Évaluation de descripteurs visuels pour l'annotation automatique d'images patrimoniales

David PICARD, Philippe-Henri GOSSELIN

ETIS UMR 8051
ENSEA / UCP / CNRS, 6 avenue du Ponceau, 95014 Cergy-Pontoise, France
picard@ensea.fr, gosselin@ensea.fr

Résumé – Dans cet article, nous évaluons plusieurs familles de descripteurs visuels couramment utilisées en classification d'images pour effectuer de l'annotation automatique dans des collections d'images patrimoniales. Dans de telles collections, les mots-clef utilisés sont souvent très précis et le nombre d'exemples d'apprentissage est relativement faible, ce qui rend le problème plus difficile que dans les traditionnels challenges de vision par ordinateur.

Abstract – In this paper, we evaluate several types of visual descriptors used in image classification when used to perform automatic labeling of cultural heritage images. In such collections, the keywords are often very precise and there are very few training samples. These two properties together make the problem of automatic labeling much more difficult than in academic computer vision benchmarks.

1 Introduction

Les collections patrimoniales conservées par les bibliothèques et les musées sont aujourd'hui numérisées sous forme d'images annotées et mises à disposition du public à travers des portails web. À titre d'exemple, les images de la bibliothèque nationale de France (BnF), de la Library of Congress américaine (LoC), ou de l'Institut royal des Arts des Pays-Bas (RKD) sont accessibles via des pages dédiées sur les sites web de ces insitutions [1].

Cependant, afin de permettre un accès facile pour les utilisateurs, chaque image doit être annotée d'une série de mots-clef caractéristiques de l'image. Ces mots-clef doivent décrire des propriétés qui dépendent de la nature de l'objet (par exemple «tableau», «sculpture»), sa période («17^{e}siècle»), son origine géographique («Suisse»), des attributs visuels («rouge», «forme géométrique») ainsi que son contenu sémantique («chèvre», «roi de France«). De part leur spécificité, ces mots-clef sont actuellement ajoutés manuellement par un opérateur à chaque numérisation d'œuvre. L'annotation manuelle est donc un frein à la vitesse de numérisation et de mise à disposition de ces collections.

Dans cet article, nous proposons donc d'évaluer les différentes méthodes issue de la classification d'images pour effectuer de l'annotation automatique ou semi-automatique dans le contexte des images patrimoniales. Afin d'étudier les propriétés des images qui sont pertinentes pour ce type d'annotation, nous nous concentrons sur les caractéristiques visuelles qui sont extraites des images et fixons le protocoles de classification. La principale contribution de cet article est une étude expérimentale approfondie des performances de plusieurs grandes familles de descripteurs visuels dans le contexte de l'annotation d'images patrimoniales.

Dans une première partie, nous présentons les descripteurs visuels qui sont utilisés dans ce travail. Puis, nous présentons notre protocole de test dans la partie suivante. Enfin, nous présentons des résultats détaillés ainsi que leurs analyses avant de conclure.

2 Descripteurs visuels pour la classification d'images

Les descripteurs visuels utilisés en classification d'images sont généralement séparés en trois grandes familles : les descripteurs globaux, les agrégats de descripteurs locaux et les architectures profondes. Dans cette partie, nous présentons les principales caractéristiques de ces familles, ainsi que les méthodes que nous utilisons dans nos expérimentations.

2.1 Descripteurs globaux

Les descripteurs globaux ont été les premiers types de descripteurs utilisés pour effectuer la classification d'images [1]. Ces descripteurs consistent à calculer des caractéristiques visuelles sur l'ensemble de l'image. Par exemple, les premiers descripteurs globaux à avoir été utilisés sont des histogrammes de couleur [2] et des histogrammes de textures [3], obtenus respectivement par quantification vectorielle de l'espace des cou-

1. http://images.bnf.fr pour la Bibliothèque nationale de France, http://www.loc.gov/pictures pour la Library of Congress et https://rkd.nl/en/ pour l'Institut Royal des Arts des Pays-Bas.

