(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

COURBEVOIE

(11) **N° de publication :** **3 048 295**

(à n'utiliser que pour les commandes de reproduction)

(21) **N° d'enregistrement national :** **16 51591**

(51) Int Cl[8] : ***G 06 F 17/30*** *(2017.01), G 06 K 9/62, 9/00*

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt :** 26.02.16.

(30) **Priorité :**

(43) **Date de mise à la disposition du public de la demande :** 01.09.17 Bulletin 17/35.

(56) **Liste des documents cités dans le rapport de recherche préliminaire :** *Se reporter à la fin du présent fascicule*

(60) **Références à d'autres documents nationaux apparentés :**

( ) **Demande(s) d'extension :**

(71) **Demandeur(s) :** *COMMISSARIAT A L'ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES* — FR et *CONSERVATOIRE NATIONAL DES ARTS ET METIERS Etablissement public* — FR.

(72) **Inventeur(s) :** TRAN THI QUYNH NHI, LE BORGNE HERVE et CRUCIANU MICHEL.

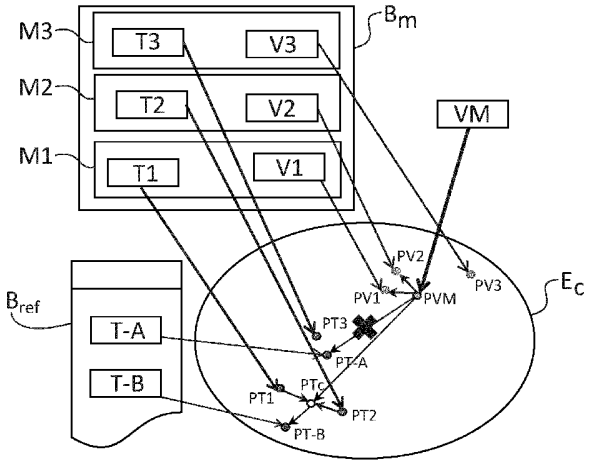
(73) **Titulaire(s) :** COMMISSARIAT A L'ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES, CONSERVATOIRE NATIONAL DES ARTS ET METIERS Etablissement public.

(74) **Mandataire(s) :** BREVALEX Société à responsabilité limitée.

(54) PROCEDE DE DESCRIPTION DE DOCUMENTS MULTIMEDIA PAR TRADUCTION INTER-MODALITES, SYSTEME ET PROGRAMME D'ORDINATEUR ASSOCIES.

(57) L'invention porte sur la génération d'une description multimodale de documents. Elle exploite un ensemble de documents multimédia (M1, M2, M3) disposant chacun d'une description (V1, V2, V3; T1, T2, T3) selon une première modalité et selon une seconde modalité, et un espace commun de représentation (Ec) de descriptions selon chacune des modalités. Les étapes suivantes sont réalisées:

- pour chacun des documents multimédia (M1, M2, M3), projection de chacune des descriptions (V1, V2, V3) du document dans ledit espace de manière à disposer d'un premier (PV1, PV2, PV3) et d'un second point (PT1, PT2, PT3);
- projection d'une description (VM) d'un document requête selon la première modalité dans ledit espace, de manière à disposer d'un point requête (PVM);
- recherche, parmi les premiers points, des k plus proches voisins (PV1, PV2) du point requête (PVM);
- détermination d'une description du document requête selon la seconde modalité à partir des k seconds points (PT1, PT2) associés aux k premiers points identifiés.



FR 3 048 295 - A1

PROCÉDÉ DE DESCRIPTION DE DOCUMENTS MULTIMEDIA PAR TRADUCTION INTER-MODALITÉS, SYSTÈME ET PROGRAMME D'ORDINATEUR ASSOCIÉS

DESCRIPTION

DOMAINE TECHNIQUE

Le domaine de l'invention est celui de la description de documents multimédias en vue d'une utilisation pour la recherche d'information par le contenu ou pour la classification supervisée de contenus multimédias. L'invention s'intéresse plus particulièrement à rapprocher un contenu décrit par une modalité (par exemple un contenu purement visuel) d'un contenu décrit par une autre modalité (par exemple un contenu purement textuel).

ÉTAT DE LA TECHNIQUE ANTÉRIEURE

La recherche d'information par le contenu et la classification supervisée de documents nécessitent une étape de description du contenu des documents.

Un document multimédia est constitué d'au moins deux médias élémentaires par exemple choisis parmi des images, des sons, des signaux vidéo, et des textes. Du fait de la nature hétérogène des modalités définissant un document multimédia, sa description s'avère délicate.

On vient ainsi généralement procéder séparément à la description des différentes modalités. Prenant l'exemple d'une image (contenu visuel) associé à un contenu textuel (des mots clés par exemple), cette étape de description transforme séparément le contenu textuel (des mots) et le contenu visuel (des pixels) en un vecteur de caractéristiques (*features* en anglais) de dimension généralement fixe. Dans le cas de la recherche par le contenu, ces vecteurs sont indexés dans une base de référence. Dans le cas de la classification supervisée, ces vecteurs sont utilisés pour apprendre un modèle au moyen d'un algorithme d'apprentissage.

