusieurs documents multimédia de référence de la base, présentant un nombre de votes supérieur au seuil de sélection.

Un tel dispositif peut notamment être intégré dans un système de détection de copies de documents multimédia.

REVENDICATIONS

1. Procédé d'identification d'un document multimédia, visant à vérifier si le document multimédia à identifier (21) est similaire ou non à au moins un document multimédia de référence référencé dans une base (22) de documents multimédia de référence, comprenant les étapes suivantes :
 - affectation (23) d'un nombre de votes à au moins un document multimédia de référence, chacun desdits votes étant significatif d'une proximité entre un descripteur dudit document multimédia de référence et un descripteur dudit document multimédia à identifier,
 - 10 - sélection (24), parmi lesdits au moins un document multimédia de référence, de documents multimédia similaires audit document multimédia à identifier,
caractérisé en ce que ladite étape de sélection comprend les sous-étapes suivantes :
 - 15 - détermination (241) d'une distribution probabiliste du nombre de votes affectés à un document multimédia de référence, en fonction du nombre total de documents référencés dans ladite base et du nombre total de votes, sous une hypothèse de vote aléatoire,
 - obtention (242) d'un seuil de sélection desdits documents multimédia
20 similaires parmi les documents multimédia de référence, à partir de ladite distribution probabiliste.
2. Procédé d'identification selon la revendication 1, caractérisé en ce que ledit seuil de sélection est défini en tenant compte du nombre de fausses alarmes possibles, estimé à partir de ladite distribution probabiliste, de façon que le
25 nombre de fausses alarmes pour ledit seuil de sélection soit inférieur à une valeur de décision prédéterminée.
3. Procédé d'identification selon la revendication 2, caractérisé en ce que ladite valeur de décision est égale à 1.
4. Procédé d'identification selon l'une quelconque des revendications 1 à 3,
30 caractérisé en ce que ladite distribution probabiliste met en œuvre une loi

binomiale $B\left(V_i; V, \frac{1}{n}\right)$, où :

- n est le nombre total de documents multimédia référencés dans ladite base ;
- V est le nombre total de votes ;
- 5 – V_i est le nombre de votes pour un document multimédia de référence i référencés dans ladite base.

5. Procédé d'identification selon la revendication 4, caractérisé en ce que ladite loi binomiale est approximée par une loi de Poisson de paramètre $L = V/n$, selon l'équation suivante :

$$10 \quad B(k; V, \frac{1}{n}) \approx \frac{L^k}{k!} \exp(-L).$$

6. Procédé d'identification selon la revendication 2 et l'une quelconque des revendications 3 à 5, caractérisé en ce que ladite étape d'obtention (242) d'un seuil de sélection met en œuvre un algorithme itératif à partir d'une valeur d'initialisation du seuil de sélection valant zéro et tant que le nombre de fausses
15 alarmes pour ledit seuil de sélection est supérieur à ladite valeur de décision.

7. Procédé d'identification selon l'une quelconque des revendications 1 à 6, caractérisé en ce que ledit seuil de sélection S est déterminé préalablement à ladite étape de sélection (24) pour différentes valeurs du nombre total de documents multimédia référencés dans ladite base et du nombre total de votes, et stocké dans
20 une table,

et en ce que ladite étape d'obtention (242) d'un seuil de sélection met en œuvre une lecture de ladite table.

8. Procédé d'identification selon l'une quelconque des revendications 1 à 7, caractérisé en ce que lesdits documents multimédia appartiennent au groupe
25 comprenant :

- une image ;
- une vidéo ;
- un contenu audio ;
- un contenu textuel.

9. Procédé d'identification selon l'une quelconque des revendications 1 à 8, caractérisé en ce que lesdits documents multimédia sont décrits par au moins deux descripteurs locaux, caractérisant un aspect et/ou une région desdits documents multimédia, un vote étant affecté à un document multimédia de référence
5 lorsqu'un des descripteurs du document multimédia à identifier est similaire à un des descripteurs dudit document multimédia de référence.

