Évaluation automatique de la qualité esthétique des photographies à l'aide de descripteurs d'images génériques

Luca Marchesotti, Florent Perronnin, Diane Larlus, Gabriela Csurka, Loïc Michallon

Évaluation automatique de la qualité esthétique des photographies à l'aide de descripteurs d'images génériques

Luca Marchesotti Florent Perronnin Diane Larlus Gabriela Csurka Loïc Michallon

Xerox Research Center Europe (XRCE)
6 chemin de Maupertuis, 38240 Meylan, France
{prenom}.{nom}@xrce.xerox.com

Résumé

Cet article s'intéresse à l'évaluation automatique des propriétés esthétiques des images. Par le passé, ce problème a été considéré à l'aide de descripteurs conçus à la main, de façon à correspondre aux bonnes pratiques photographiques (par exemple : "est-ce que cette photo est une macro ?"). Nous nous différencions fondamentalement de cette approche, et proposons d'utiliser des descripteurs d'images génériques pour l'évaluation de la qualité esthétique. Nous montrons expérimentalement que les descripteurs utilisés, qui accumulent des statistiques locales de bas niveau, encodent implicitement les propriétés utilisées explicitement par les méthodes actuellement les plus performantes, et surpassent celles-ci de façon significative.

Mots Clef

descripteurs d'image, esthétique, catégorisation d'images

Abstract

*In this paper, we automatically assess the aesthetic properties of images. In the past, this problem has been addressed by hand-crafting features which would correlate with best photographic practices (*e.g.*"Does this image respect the rule of thirds ?") or with photographic techniques (*e.g.*"Is this image a macro ?"). We depart from this line of research and propose to use generic image descriptors to assess aesthetic quality. We experimentally show that the descriptors we use, which aggregate statistics computed from low-level local features, implicitly encode the aesthetic properties explicitly used by state-of-the-art methods and outperform them by a significant margin.*

Keywords

image descriptors, aesthetics, image categorization

1 Introduction

Dans cet article, nous nous intéressons à la tâche ambitieuse qui consiste à concevoir des méthodes capables de prédire de façon automatique des scores de qualité esthétique des images [32]. De tels modèles ont de nombreuses applications. Les moteurs de recherche d'images pourraient ainsi réordonner les résultats en fonction des

FIG. 1 – Photos très bien notées dans un site communautaire de partage de photos en ligne (*photo.net*).

propriétés eshtétiques. Ces modèles pourraient permettre de sélectionner les meilleures images d'une collection personnelle de photos pour aider à la construction d'un album. Ils pourraient également être déployés directement sur un appareil photo numérique pour faire des suggestions en temps réel.

La question de la qualité esthétique d'une image peut être vue comme un problème de classification binaire. Cette image est-elle esthétiquement plaisante ou non ? C'est un problème difficile pour plusieurs raisons. Premièrement, les données visuelles sont riches et ambiguës. Deuxièmement, notre jugement d'une photographie dépend beaucoup de nos goûts personnels. Enfin, même s'il est facile de s'accorder pour dire que les dégradations bas niveau (par exemple, une image floue ou bruitée) sont, en général, de bons indicateurs d'une qualité médiocre, il est plus difficile de trouver un consensus sur les propriétés visuelles de plus haut niveau, comme l'harmonie des couleurs, la composition ou l'illumination d'une photo. Face à de telles difficultés, la possibilité même de construire des modèles génériques qui encodent les préférences photographiques peut être remise en question.

Cependant, il existe un consensus entre les professionnels sur certaines bonnes pratiques photographiques (comme la règle des tiers), ainsi que sur des techniques photographiques (par exemple la macro). En se basant sur ces

connaissances, des premières méthodes d'estimation esthétique ont été proposées. Elles utilisent des descripteurs visuels d'images *conçus à la main* afin de reproduire au mieux ces règles photographiques. Combinées avec des classifieurs supervisés, ces méthodes ont obtenu de bons résultats dans la prédiction de la qualité des images. Bien que ces descripteurs soient motivés par des règles esthétiques, ils ont un certain nombre de désavantages : i) ils ne sont pas exhaustifs : ils ne pourront jamais couvrir tous les principes photographiques, ii) ils sont coûteux en temps de calcul, et iii) ils sont basés sur des heuristiques qui risquent de mal se généraliser à d'autres applications similaires.

C'est pourquoi, dans cet article, nous poursuivons une approche différente et considérons des *descripteurs génériques*, qui ont obtenu d'excellents résultats sur des tâches sémantiques. Alors que les méthodes basées sur des descripteurs conçus à la main tentent de modéliser les meilleures pratiques photographiques de façon *explicite*, nous pensons que la même information est déjà contenue, au moins partiellement, de façon *implicite* dans les descripteurs génériques d'images. Nous montrons dans cet article que ces descripteurs, indépendants de la tâche, surpassent les descripteurs de l'état de l'art, sur deux bases esthétiques d'images. Des résultats qualitatifs confirment notre intuition initiale que les descripteurs génériques sont capables d'apprendre des propriétés esthétiques visuelles de façon efficace et implicite.