Un premier problème est que le contenu textuel n'est pas décrit par le même type de vecteur que le contenu visuel. En particulier, ces vecteurs ne sont

généralement pas de même dimension. Et même s'ils sont par hasard de même dimension, ces vecteurs n'engendrent pas le même sous-espace. Dans tous les cas, ils ne peuvent pas être comparés directement. Ils ne peuvent donc pas être indexés de la même manière pour la recherche par le contenu, et ne peuvent pas servir à apprendre le même modèle pour la classification supervisée.

Une solution à ce problème est de considérer un espace commun de représentation entre les deux modalités de contenu. Cet espace peut faire l'objet d'un apprentissage, par exemple au moyen d'une Analyse canonique des corrélations (CCA pour *Canonical Correlation Analysis*) ou sa version non linéaire à noyau (KCCA pour *Kernel Canonical Correlation Analysis*).

Cette solution est par exemple décrite dans l'article de T. Q. N. Tran, H. Le Borgne, et M. Crucianu intitulé « Combining Generic and Specific Information for Cross-modal Retrieval », In Proc. *ACM International Conference on Multimedia Retrieval (ICMR 2015)*, Shanghai, China, June 23-26, 2015.

Comme représenté sur la figure 1, un document multimédia bi-modal M comporte par exemple un contenu textuel T et un contenu visuel V. Le contenu textuel T est soumis à une extraction de caractéristiques textuelles Ext_T qui fournit un vecteur de caractéristiques textuelles Vt. Le contenu visuel est soumis à une extraction de caractéristiques visuelles Ext_V qui fournit un vecteur de caractéristiques textuelles Vv. Chacun de ces vecteurs Vt, Vv se projettent en un point PT, PV dans l'espace commun de représentation Ec.

Une fois un tel espace appris, un document au contenu purement textuel T1 est projeté en un point PT1, un document au contenu purement visuel V2 est projeté en un point PV2. Les points PT1 et PV2 sont dans le même espace et peuvent donc être comparés directement. En particulier, il s'avère ainsi possible, partant d'une description uni-modale d'un document, par exemple une description d'un contenu textuel T1, d'identifier une description associée selon une autre modalité, par exemple une description d'un contenu visuel V2. La projection de la description du contenu visuel V2 correspond par exemple au plus proche voisin de la projection de la description du contenu textuel T1 dans l'espace commun de représentation.

On peut ainsi procéder à une illustration automatique de textes et, symétriquement, à une annotation automatique d'images. Il est par ailleurs possible de réaliser un apprentissage de classificateurs à partir de documents représentés selon une modalité (par exemple visuelle) et s'appliquant à des documents (par exemple des textes) ne présentant pas cette modalité. Et ces classificateurs peuvent pareillement être évalués au moyen de documents ne présentant pas une modalité (par forcément la même que pour l'apprentissage, par exemple textuelle).

Dans un cas idéal représenté en traits pointillés sur la figure 1, les contenus T et V étant alignés (issus du même document), les projections dans l'espace commun de représentation Ec devraient être confondues, ou en tout état de cause plus proches l'une de l'autre que de n'importe quel autre point dans l'espace commun de représentation. En réalité, comme représenté en traits pleins sur la figure 1, l'apprentissage de l'espace commun est imparfait et les points PT et PV sont distants.

La comparaison dans l'espace commun de ces projections est donc approximative, limitant de facto la qualité des rapprochements que l'on souhaite effectuer entre la description d'une modalité (par exemple visuelle) d'un document et celle d'une autre modalité (par exemple textuelle) du même document ou d'un autre document.

EXPOSÉ DE L'INVENTION

L'invention vise à remédier aux inconvénients résultants des imperfections de l'espace commun de représentation. Elle propose pour ce faire un procédé de génération, dans un dispositif informatique, d'une description multimodale d'un document requête à partir d'une description du document requête selon une première modalité. Le procédé exploite une base multimodale constituée d'un ensemble de documents multimédia disposant chacun d'une description selon la première modalité et d'une description selon une seconde modalité, et un espace commun de représentation à la fois de descriptions selon la première modalité et de descriptions selon la seconde modalité.

Le procédé comprend les étapes suivantes :

- pour chaque document multimédia de la base multimodale, projection de la description du document selon la première modalité dans l'espace commun de représentation de manière à disposer d'un premier point, et projection de la description du document selon la seconde modalité dans l'espace commun de représentation, de manière à disposer d'un second point associé au premier point ;

- projection de la description du document requête selon la première modalité dans l'espace commun de représentation, de manière à disposer d'un point requête.

Le procédé est caractérisé en ce qu'il comprend les étapes suivantes :

- recherche, parmi les premiers points, des k plus proches voisins du point requête dans l'espace commun de représentation, de manière à identifier k premiers points, k étant un entier supérieur ou égal à 1 ;

- identification, parmi les seconds points, des k seconds points associés au k premiers points identifiés ;

- détermination d'une description du document requête selon la seconde modalité à partir des k seconds points associés aux k premiers points identifiés.