10. Procédé d'identification selon l'une quelconque des revendications 1 à 8, caractérisé en ce que lesdits documents multimédia sont décrits par un descripteur vectoriel global comprenant au moins deux composantes, un vote étant affecté à
10 un document multimédia de référence lorsqu'une des composantes du descripteur du document multimédia à identifier est similaire à une des composantes du descripteur dudit document multimédia de référence.

11. Produit programme d'ordinateur téléchargeable depuis un réseau de communication et/ou enregistré sur un support lisible par ordinateur et/ou
15 exécutable par un processeur, caractérisé en ce qu'il comprend des instructions de code de programme pour la mise en œuvre du procédé d'identification selon l'une au moins des revendications 1 à 10.

12. Dispositif d'identification d'un document multimédia, visant à vérifier si le document multimédia à identifier (21) est similaire ou non à au moins un
20 document multimédia de référence référencé dans une base (22) de documents multimédia de référence, comprenant :

- des moyens d'affectation (23) d'un nombre de votes à au moins un document multimédia de référence, chacun desdits votes étant significatif d'une proximité entre un descripteur dudit document multimédia de
25 référence et un descripteur dudit document multimédia à identifier,
- des moyens de sélection (24), parmi lesdits au moins un document multimédia de référence, de documents multimédia similaires audit document multimédia à identifier,

caractérisé en ce que lesdits moyens de sélection comprennent :

- des moyens de détermination (241) d'une distribution probabiliste du nombre de votes affectés à un document multimédia de référence, en fonction du nombre total de documents référencés dans ladite base et du nombre total de votes, sous une hypothèse de vote aléatoire,
- 5 - des moyens d'obtention (242) d'un seuil de sélection desdits documents multimédia similaires parmi les documents multimédia de référence, à partir de ladite distribution probabiliste.