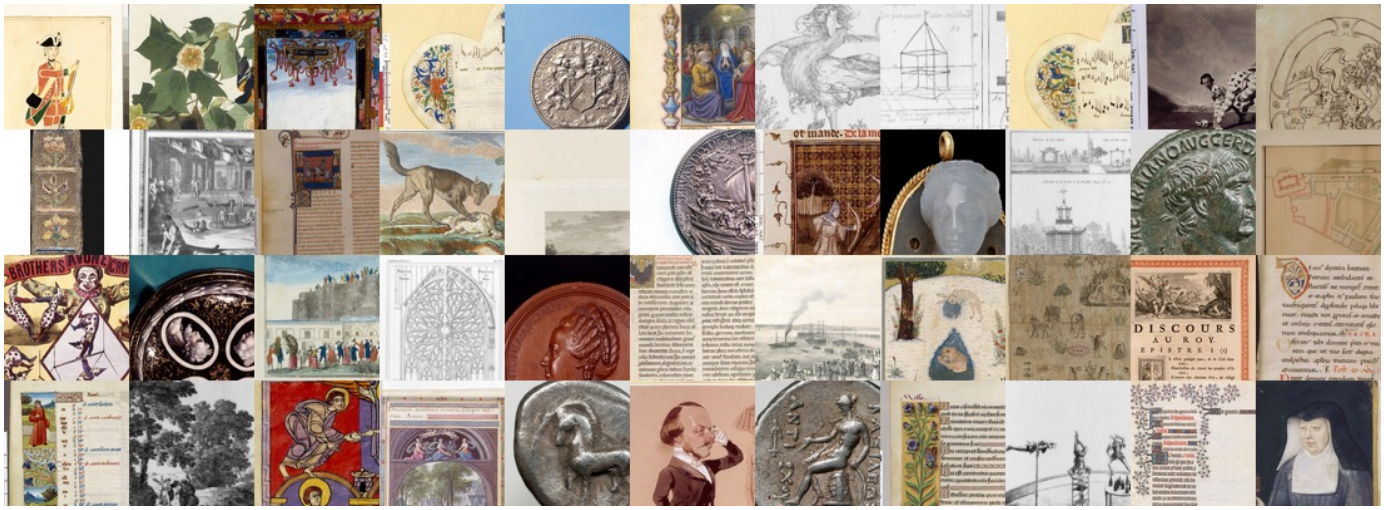

FIGURE 1 – Collage d'un ensemble d'images de la Bibliothèque nationale de France utilisées pour l'évaluation. On peut remarquer la diversité tant sur les sujets que sur les types d'objets photographiés. Toutes les images utilisées sont disponibles sur `http://images.bnf.fr`.

leurs ou de l'espace engendré par un banc de filtres.

Cependant, de tels descripteurs ne prennent pas en compte la disposition spatiale du contenu visuel dans l'image. Par exemple, un motif à 2 bandes et un motif en damier ont à peu près les mêmes descripteurs malgré leur apparence visuelle complètement différente. Certains descripteurs proposent de prendre en compte l'information de disposition spatiale par découpage de l'image en blocs [4]. Un descripteur est alors calculé dans chaque bloc. Le descripteur de l'image est la concaténation de chacun des descripteurs.

Dans le cas des collections patrimoniales, une bonne partie des images sont centrées sur l'objet, ce qui est cohérent avec l'utilisation du découpage en blocs. Cependant, certaine annotation concernent des détails précis dans l'image (par exemple un animal présent dans une enluminure) qui peuvent être trop complexes pour être modélisés par un histogramme de couleurs ou de sorties de filtres. D'autre part, la position de ces détails dans l'objet peut énormément varier et être ainsi très mal représentée par le découpage en bloc.

2.2 Agrégats de descripteurs locaux

Pour résoudre le problème des annotations basées sur des régions précise de l'image, les représentations par agrégation de descripteurs locaux ont été proposées avec succès dans la littérature [5]. On parle généralement de «*sacs de mots*», terme auquel nous préférons l'expression «*agrégats de descripteurs locaux*». L'idée est d'utiliser des descripteurs locaux très discriminants, par exemple SIFT [6] ou SURF [7], et de compter le nombre d'appariements des ces descripteurs entre deux images.

Pour éviter d'avoir à effectuer un appariement coûteux (car dépendant du carré du nombre descripteurs extraits par image) à chaque comparaison d'images, les descripteurs peuvent être agrégés dans un unique vecteur de telle sorte que la comparaison de ces vecteurs est proche du résultat obtenu par appariement. Un manière très simple proposée dans [5] consiste à calculer un dictionnaire de prototypes de descripteurs à l'aide de l'algorithme k-means, puis de calculer l'histogramme d'occurrences de ces prototypes dans l'image. Afin d'améliorer les performances en reconnaissance d'objets, des signatures calculant des statistiques d'ordres plus élevés telles que VLAD [8] (ordre 1) ou Fisher Vector [9, 10] (ordre 2) ont été proposées avec succès.

Dans les cas où la disposition spatiale des motifs est importante (par exemple une représentation d'une scène mythologique), le découpage en blocs a aussi été proposé sous la forme de pyramide spatiale [4].