Certains aspects préférés mais non limitatifs de ce procédé sont les suivants :

- la détermination d'une description du document requête selon la seconde modalité comprend le calcul d'une moyenne pondérée des k seconds points associés au k premiers points identifiés, de manière à fournir un point cible ;

- l'espace commun de représentation est divisé en une pluralité de régions, chaque région étant représentée par un mot de code de quantification, et le point requête et les k seconds points associés au k premiers points identifiés sont codés conformément à un dictionnaire formé par les mots de code de quantification ;

- le codage d'un point conformément au dictionnaire correspond aux différences par composante du point avec les mots de code les plus proches dudit point dans l'espace commun de représentation ;

- la détermination d'une description du document requête selon la seconde modalité comprend le calcul d'une moyenne pondérée des codages des k seconds points associés au k premiers points identifiés ;

- le poids associé à un second point dans le calcul de la moyenne pondérée est fonction de la distance entre le point requête et le premier point associé au second point sur l'espace commun de représentation.

L'invention vise également un produit programme d'ordinateur comprenant des instructions de code de programme permettant d'effectuer les étapes du procédé lorsque ledit programme est exécuté sur un ordinateur. Elle s'étend en outre à un système configuré de manière à permettre d'effectuer les étapes de ce procédé.

BRÈVE DESCRIPTION DES DESSINS

D'autres aspects, buts, avantages et caractéristiques de l'invention apparaîtront mieux à la lecture de la description détaillée suivante de formes de réalisation préférées de celle-ci, donnée à titre d'exemple non limitatif, et faite en référence aux dessins annexés sur lesquels, outre la figure 1 déjà discutée précédemment :

- la figure 2 est un schéma illustrant les différentes étapes du procédé selon l'invention ;
- la figure 3 est un schéma illustrant une quantification de l'espace commun de représentation pouvant être mise en œuvre dans un mode de réalisation possible de l'invention.

EXPOSÉ DÉTAILLÉ DE MODES DE RÉALISATION PARTICULIERS

En référence à la figure 2, l'invention porte sur un procédé de génération, dans un dispositif informatique, d'une description multimodale d'un document, appelé document requête, à partir d'une description du document selon une première modalité, par exemple une modalité visuelle VM. On prendra dans ce qui suit la génération d'une description bi-modale par souci de simplicité, sans que l'invention n'y soit limitée.

Le document requête peut ne pas disposer d'une description selon une seconde modalité (le document est par exemple mono-média), ou bien on peut ignorer

une description selon une seconde modalité du document (ici multimédia) pour en déterminer une conformément au procédé selon l'invention.

Dans le cadre de l'invention, on entend par description d'une modalité un vecteur de caractéristiques représentatives de ladite modalité dans le document. Prenant l'exemple d'un document disposant d'un contenu textuel et visuel (image), un vecteur de caractéristiques x^T est extrait de son contenu textuel et un autre vecteur de caractéristiques x^I est extrait de son contenu visuel.

Le procédé exploite un espace commun de représentation Ec à la fois de descriptions selon la première modalité et de descriptions selon une seconde modalité.

Cet espace peut être déterminé à partir d'une analyse canonique des corrélations de type KCCA exploitant une base d'apprentissage constitué d'un ensemble de documents bi-modaux disposant chacun d'une description selon la première modalité et d'une description selon une seconde modalité. Cette base d'apprentissage est composée d'un ensemble de N couples de vecteurs de caractéristiques $(\mathrm{x}_\mathrm{i}^I, \mathrm{x}_\mathrm{i}^T), \mathrm{i} = 1 \ldots \mathrm{N}$.

Dans l'espace commun de représentation, chaque document, ici assimilé à un couple de vecteurs de caractéristiques (x^I, x^T), est représenté par deux points : p^I qui correspond à la projection de x^I, et p^T qui correspond à la projection de x^T.

A titre d'exemple purement illustratif, les vecteurs de caractéristiques textuelles x^T sont de dimension 300, les vecteurs de caractéristiques visuelles x^I sont de dimension 4096 et l'espace commun de représentation, déterminé au moyen d'une base d'apprentissage de plus de 5000 documents, est de dimension d = 150.

Le procédé exploite par ailleurs une base multimodale Bm constituée d'un ensemble de documents multimédia M1, M2, M3 disposant chacun d'une description V1, V2, V3 selon la première modalité et d'une description T1, T2, T3 selon une seconde modalité. Cette base permet de fournir un ensemble de points pivots bi-modaux à même de refléter les imperfections de l'espace commun de représentation. En pratique, cette base bimodale peut correspondre à la base d'apprentissage, sans pour autant que cela ne soit nécessaire.

Dans le cadre de l'invention, les descriptions bi-modales des documents de la base multimodale Bm sont projetées dans l'espace commun de représentation Ec.

Le procédé comprend ainsi une étape consistant, pour chaque document multimédia M1, M2, M3 de la base multimodale Bm, à réaliser la projection de la description V1, V2, V3 du document selon la première modalité dans l'espace commun de représentation de manière à disposer d'un premier point PV1, PV2, PV3, et à réaliser la projection de la description T1, T2, T3 du document selon la seconde modalité dans l'espace commun de représentation, de manière à disposer d'un second point PT1, PT2, PT3 associé au premier point.