1/2

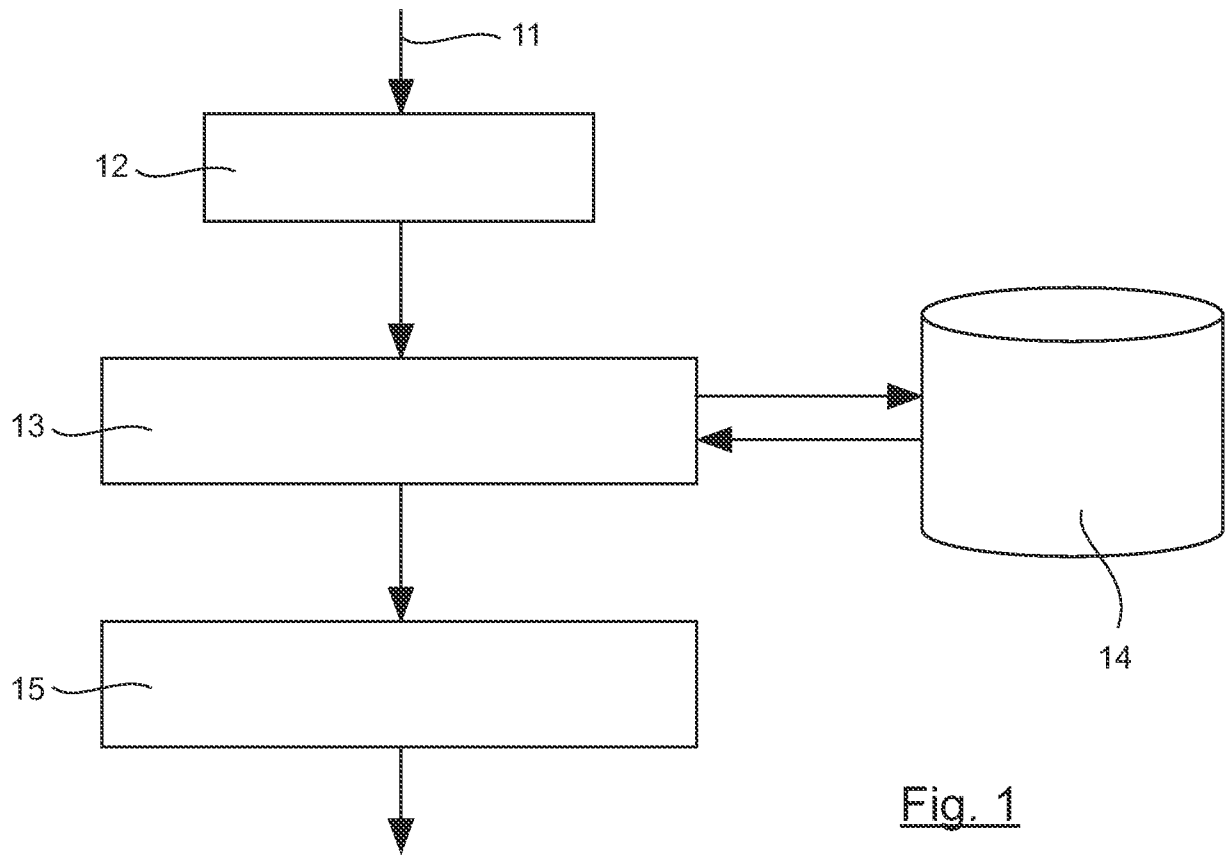


Fig. 1

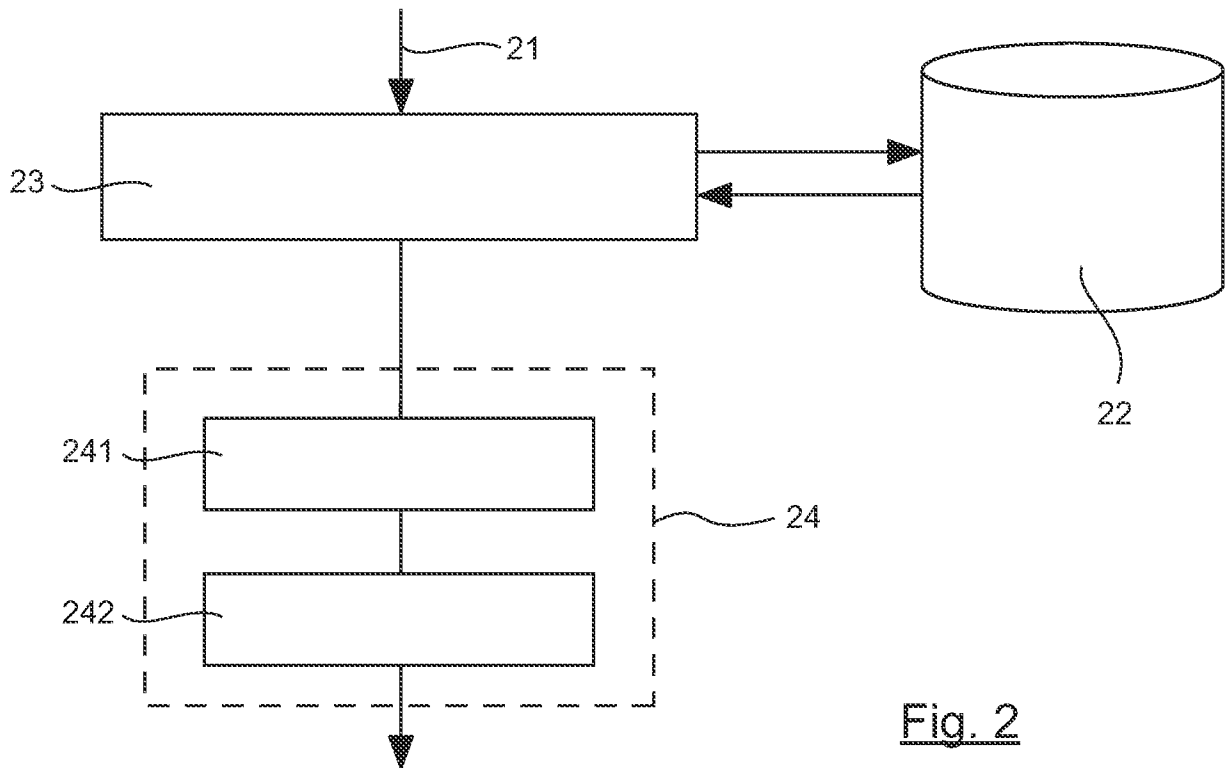
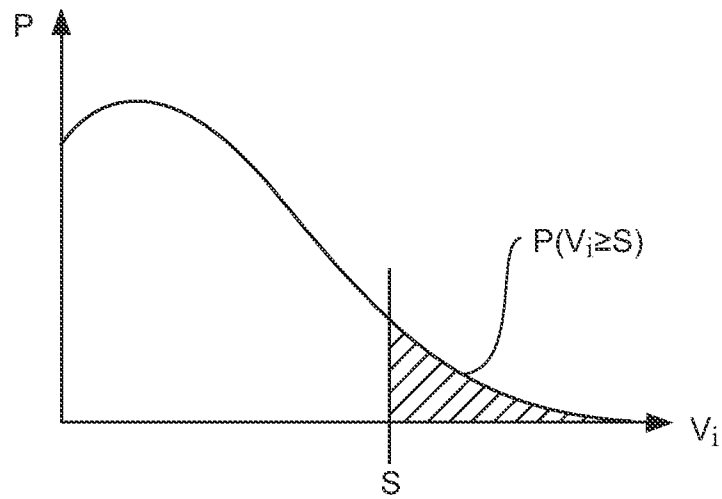
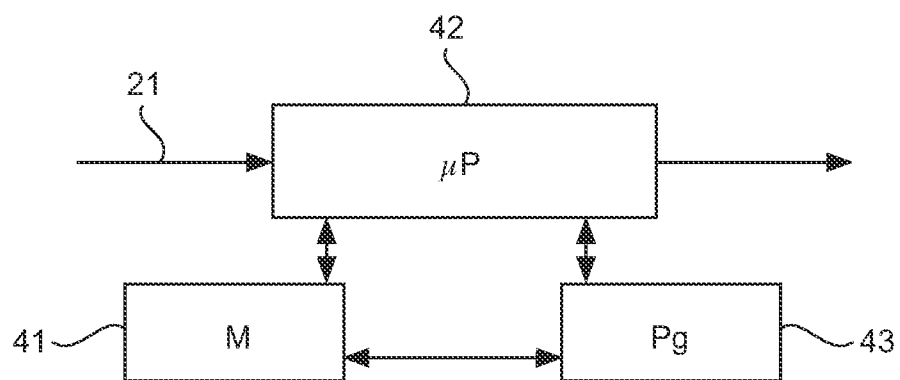