Pour les collections patrimoniales, les agrégats de descripteurs locaux ont l'avantage de pouvoir capturer certains détails de l'image correspondants à des annotations sémantiques très précises. Cependant, ces méthodes sont très dépendantes des descripteurs locaux utilisés. De fait, ceux-ci ne sont pas forcément bien adaptés à certains motifs présents dans les collections patrimoniales comme le matériaux de l'objet ou bien des décors présents en bordure.

2.3 Architectures profondes

Afin d'adapter les descripteurs aux spécificités de la collection considérée, les réseaux de neurones à architecture profonde ont récemment connu plusieurs succès [11, 12, 13, 14, 15, 16]. Les réseaux les plus utilisés sont des réseaux dit convolutifs, c'est à dire que les poids d'une couche sont partagés entre les régions auxquelles ils s'appliquent. On peut alors voir les activations d'une couche comme la convolution de la couche d'entrée par un certain nombre de filtres.

Dans les réseaux convolutifs, un couche de convolution est généralement suivie d'une couche d'activation non linéaire, typiquement une simple rectification ($\max(0, output)$), puis d'une

couche d'agrégation (*max pooling* en anglais), où la valeur maximale de sortie est conservée dans un voisinage local. Ces 3 couches sont ainsi répétées plusieurs fois. En fin de réseau, des couches intégralement connectées permettent d'avoir une représentation sur la totalité de l'image.

L'apprentissage des poids des différentes couches peut se faire de manière non supervisée, comme c'est le cas dans les auto-encodeurs. Lorsqu'une vérité terrain est disponible, la rétropropagation du gradient de l'erreur de sortie permet de d'affiner les poids du réseau. Cet entraînement supervisé peut avoir un impact positif dans le cas des collections patrimoniales en intégrant au sein des descripteurs une information sémantique plus importante.

3 Expérimentations et résultats

Dans cette section, nous détaillons dans un premier temps notre protocole expérimental et les descripteurs que nous avons utilisés. Puis, nous présentons les résultats et leur analyse.

3.1 Protocole expérimental

Pour notre évaluation, nous utilisons les images de la bibliothèques nationale de France décrites dans [17]. Cette collection contient 3016 images annotées dont un collage est présenté dans la figure 1. Les annotations sont puisées dans un vocabulaire de plus de 400 concepts non exclusifs. Les concepts sont découpés en catégories correspondant à différent types d'information : visuelle, sémantique, géographique, historique et physique. Seuls les concepts de plus de 10 images ont été gardés. L'ensemble des programmes nécessaire à faire tourner le challenge ainsi que les images sont disponibles en ligne [2]. Nous évaluons les résultats en termes de précision et de score F1 (qui est la moyenne harmonique entre précision et rappel), en utilisant un validation croisée à découpage en cinq parties (l'entrainement se fait donc sur 80% du concept et l'évaluation sur les 20% restant, et ce cinq fois). Étant donné le faible nombre d'images par mot-clef, la justesse (*accuracy* en anglais) de classification est peu informative car un classificateur répondant systématiquement "faux" a peu de chance de se tromper.

Les caractéristiques utilisées dans cette évaluation correspondent aux 3 catégories décrites : descripteurs globaux, agrégats de descripteurs locaux et architecture profondes. Pour les descripteurs globaux, nous comparons des histogrammes de couleur obtenus par quantification vectorielle de l'espace *Lab* en 256 valeurs. Nous utilisons aussi des histogrammes de texture obtenus par quantification vectorielle en 256 valeurs d'un espace formé par les sorties d'un banc de filtres correspondant à une transformée quaternionique que nous appelons «qw». Enfin, les descripteurs GIST [1] complètent notre ensemble de descripteurs globaux.

2. http://perso-etis.ensea.fr/~picard/bnf_bench

Des.	Phy.	Vis.	Sém.	His.	Géo.	Moy.
lab	4.8%	4.0%	2.1%	3.5%	8.5%	4.6%
qw	7.2%	10.0%	3.9%	3.8%	12.2%	7.4%
gist	19.3%	21.8%	9.0%	10.9%	16.6%	15.5%
llc	8.3%	19.0%	5.2%	3.0%	13.7%	9.8%
fv4k	24.6%	30.3%	13.6%	13.5%	18.5%	20.1%
dl8	34.6%	31.1%	24.7%	13.1%	30.1%	24.7%
dl19	30.9%	29.1%	15.5%	17.2%	31.0%	24.8%

TABLE 1 – Précision moyenne pour chaque groupe de catégories en fonction des descripteurs visuels.

Pour les agrégats de descripteurs locaux, nous utilisons les Fisher Vectors avec un dictionnaire visuel de 4096 mots visuels agrégeant des descripteurs locaux de type SIFT [10]. Nous utilisons également un descripteur de type «Sac de Mots» améliorée appelée LLC [18] dans lequel les sift sont codés sur le dictionnaire de manière à minimiser l'erreur de reconstruction tout en conservant des propriétés de parcimonie. Ce descripteur utilise un découpage en bloc, comme pour le GIST.