On peut noter $\mathcal{A}$ l'ensemble des couples de premier point q_i^I et second point q_i^T résultant de cette étape de projection des descriptions des documents de la base bimodale : $\mathcal{A} = \{(q_i^I, q_i^T)\},\ q_i^I \in \mathcal{A}^I,\ q_i^T \in \mathcal{A}^T,\ i = 1 \dots m.$

Le procédé selon l'invention comprend par ailleurs une étape consistant à réaliser la projection de la description VM (également notée r^I) du document requête selon la première modalité dans l'espace commun de représentation Ec, de manière à disposer d'un point requête PVM. L'objectif est alors de déterminer, à partir du point requête PVM, un ou plusieurs points cible PTc de l'espace commun de représentation permettant de compléter la description (notée r^T) de l'autre modalité du document requête.

Une approche naïve pourrait consister à identifier pour points cibles les k plus proches voisins de PVM parmi les points résultant d'une projection d'une description selon la seconde modalité (cet ensemble de points est noté $NN_{\mathcal{A}^T}^k(r^I)$).

Sur l'exemple de la figure 2, cette approche conduirait, partant de PVM, à identifier le point PT-A renvoyant à un contenu textuel T-A stocké dans une base de référence Bref qui est a priori différente de la base d'apprentissage ayant permis de déterminer l'espace commun de représentation.

L'invention propose une autre approche selon laquelle on vient rechercher les plus proches voisins du point requête r^I dans l'espace commun de représentation, non pas parmi les seconds points, mais parmi les premiers points $q_i^I \in \mathcal{A}^I$ (points de même modalité). On vient ainsi identifier k premiers points, k étant un entier supérieur ou égal à 1. Il s'agit ainsi des points q_i^Itels que $q_i^I \in NN_{\mathcal{A}^I}^k(r^I)$. L'entier k est

typiquement supérieur à 10. Il est de préférence supérieur à 20. La métrique d'identification des plus proches voisins est par exemple une distance euclidienne.

Dans l'exemple de la figure 2, cette étape permet d'identifier les deux premiers voisins du point requête PVM dans la même modalité, à savoir PV1 et PV2.

Puis le procédé comprend une étape consistant à identifier, parmi les seconds points, les k seconds points associés au k premiers points identifiés. Il s'agit ainsi de l'ensemble $\mathcal{M}_c(r^I) = \{q_i^T\}$ tels que $q_i^I \in NN_{\mathcal{A}^I}^k(r^I)$ et $(q_i^I, q_i^T) \in \mathcal{A}$ (cette dernière condition implique que q_i^I et q_i^T sont les projections des deux descriptions du même document multimédia de la base multimodale).

Dans l'exemple de la figure 2, cette étape permet d'identifier les points PT1 et PT2 qui sont les points complémentaires (i.e. ils correspondent à l'autre modalité) des premiers voisins du point requête PVM dans la même modalité, à savoir PV1 et PV2.

Le procédé comprend ensuite une étape consistant à déterminer une description r^T du document requête selon la seconde modalité à partir des k seconds points associés aux k premiers points identifiés.

Dans un premier mode de réalisation possible, la détermination de la description du document requête selon la seconde modalité peut comprendre le calcul d'une moyenne pondérée des k seconds points associés au k premiers points identifiés, de manière à fournir un point cible PTc.

Prenant l'exemple d'une simple moyenne, le point cible est $\frac{1}{k}\sum_{q_j^T \epsilon \mathcal{M}_c(r^I)} q_j^T$.

Dans une variante de réalisation, le poids associé à un second point $q_j^T \epsilon \mathcal{M}_c(r^I)$ dans le calcul de la moyenne pondérée est fonction de la distance entre le point requête r^I et le premier point $q_j^I \in NN_{\mathcal{A}^I}^k(r^I)$ (l'un des plus proches voisins de r^I dans la même modalité) associé au second point sur l'espace commun de représentation. On adopte typiquement une fonction décroissance de la distance qui exprime une similarité entre le point requête et le premier point, par exemple une exponentielle décroissante du type exp (-α*de) où de est la distance euclidienne entre le point requête et le premier point, et α une constante.

Une représentation « complétée » du document requête dans l'espace commun de représentation peut alors correspondre à la moyenne ou à la somme du point requête PVM et du point cible PTc. En variante, la représentation « complétée » peut correspondre à la concaténation du point requête PVM et du point cible PTc.

Une fois le point cible PTc connu, le procédé peut comprendre l'identification d'un ou plusieurs documents disposant d'une description, par exemple selon la seconde modalité, dont la projection dans l'espace commun de représentation est la plus proche du point cible. Ces documents sont typiquement stockés dans la base de référence Bref. Selon l'exemple de la figure 2, cette étape permet d'identifier le contenu textuel T-B dont la projection PT-B est proche du point cible PTc. La base de référence peut être une base bi-modale texte-image ou une base mono-modale texte ou visuelle. Prenant l'exemple d'une requête texte et d'une base de référence mono-modale texte, l'invention permet de prendre en compte un aspect multimédia. Par exemple, la requête « hawai » et le texte « floride » peuvent être rapprochés parce que des images (de la base multimodale Bm) taguées par ces mots (ou des mots proches de ces mots) se ressemblent.

