Association d'un détecteur de visages et d'un détecteur d'entités nommées pour l'annotation automatique d'images

Pierre Tirilly* — **Emmanuelle Martienne**** — **Vincent Claveau*** — **Patrick Gros*****

* *IRISA / CNRS*
Campus de Beaulieu
35042 RENNES, France
{ptirilly,vclaveau}
@irisa.fr

** *IRISA / Université*
de Rennes 2
Campus de Villejean
35043 RENNES, France
emartien@irisa.fr

*** *IRISA / INRIA*
Campus de Beaulieu
35042 RENNES, France
pgros@irisa.fr

RÉSUMÉ. *Dans cet article, nous proposons une méthode d'annotation d'images de visages dans un grand corpus réel de documents texte-images. Cette méthode s'appuie sur l'utilisation conjointe d'un détecteur de visages et d'un détecteur d'entités nommées : les images contenant des visages sont annotées par les entités nommées les plus fréquentes dans le texte accompagnant les images. Bien que basique, cette méthode donne de bons résultats. Elle constitue un premier pas vers des méthodes d'indexation textuelle des images plus élaborées et basées sur des caractéristiques haut-niveau des documents.*

ABSTRACT. *In this paper, we present a method for the annotation of faces in a large real corpus. We use a face detector and a named entities detector together: pictures containing faces can be annotated by the most frequent named entities found in the text surrounding the pictures. Although this method is quite simple, it yields good results. This is a first step towards more intelligent image annotation techniques, using high-level features of the documents.*

MOTS-CLÉS : *Annotation d'images, détection de visages, entités nommées, fossé sémantique*

KEYWORDS: *Image annotation, face detection, named entities, semantic gap*

1. Introduction

La multiplication des documents numériques pose le problème de l'accès à ces documents : comment les indexer pour accéder, à partir d'une requête en langage naturel, aux documents souhaités ? Dans le cas des images, les techniques de Recherche d'Information (RI) par le contenu ne permettent de décrire les images que par des descripteurs numériques représentant leurs caractéristiques visuelles (couleur, texture...). Il est alors difficile de rapprocher ces descripteurs des concepts exprimés par les mots-clefs de la requête. Ce problème, appelé «fossé sémantique», est caractéristique de l'indexation de documents multimédias : on souhaiterait décrire les documents par le sens qu'ils portent et non par leur contenu numérique.

Les techniques d'annotation automatique d'images par des mots-clefs tentent de résoudre ce problème, mais il n'existe pas encore de cadre général permettant de décrire sémantiquement tous types d'images. En revanche, on dispose de techniques avancées permettant d'extraire certaines informations précises des images, par exemple détecter les visages dans des images. En se plaçant dans un cadre précis, on dispose donc d'outils dédiés qui facilitent la tâche d'annotation automatique. Dans cet article[1], nous proposons une méthode s'appuyant sur des outils spécifiques (un détecteur de visages et un détecteur d'entités nommées) pour annoter les images de visages d'un grand corpus de documents texte-images. Ce travail constitue une première étape vers l'indexation textuelle d'images utilisant des caractéristiques textuelles et visuelles de haut niveau, mais répond également à un besoin réel des personnes travaillant sur de grandes bases d'images (agences de presse, photothèques...).

2. État de l'art

Parmi les travaux actuels sur l'annotation d'images de personnes, on distingue deux grandes tendances. La première concerne les annotations d'albums photographiques personnels (Kuchinsky *et al.,* 1999, Zhang *et al.,* 2004, Davis *et al.,* 2005). Dans ce contexte, il s'agit plutôt d'offrir à l'utilisateur des outils lui permettant d'annoter et de classer plus facilement ses photographies, par exemple en propageant automatiquement des mots-clefs associés manuellement. Dans ce cadre, l'utilisateur est au centre du processus d'annotation, qui est donc faiblement automatisé. La seconde tendance concerne l'annotation de documents de presse. En particulier, (Berg *et al.,* 2004) propose un travail sur l'annotation de visages dans des images de presse, et (Satoh *et al.,* 1999) dans des vidéos d'informations.