Fig. 2

2/2

Fig. 3Fig. 4

INTERNATIONAL SEARCH REPORT

International application No
PCT/FR2009/050129

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06F17/30

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)

G06F

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

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

EPO-Internal, WPI Data, INSPEC, IBM-TDB, COMPENDEX

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ALEXIS JOLY ET AL: "Content-Based Copy Retrieval Using Distortion-Based Probabilistic Similarity Search" IEEE TRANSACTIONS ON MULTIMEDIA, IEEE SERVICE CENTER, PISCATAWAY, NJ, US, vol. 9, no. 2, 1 February 2007 (2007-02-01), pages 293-306, XP011157474 ISSN: 1520-9210 abstract page 293, left-hand column, line 19 - line 28 page 295, right-hand column, line 26 - line 45 page 299, left-hand column, line 47 - page 300, left-hand column, line 6 ----- -/--	1-12

☒ 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 document 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

9 juin 2009

Date of mailing of the international search report

17/06/2009

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Deane, Inigo

INTERNATIONAL SEARCH REPORT

International application No

PCT/FR2009/050129

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2006/239503 A1 (PETROVIC RADE [US] ET AL) 26 October 2006 (2006-10-26) paragraph [0178] - paragraph [0180] -----	5
A	JOLY A ET AL: "Content-Based Video Copy Detection in Large Databases: A Local Fingerprints Statistical Similarity Search Approach" IMAGE PROCESSING, 2005. ICIP 2005. IEEE INTERNATIONAL CONFERENCE ON GENOVA, ITALY 11-14 SEPT. 2005, PISCATAWAY, NJ, USA, IEEE, vol. 1, 11 September 2005 (2005-09-11), pages 505-508, XP010850797 ISBN: 978-0-7803-9134-5 page 506, left-hand column, line 34 - right-hand column, line 40 -----	
A	JULIEN LAW-TO ET AL: "Video Copy Detection on the Internet: The Challenges of Copyright and Multiplicity" MULTIMEDIA AND EXPO, 2007 IEEE INTERNATIONAL CONFERENCE ON, IEEE, PI, 1 July 2007 (2007-07-01), pages 2082-2085, XP031124067 ISBN: 978-1-4244-1016-3 abstract page 293, left-hand column, line 19 - line 28 page 295, right-hand column, line 26 - line 45 page 299, left-hand column, line 47 - page 300, left-hand column, line 6 -----	1,12