Dans un second mode de réalisation possible illustré par la figure 3, l'espace commun de représentation Ec est divisé en une pluralité de régions, chaque région étant représentée par un mot de code de quantification C1-C8. Les différents points (notamment le point requête et les k seconds points associés aux k premiers points identifiés) sont codés conformément à un dictionnaire formé par les mots de code de quantification. Cette division de l'espace commun de représentation peut être réalisée au moyen d'un algorithme de partitionnement en K-moyennes qui exploite l'ensemble des projections de la base d'apprentissage, provenant à la fois de descriptions selon la première modalité et de projections selon la seconde modalité. Le partitionnement fournit trois types de mots de code (qui sont les centres des partitions). Certains sont représentatifs de la première modalité seulement, d'autres sont représentatifs de la seconde modalité seulement, tandis que certains contiennent des projections de descriptions à la fois selon la première et selon la seconde modalité.

Le codage peut être réalisé au moyen de techniques connues de l'homme du métier, telles celles passées en revue dans l'article Yongzhen Huang, Zifeng Wu, Liang Wang, Tieniu Tan, "Feature Coding in Image Classification: A Comprehensive Study," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 493-506, March, 2014.

Le codage d'un point conformément au dictionnaire peut notamment correspondre aux différences par composante (gradient) du point avec les mots de code les plus proches dudit point dans l'espace commun de représentation. Dans l'exemple de la figure 3, le point PT a pour mots de code les plus proches les mots de code C2, C7 et C8, tandis que le point PV a pour mots de code les plus proches les mots de code C6, C5 et C7.

Considérant un document multimédia assimilé à un couple de vecteurs de caractéristiques (x^I, x^T), ce dernier est projeté dans l'espace commun de représentation Ec de dimension d aux points p^I et p^T. Chacun de ces points est alors encodé en v^I et v^T par ces différences aux n plus proches mots de code (par exemple au sens d'une distance euclidienne) :

$$v^I = [v_1^I, \ldots., v_i^I, \ldots, v_l^I] \; ; \; v^T = [v_1^T, \ldots., v_i^T, \ldots, v_l^T]$$

Avec v_i^I et v_i^T des vecteurs de dimension d tels que :

- $v_i^I = (p^I - c_i)\, 1_{NN^n(p^I)}(c_i)$
- $v_i^T = (p^T - c_i)\, 1_{NN^n(p^T)}(c_i)$

où 1_A est la fonction indicatrice telle que $1_A(x) = 1$ si $x \in A$ et $1_A(x) = 0$ sinon,
et où l correspond à la taille du dictionnaire et $NN^n(p)$ est l'ensemble des n plus proches voisins du point p.

A titre d'exemple illustratif, on peut retenir l=16 et n=5.

Dans le cadre de ce second mode de réalisation, la détermination d'une description du document requête selon la seconde modalité est réalisée à partir du codage conformément au dictionnaire de chacun des k seconds points associés aux k premiers points identifiés. Comme pour le premier mode de réalisation, on peut réaliser une moyenne pondérée de ces codages, de manière à fournir une description codée d'un point cible PTc.

Prenant l'exemple d'une simple moyenne, et d'une seconde modalité de type T (texte), on vient ainsi déterminer $v^T = [v_1^T, \ldots., v_i^T, \ldots, v_l^T]$, avec $v_i^T = \frac{1}{k}\sum_{q_j^T \epsilon \mathcal{M}_c(r^I)} (q_j^T - c_i)\ 1_{NN^n\left(q_j^T\right)}(c_i)$.

La pondération du codage d'un second point peut notamment prendre en compte la distance entre le point requête PVM et le premier point PV1, PV2 associé au second point sur l'espace commun de représentation.

Une représentation « complétée » du document requête dans l'espace commun de représentation peut alors correspondre à la moyenne ou à la somme de la description (codée selon le dictionnaire) du point requête PVM et du point cible PTc. Cette représentation complétée d'un document initialement décrit par sa modalité visuelle peut ainsi s'exprimer selon :

$$v = [v_1, \ldots., v_i, \ldots, v_l]$$

avec $v_i = (p^I - c_i)\ 1_{NN^n(p^I)}(c_i) + \frac{1}{k}\sum_{q_j^T \epsilon \mathcal{M}_c(r^I)} (q_j^T - c_i)\ 1_{NN^n\left(q_j^T\right)}(c_i)$.

L'invention n'est pas limitée au procédé tel que précédemment, mais s'étend également à un produit programme d'ordinateur comprenant des instructions de code de programme permettant d'effectuer les étapes du procédé tel que précédemment décrit lorsque ledit programme est exécuté sur un ordinateur.

Et l'invention s'étend également à un système pour la génération d'une description multimodale d'un document requête disposant d'une description VM selon une première modalité. Le système comprend une base de données dans laquelle sont stockés un modèle de l'espace commun de représentation Ec et les documents multimédia M1, M2, M3 de la base multimodale Bm. Il comprend par ailleurs un processeur configuré :

- pour réaliser, pour chaque document multimédia M1, M2, M3 de la base multimodale Bm, la projection de la description V1, V2, V3 du document selon la première modalité dans l'espace commun de représentation de manière à disposer d'un premier point PV1, PV2, PV3, et la projection de la description T1, T2, T3 du document selon la seconde modalité dans l'espace commun de représentation, de manière à disposer d'un second point PT1, PT2, PT3 associé au premier point ;

- pour réaliser la projection de la description VM du document requête selon la première modalité dans l'espace commun de représentation, de manière à disposer d'un point requête PVM;
- pour rechercher, parmi les premiers points, les k plus proches voisins du point requête PVM dans l'espace commun de représentation, de manière à identifier k premiers points PV1, PV2, k étant un entier supérieur ou égal à 1 ;
- identifier, parmi les seconds points, les k seconds points PT1, PT2 associés au k premiers points identifiés ;
- déterminer une description du document requête selon la seconde modalité à partir des k seconds points PT1, PT2 associés au k premiers points identifiés.