Tous ces travaux se situent dans des contextes précis ; les approches utilisées s'appuient sur certaines propriétés connues de ces contextes pour fonctionner. Par exemple, (Zhang *et al.,* 2004) s'appuie sur des indices temporels pour améliorer la reconnaissance des visages, (Berg *et al.,* 2004) utilise des patrons connus a priori pour interpréter les légendes des images, (Satoh *et al.,* 1999) utilise une approche proche de la nôtre mais exploite la structure spécifique des vidéos d'informations pour affiner les annotations. Ces systèmes manquent donc de généricité par rapport aux données qu'ils

1. Ce travail a été réalisé avec le soutien du réseau d'excellence européen MUSCLE du 6e P.C.R.D.T., de la région Bretagne et du C.N.R.S.

peuvent traiter : les méthodes développées sont souvent dédiées aux corpus utilisés et parfois peu adaptables à d'autres données.

3. Annotation à l'aide d'entités nommées

Nous proposons ici une méthode d'annotation d'images de visages dans de grands corpus de documents bimodaux, composés d'un texte et d'une ou plusieurs images. Notre approche se veut générique dans le sens où elle peut s'appliquer à n'importe quel corpus de documents texte-images et s'appuie sur des outils facilement disponibles.

3.1. *Principe de l'annotation*

L'idée est d'extraire des documents les noms de personnes contenus dans le texte et les images contenant des visages. Il est alors possible d'annoter les images retenues à l'aide des noms les plus représentés dans le document. On annote donc des images, mais pas chaque visage indépendamment. On suppose de plus que les personnes qui apparaissent sur les images sont aussi celles dont le nom apparaît de manière significative dans le texte. Cette hypothèse semble raisonnable puisque d'une manière générale les images d'un document illustrent les thèmes importants abordés dans le texte.

Pour sélectionner, parmi les noms extraits du texte, ceux qui seront retenus pour annoter les images du document, nous utilisons une méthode simple : on associe à chaque image autant de noms que l'on y détecte de visages. Les noms de personnes retenus pour l'annotation sont ceux dont le nombre d'occurences dans le texte dépasse un seuil S donné (il peut donc y avoir des images annotées avec moins de noms que de visages détectés). En faisant varier la valeur de S, on peut influer sur la quantité d'image annotées et l'exactitude des annotations : moins S est élevé, plus on retient de mots-clefs et plus on annote d'images, mais le risque que les annotations soient fausses est élevé. Inversement, plus S est élevé, moins il y aura d'images annotées, mais les annotations seront probablement plus justes. On peut donc, en faisant varier S, observer l'évolution de la qualité des annotations en fonction de la proportion d'images annotées. Enfin, en cas d'ambiguité (plus de mots-clefs retenus pour l'annotation que de visages détectés sur l'image), les noms les plus fréquents sont choisis.

3.2. *Extraction des entités nommées*

Pour effectuer la détection des entités nommées, nous utilisons le logiciel NEMESIS (Fourour, 2002). Il permet d'extraire et catégoriser certaines entités nommées (anthroponymes, toponymes...). Ses performances atteignent un rappel de 90% et une précision de 95% sur le corpus de test proposé par ses concepteurs. Dans notre cas, seuls les anthroponymes (noms de personnes) sont extraits des textes. Ils sont ensuite normalisés de manière à associer une forme unique à chaque entité nommée (par exemple réunir les formes *le président Chirac* et *Jacques Chirac* sous une forme unique *Chirac*). En revanche aucun traitement n'est appliqué aux synonymes dont tous les termes sont distincts (par exemple *Jacques Chirac* et *le président français*).

3.3. *Détection des visages*

Pour la détection des visages dans les images, nous utilisons la bibliothèque OPENCV développée par Intel (Intel, 2006). Elle combine deux détecteurs de visages, un pour les visages de face et un pour les visages de profil. Ces détecteurs sont basés sur des caractéristiques de Haar (*Haar-like features*) et des classifieurs en cascade. Cette méthode donne de bonnes performances pour la reconnaissance de visages (Lienhart *et al.,* 2003). Dans notre expérience, ce système nous permet de compter le nombre de visages présents sur chaque image.