Pour les architectures profondes, nous comparons un réseau à 8 couches [12] et un réseau très profond à 19 couches [15], tous deux entraînés sur ImageNet. Nous considérons à chaque fois l'avant-dernière couche que nous utilisons comme descripteur visuel. Il est a noter que ces réseaux ont été entraînés sur une autre base d'images et que l'étape de *fine-tuning* n'a pas été effectuée.

3.2 Résultats

Nous présentons dans le tableau 1 la précision obtenue pour chaque catégorie et chaque descripteur. Le tableau est composé de trois groupes de lignes correspondant aux familles de descripteurs visuels. Comme nous pouvons le voir, les réseaux convolutifs offrent les meilleurs performances, en particulier pour les catégories sémantiques. Cela peut s'expliquer par le fait qu'ils sont entraînés de manière supervisée sur un autre jeu de données avec une information sémantique. Au mieux, la précision moyenne est en dessous de 25%, ce qui veut dire que beaucoup de résultats sont des faux positifs.

Nous présentons dans le tableau 2 les scores F1 pour chaque famille de descripteur et chaque catégorie. Le score F1 correspond à la moyenne harmonique entre la précision et le rappel et est donc un bon compromis entre le taux de faux positifs et le taux de vrai négatifs. Comme on peut le voir, les scores F1 sont assez faibles pour la très grande majorité des descripteurs et des catégories. Seuls les réseaux profonds arrivent aux environs de 20% en moyenne. En terme de catégories, les catégories sémantiques et historiques semblent particulièrement difficiles à annoter. Cela peut s'expliquer par le fait que beaucoup de concepts dans ces catégories ont peu d'exemples avec pourtant une grande variété de contenu, rendant le rappel très difficile à maximiser.

Desc	Phy.	Vis.	Sém.	His.	Géo.	Moy.
lab	7.1%	6.2%	3.2%	4.8%	11.8%	6.6%
qw	9.1%	13.0%	5.4%	5.2%	15.1%	9.6%
gist	17.1%	17.6%	6.8%	7.5%	16.3%	13.1%
llc	6.7%	11.5%	3.3%	0.8%	12.4%	7.0%
fv4k	14.7%	20.7%	8.4%	8.7%	14.4%	13.4%
dl8	23.6%	25.6%	11.1%	10.3%	25.3%	19.2%
dl19	26.4%	24.3%	12.3%	13.8%	27.7%	20.9%

TABLE 2 – Score F1 pour chaque groupe de catégories en fonction des descripteurs visuels.

4 conclusion

Dans cet article, nous avons présenté une évaluation de plusieurs familles de descripteurs visuels populaires dans le contexte de la classification d'images. Nous avons effectué notre évaluation sur une collection d'images patrimoniales issues de la Bibliothèque nationale de France dont la particularité est d'avoir des annotations très précises et pour lesquelles peu d'exemples sont disponibles.

Notre évaluation montre que les descripteurs visuels issus de réseaux de neurones à architecture profonde donnent les meilleurs résultats. Cependant, les scores de précision et les scores F1 sont trop faibles pour pouvoir utiliser ces descripteurs en conditions réelles.

Remerciements

Ce travail a été effectué dans le cadre du projet ASAP financé par le labex Patrima et la Fondation des Sciences du Patrimoine.

Références

[1] A. Oliva and A. Torralba, "Modeling the shape of the scene : A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[2] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[3] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 8, pp. 837–842, 1996.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.

[5] J. Sivic and A. Zisserman, "Video google : a text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, vol. 2, October 2003, pp. 1470–1477.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, 2004, pp. 91–110.

[7] H. Bay, T. Tuytelaars, and L. J. V. Gool, "Surf : Speeded up robust features," in *European Conference on Computer Vision*, 2006, pp. 404–417.

[8] R. Arandjelović and A. Zisserman, "All about vlad," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[9] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, 2010, pp. 143–156.

[10] P. H. Gosselin, N. Murray, H. Jégou, and F. Perronnin, "Revisiting the fisher vector for fine-grained classification," *Pattern Recognition Letters*, vol. 49, pp. 92–98, 2014.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.

[13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.

[14] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587. [Online]. Available : http ://arxiv.org/pdf/1311.2524v5.pdf

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers : Surpassing human-level performance on imagenet classification," *CoRR*, vol. abs/1502.01852, 2015.

[17] D. Picard, P.-H. Gosselin, and M.-C. Gaspard, "Challenges in content-based image indexing of cultural heritage collections," *Signal Processing Magazine*, July 2015.

[18] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3360–3367.