Comme illustré dans le tableau ci-dessous, l'invention permet d'améliorer les performances dans certains cas par comparaison aux techniques existantes et rend possible la résolution de certains problèmes de reconnaissance.

		Base de référence		
		Bi-modale texte image	Mono-modale texte	Mono-modale visuel
Requête	Bi-modale texte image	Idem ou mieux que l'existant	Améliore l'existant	Améliore l'existant
	Mono-modale texte	Améliore l'existant	Idem ou mieux que l'existant	Rend possible
	Mono-modale visuel	Améliore l'existant	Rend possible	Idem ou mieux que l'existant

REVENDICATIONS

1. Procédé de génération, dans un dispositif informatique, d'une description multimodale d'un document requête à partir d'une description (VM) du document requête selon une première modalité, le procédé exploitant une base multimodale (Bm) constituée d'un ensemble de documents multimédia (M1, M2, M3) disposant chacun d'une description (V1, V2, V3) selon la première modalité et d'une description (T1, T2, T3) selon une seconde modalité, et un espace commun de représentation (Ec) à la fois de descriptions selon la première modalité et de descriptions selon la seconde modalité,

le procédé comprenant les étapes suivantes :

- pour chaque document multimédia (M1, M2, M3) de la base multimodale (Bm), projection de la description (V1, V2, V3) du document selon la première modalité dans l'espace commun de représentation de manière à disposer d'un premier point (PV1, PV2, PV3), et projection de la description (T1, T2, T3) du document selon la seconde modalité dans l'espace commun de représentation, de manière à disposer d'un second point (PT1, PT2, PT3) associé au premier point ;
- projection de la description (VM) du document requête selon la première modalité dans l'espace commun de représentation, de manière à disposer d'un point requête (PVM);

le procédé étant caractérisé en ce qu'il comprend les étapes suivantes :

- recherche, parmi les premiers points, des k plus proches voisins du point requête (PVM) dans l'espace commun de représentation, de manière à identifier k premiers points (PV1, PV2), k étant un entier supérieur ou égal à 1 ;
- identification, parmi les seconds points, des k seconds points (PT1, PT2) associés au k premiers points identifiés ;
- détermination d'une description du document requête selon la seconde modalité à partir des k seconds points (PT1, PT2) associés aux k premiers points identifiés.

2. Procédé selon la revendication 1, dans lequel la détermination d'une description du document requête selon la seconde modalité comprend le calcul d'une moyenne

pondérée des k seconds points associés au k premiers points identifiés, de manière à fournir un point cible (PTc).

3. Procédé selon la revendication 1, dans lequel l'espace commun de représentation est divisé en une pluralité de régions, chaque région étant représentée par un mot de code de quantification, et dans lequel le point requête et les k seconds points associés au k premiers points identifiés sont codés conformément à un dictionnaire formé par les mots de code de quantification

4. Procédé selon la revendication 3, dans lequel le codage d'un point conformément au dictionnaire correspond aux différences par composante du point avec les mots de code les plus proches dudit point dans l'espace commun de représentation.

5. Procédé selon l'une des revendications 3 et 4, dans lequel la détermination d'une description du document requête selon la seconde modalité comprend le calcul d'une moyenne pondérée des codages des k seconds points associés au k premiers points identifiés.

6. Procédé selon l'une des revendications 2 et 5, dans lequel le poids associé à un second point dans le calcul de la moyenne pondérée est fonction de la distance entre le point requête et le premier point associé au second point sur l'espace commun de représentation.

7. Procédé selon l'une des revendications 1 à 6, comprenant une étape préalable de détermination de l'espace commun de représentation par apprentissage au moyen de descriptions de documents selon la première et la seconde modalité.

8. Produit programme d'ordinateur comprenant des instructions de code de programme permettant d'effectuer les étapes du procédé selon l'une quelconque des revendications 1 à 7 lorsque ledit programme est exécuté sur un ordinateur.

9. Système pour la génération d'une description multimodale d'un document requête disposant d'une description (VM) selon une première modalité, comprenant :

une base de données dans laquelle sont stockés un modèle d'une espace commun de représentation (Ec) à la fois de descriptions selon la première modalité et de descriptions selon la seconde modalité, et un ensemble de documents multimédia (M1, M2, M3) disposant chacun d'une description (V1, V2, V3) selon la première modalité et d'une description (T1, T2, T3) selon une seconde modalité ; et

un processeur configuré :

pour réaliser, pour chaque document multimédia (M1, M2, M3) dudit ensemble, la projection de la description (V1, V2, V3) du document selon la première modalité dans l'espace commun de représentation de manière à disposer d'un premier point (PV1, PV2, PV3), et la projection de la description (T1, T2, T3) du document selon la seconde modalité dans l'espace commun de représentation, de manière à disposer d'un second point (PT1, PT2, PT3) associé au premier point ;

pour réaliser la projection de la description (VM) du document requête selon la première modalité dans l'espace commun de représentation, de manière à disposer d'un point requête (PVM);

le système étant caractérisé en ce que le processeur est en outre configuré pour mettre en œuvre les étapes suivantes :

- recherche, parmi les premiers points, des k plus proches voisins du point requête (PVM) dans l'espace commun de représentation, de manière à identifier k premiers points (PV1, PV2), k étant un entier supérieur ou égal à 1 ;

- identification, parmi les seconds points, des k seconds points (PT1, PT2) associés au k premiers points identifiés ;

- détermination d'une description du document requête selon la seconde modalité à partir des k seconds points (PT1, PT2) associés aux k premiers points identifiés.