4. Expérimentation

4.1. *Présentation du corpus*

Le corpus utilisé pour cette expérience est composé d'articles de presse téléchargés sur le site www.tv5.org entre mars et novembre 2006. Les documents sont composés d'un texte et d'une ou plusieurs images. Le texte est composé uniquement du corps de l'article, *i.e.* on n'a pas cherché à exploiter certaines parties spécifiques (titre, légendes). Les légendes peuvent ainsi nous servir de vérité terrain pour évaluer notre système. Les images n'ont subi aucun prétraitement. Le corpus original était composé de 26289 textes et 41532 images. À l'issue de la détection de visages, nous avons conservé les 25533 images repérées comme contenant un ou plusieurs visages.

4.2. *Protocole expérimental*

L'expérience a été réalisée de la manière suivante pour chaque document :

1) sélection des images contenant des visages grâce à OPENCV.

2) extraction des noms de personnes du texte avec NEMESIS.

3) annotation des images sélectionnées par les noms extraits, en fonction du seuil S et du nombre de visages dans les images.

4) évaluation de la qualité des annotations.

Nous avons répété l'expérience en faisant varier le seuil S de 1 à 15. Cela permet d'observer l'évolution du nombre d'images annotées et de la précision des annotations en fonction de la dureté du critère d'annotation S.

4.3. *Évaluation*

Pour tester la précision de nos annotations, nous ne pouvons, compte-tenu de la taille du corpus, effectuer d'évaluation manuelle. Nous comparons donc nos annotations à des annotations de référence : les légendes associées aux images. Si toutes les entités nommées utilisées pour annoter une image sont présentes dans la légende de cette image, alors nous considérons l'annotation comme juste. Cette méthode est réaliste dans la mesure où les légendes, dans notre corpus, sont très courtes : les seuls noms que l'on y trouve se réfèrent aux personnes présentes sur l'image.

5. Résultats

La figure 1 résume les résultats de l'expérience : elle indique l'exactitude des annotations réalisées en fonction de la proportion d'images annotées. Lorsque toutes les images sont annotées, la précision des annotations est de l'ordre de 40% d'images correctement annotées. Bien que cela corresponde au critère d'annotation le plus souple ($S = 1$), cette précision n'est pas mauvaise. De plus, la précision atteint 90% lorsque le seuil S est suffisamment élevé ($S = 10$). Dans ce cas, seulement 2% des images sont annotées. Pour $S > 10$, cette précision diminue légèrement car ce sont majoritairement des images correctement annotées qui ne sont plus annotées, mais pour $S = 15$, la précision des annotations atteint 100%. Ces résultats montrent que notre méthode, bien que simple, peut être fiable, pour peu que le seuil S soit suffisamment élevé. Enfin, le seuil optimal en terme de F_1-mesure est $S = 2$.

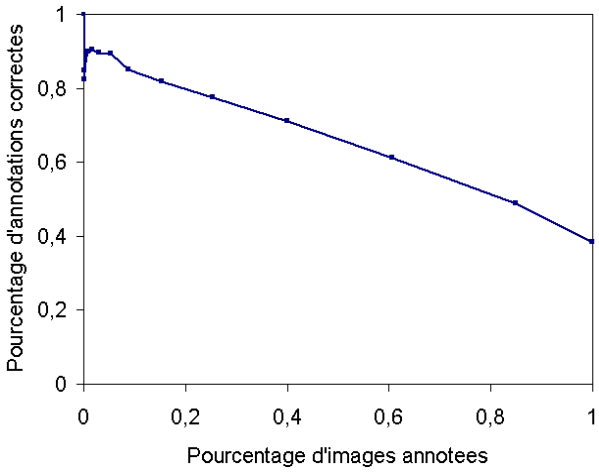

Figure 1. *Précision des annotations en fonction du taux d'images annotées*

Ces résultats sont encourageants : ils montrent que l'association de la détection de visages et de la détection d'entités nommées permet d'annoter des images de manière assez efficace, malgré un critère d'annotation trivial. On peut ainsi espérer que l'utilisation d'un critère plus élaboré pour annoter les images améliorerait considérablement les résultats. L'utilisation d'outils dédiés permet donc, dans le cadre qui leur est propre, de réaliser efficacement l'annotation sémantique d'images.