Information on patent family members

PCT/FR2009/050129

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2006239503 A1	26-10-2006	JP 2008539671 T	13-11-2008
		US 2008310673 A1	18-12-2008

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°

PCT/FR2009/050129

A. CLASSEMENT DE L'OBJET DE LA DEMANDE

INV. G06F17/30

Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)

G06F

Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si cela est réalisable, termes de recherche utilisés)

EPO-Internal, WPI Data, INSPEC, IBM-TDB, COMPENDEX

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	ALEXIS JOLY ET AL: "Content-Based Copy Retrieval Using Distortion-Based Probabilistic Similarity Search" IEEE TRANSACTIONS ON MULTIMEDIA, IEEE SERVICE CENTER, PISCATAWAY, NJ, US, vol. 9, no. 2, 1 février 2007 (2007-02-01), pages 293-306, XP011157474 ISSN: 1520-9210 abrégé page 293, colonne de gauche, ligne 19 - ligne 28 page 295, colonne de droite, ligne 26 - ligne 45 page 299, colonne de gauche, ligne 47 - page 300, colonne de gauche, ligne 6 ----- -/--	1-12

☒ Voir la suite du cadre C pour la fin de la liste des documents

☒ Les documents de familles de brevets sont indiqués en annexe

* Catégories spéciales de documents cités:

"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent

"E" document antérieur, mais publié à la date de dépôt international ou après cette date

"L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée)

"O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens

"P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée

"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention

"X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément

"Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier

"&" document qui fait partie de la même famille de brevets

Date à laquelle la recherche internationale a été effectivement achevée

9 juin 2009

Date d'expédition du présent rapport de recherche internationale

17/06/2009

Nom et adresse postale de l'administration chargée de la recherche internationale

Office Européen des Brevets, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Fonctionnaire autorisé

Deane, Inigo

C(suite). DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
A	US 2006/239503 A1 (PETROVIC RADE [US] ET AL) 26 octobre 2006 (2006-10-26) alinéa [0178] - alinéa [0180] -----	5
A	JOLY A ET AL: "Content-Based Video Copy Detection in Large Databases: A Local Fingerprints Statistical Similarity Search Approach" IMAGE PROCESSING, 2005. ICIP 2005. IEEE INTERNATIONAL CONFERENCE ON GENOVA, ITALY 11-14 SEPT. 2005, PISCATAWAY, NJ, USA, IEEE, vol. 1, 11 septembre 2005 (2005-09-11), pages 505-508, XP010850797 ISBN: 978-0-7803-9134-5 page 506, colonne de gauche, ligne 34 - colonne de droite, ligne 40 -----	
A	JULIEN LAW-TO ET AL: "Video Copy Detection on the Internet: The Challenges of Copyright and Multiplicity" MULTIMEDIA AND EXPO, 2007 IEEE INTERNATIONAL CONFERENCE ON, IEEE, PI, 1 juillet 2007 (2007-07-01), pages 2082-2085, XP031124067 ISBN: 978-1-4244-1016-3 abrégé page 293, colonne de gauche, ligne 19 - ligne 28 page 295, colonne de droite, ligne 26 - ligne 45 page 299, colonne de gauche, ligne 47 - page 300, colonne de gauche, ligne 6 -----	1,12

RAPPORT DE RECHERCHE INTERNATIONALE

Renseignements relatifs aux membres de familles de brevets

Demande internationale n°

PCT/FR2009/050129

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
US 2006239503 A1	26-10-2006	JP 2008539671 T	13-11-2008
		US 2008310673 A1	18-12-2008