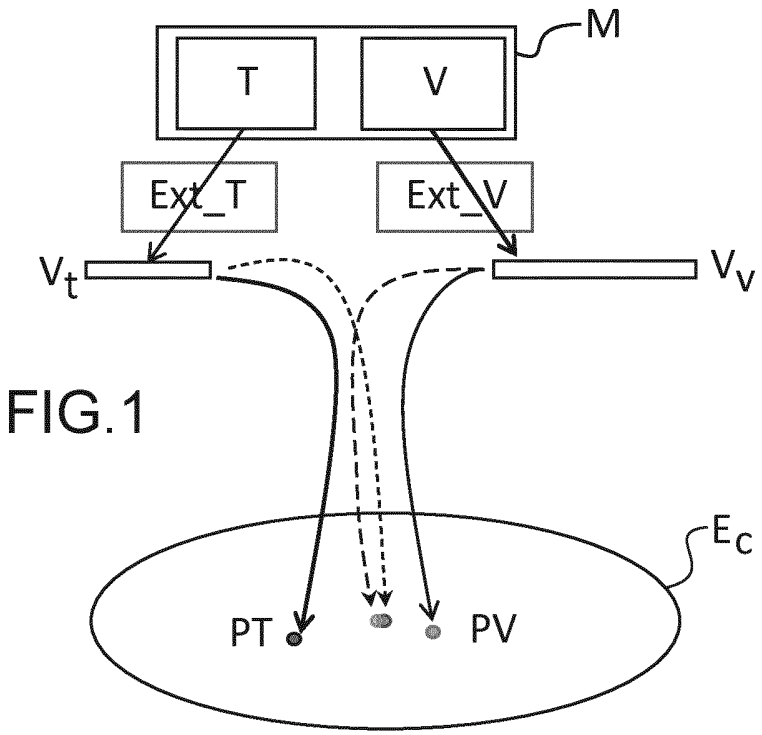

M
T
V
Ext_T
Ext_V
V_t
V_v
E_c
PT
PV

FIG.1

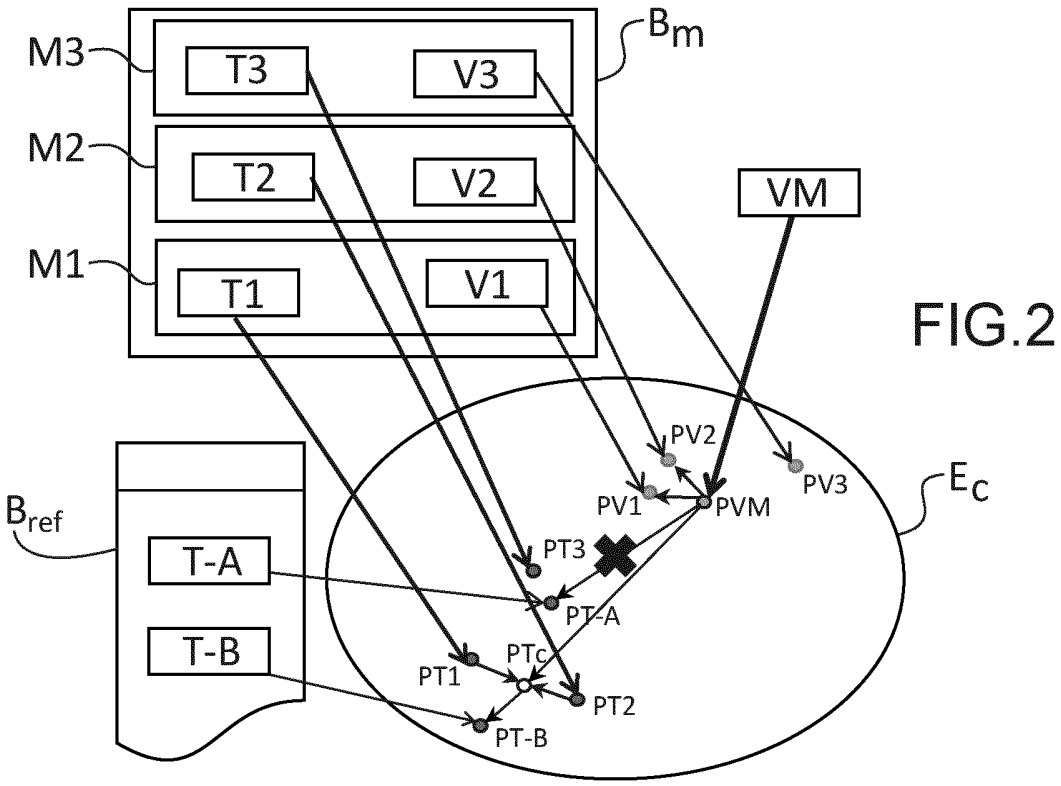

M3
M2
M1
T3
V3
T2
V2
T1
V1
B_m
VM
B_{ref}
T-A
T-B
PV2
PV1
PVM
PV3
E_c
PT3
PT-A
PTc
PT1
PT2
PT-B

FIG.2

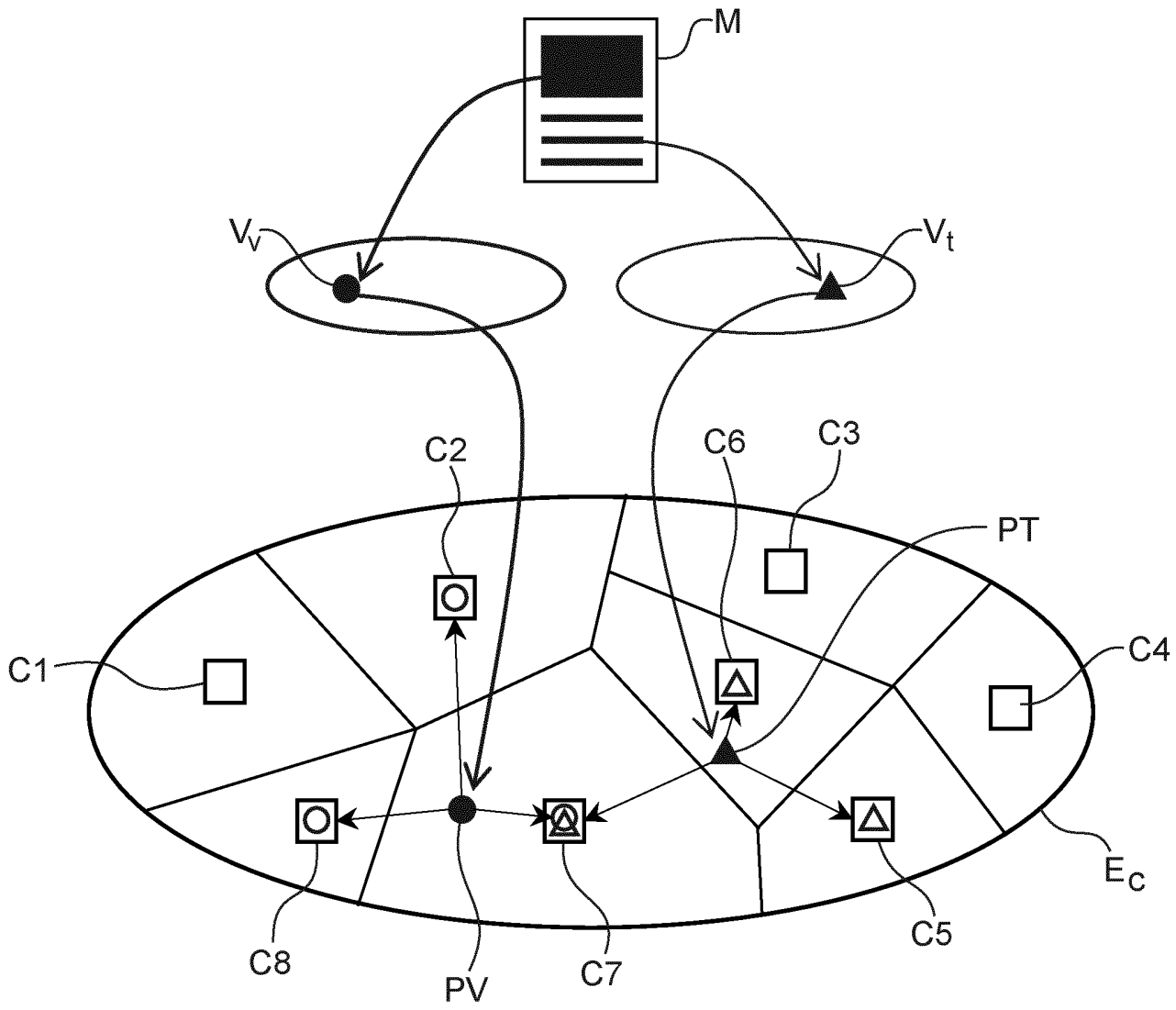


FIG.3

RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

N° d'enregistrement national

FA 823138
FR 1651591

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X,D	THI QUYNH NHI TRAN ET AL: "Combining Generic and Specific Information for Cross-modal Retrieval", PROCEEDINGS OF THE 5TH ACM ON INTERNATIONAL CONFERENCE ON MULTIMEDIA RETRIEVAL, ICMR '15, 1 janvier 2015 (2015-01-01), pages 551-554, XP055294595, New York, New York, USA DOI: 10.1145/2671188.2749348 ISBN: 978-1-4503-3274-3 * pages 551,552 * -----	1-9	G06F17/30 G06K9/62 G06K9/00
			DOMAINES TECHNIQUES RECHERCHÉS (IPC)
			G06F

Date d'achèvement de la recherche	Examinateur
10 août 2016	Michalski, Stéphane

CATÉGORIE DES DOCUMENTS CITÉS

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons
.....
& : membre de la même famille, document correspondant

EPO FORM 1503 12.99 (P04C14)