6. Conclusions et perspectives

Nous avons présenté une méthode d'annotation d'images contenant des visages dans un grand corpus de documents texte/images. Cette méthode se base sur l'utilisation conjointe d'un outil de détection de visages dans les images (OPENCV) pour repérer les images contenant une personne, et d'un outil de détection des entités nommées (NEMESIS) pour extraire les noms cités dans le texte. Les images peuvent ainsi

être annotées par les noms dont la fréquence dans le texte est significative. Les résultats de l'expérience montrent que cette méthode, bien que simple, donne de bonnes performances en termes de précision des annotations réalisées.

Plusieurs améliorations peuvent être apportées à ce système d'annotation. D'une part, on pourrait affiner l'extraction des entités nommées en tenant compte de la structure du texte, par exemple en accordant plus de poids aux entités présentes dans le titre des articles. D'autre part, l'ajout d'un système de reconnaissance des visages permettrait de propager les mots-clefs des images annotées vers des images dont l'annotation est peu fiable. Enfin, l'utilisation d'un critère d'annotation plus élaboré que la fréquence des termes, s'inspirant par exemple des pondérations utilisées en RI, pourrait améliorer les résultats. De plus, l'évaluation de notre méthode pourrait être améliorée, notamment en comparant ses performances à celles d'autres méthodes et en constituant manuellement un jeu de test permettant d'évaluer le rappel de l'annotation (dépendant uniquement du rappel du détecteur de visages). Ce travail constitue donc une première étape vers l'élaboration de techniques d'indexation textuelle des images plus intelligentes, se basant sur des caractéristiques haut-niveau des documents. Il peut également être adapté à l'annotation de documents vidéos accompagnés de texte (télétexte, transcription de la bande sonore...).

7. Bibliographie

Berg T. L., Berg A. C., Edwards J., Maire M., White R., Teh Y.-W., Learned-Miller E., Forsyth D. A., « Names and Faces in the News », *CVPR'04 : Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, États-Unis, p. 848-854, 2004.

Davis M., Smith M., Canny J., Good N., King S., Janakiraman R., « Towards context-aware face recognition », *MULTIMEDIA '05 : Proceedings of the 13th annual ACM international conference on Multimedia*, Singapore, p. 483-486, 2005.

Fourour N., « Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français », *TALN'02 : Actes de la conférence sur le Traitement Automatique des Langues Naturelles*, Nancy, France, p. 265-274, 2002.

Intel, « OpenCV : Open Source Computer Vision Library », 2006.
http ://www.intel.com/technology/computing/opencv/overview.htm.

Kuchinsky A., Pering C., Creech M. L., Freeze D., Serra B., Gwizdka J., « FotoFile : a consumer multimedia organization and retrieval system », *CHI '99 : Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 496-503, 1999.

Lienhart R., Kuranov A., Pisarevsky V., « Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection », *DAGM'03 : Proceedings of the 25th Symposium for Pattern Recognition*, Magdeburgh, Allemagne, p. 297-304, 2003.

Satoh S., Nakamura Y., Kanade T., « Name-It : Naming and Detecting Faces in News Videos », *IEEE MultiMedia*, vol. 6, n° 1, p. 22-35, 1999.

Zhang L., Hu Y., Li M., Ma W., Zhang H., « Efficient propagation for face annotation in family albums », *MULTIMEDIA '04 : Proceedings of the 12th annual ACM international conference on Multimedia*, New-York, États-Unis, p. 716-723, 2004.