En particulier, nous considérons deux familles de descripteurs génériques, l'approche par sac-de-mots (*Bag-of-visual-words* [4]) et le vecteur de Fisher (*Fisher Vector* ou FV [21]), qui encodent tous deux des statistiques locales des images. Ces descripteurs se sont montrés performants sur un large spectre de tâches sémantiques (classification d'objets, de scènes, détection et segmentation d'images, ..).

Dans la section suivante, nous proposons un aperçu des travaux déjà effectués dans ce domaine. Dans la section 3, nous décrivons les descripteurs d'images génériques utilisés. La section 4 décrit les bases d'images. Dans la section 5, nous comparons les descripteurs esthétiques aux descripteurs génériques sur ces bases. Enfin, nous proposons une discussion des résultats obtenus.

2 Travaux antérieurs

Les techniques les plus récentes pour l'évaluation de la qualité esthétique des images sont basées sur l'apprentissage automatique. Une fonction de prédiction est apprise à partir d'un ensemble d'apprentissage étiqueté [7, 1, 19].

Les efforts se concentrent actuellement sur la conception de descripteurs appropriés pour ce problème d'apprentissage. Ces descripteurs doivent capturer les propriétés visuelles qui font qu'une image donnée est plaisante pour la majorité de ses spectateurs. Typiquement, des propriétés bas niveau et haut niveau sont utilisées de façon conjointe. Les propriétés bas niveau incluent l'exposition, le contraste, la gamme de couleur et la texture des photographies. Les propriétés haut niveau sont principalement liées à l'analyse de la composition de l'image.

Les premiers travaux d'analyse esthétique ont utilisé des descripteurs bas niveau, inspirés par des métriques liées à la perception humaine [25, 12, 5]. Les plus populaires [27] sont basées sur de simples statistiques (comme la moyenne et l'écart type) calculées sur la totalité de l'image, de façon à caractériser les dégradations du signal, comme le bruit, aléatoire ou structuré [26]. D'autres descripteurs sont basés sur des techniques d'estimation du flou, où celui-ci est modélisé comme une filtre gaussien [29].

Actuellement, les chercheurs s'intéressent à des représentations d'un plus haut niveau d'abstraction, suivant l'intuition que les images de grande qualité respectent un certain nombre de règles photographiques utilisées par les photographes professionnels (comme le flou d'arrière plan ou "bokeh", le clair-obscur, le flou de mouvement, la photographie à haute vitesse, la règle des tiers, la macrophotographie, ou la presence de lignes directrices) [13].

[1] proposent donc des descripteurs conçus à la main afin de détecter ces règles photographiques. Des techniques de segmentation sont souvent utilisées pour obtenir une description plus haut niveau de l'image, comme par exemple pour la règle des tiers où il faut déterminer la position du sujet principal. Une simple segmentation en régions peut être utilisée [25], ou des méthodes plus complexes peuvent repérer les objets dans la plage de netteté en détectant les contours flous, correspondant au fond [17]. Cette dernière méthode est particulièrement utile pour détecter les images capturées avec une technique populaire faisant intervenir une grande ouverture.

Une autre approche consiste à localiser certains objets spécifiques, par exemple, [15] utilisent le détecteur de Viola & Jones pour la détection de visage, pour ensuite localiser les yeux et la bouche. La forme des objets segmentés peut être caractérisée à l'aide de descripteurs géométriques [1]. De plus, la position absolue de ces objets peut être comparée aux positions considérées comme importantes d'un point de vue esthétique (par exemple, les points désignés par la règle des tiers, ou par le nombre d'or [13]). Enfin, des métriques ont été définies [19] pour décrire des compositions visuellement agréables en terme de positions relatives entre les objets, présence de points de fuite ou de perspectives.

Dans cet article, nous dévions de cette ligne de recherche, et proposons d'utiliser des descripteurs d'images génériques [4, 21]. Ces descripteurs ont été appliqués avec succès à des tâches sémantiques, comme la recherche d'objets ou de scènes [28, 22], la classification et l'annotation d'images [4, 23] et la localisation d'objets [10]. Ils se sont montrés polyvalents, obtenant de bons résultats pour des scénarios avec un grand nombre de classes, et passant à l'échelle de grandes bases de données. Une première version de ces travaux sera publiée dans [18].

3 Descripteurs génériques d'images

Nous proposons d'utiliser des signatures d'images génériques, basées sur le contenu, pour l'évaluation esthétique

des images. Nous considérons d'abord le descripteur par sac-de-mots (*Bag-of-Visual-words* ou BOV) [4, 28], qui est probablement le descripteur visuel le plus utilisé pour l'analyse sémantique. Nous utilisons aussi une de ses extensions récentes, le vecteur de Fisher (FV) [21, 23].
Notre motivation est la suivante : au lieu d'essayer d'encoder les règles photographiques de façon explicite, nous les encodons de façon implicite dans des descripteurs génériques basés sur le contenu, comme BOV ou FV, qui décrivent les images comme des descriptions locales de patchs (BOV : une distribution discrète, FV : une distribution continue). En effet, chaque patch (par exemple décrit par SIFT, ou par un descripteur couleur) peut nous en dire beaucoup sur les propriétés locales des images ("ce patch contient-il des contours nets ?" ou "la couleur de ce patch est-elle saturée ?"). De plus, en accumulant l'information au niveau des patchs en un descripteur de l'image toute entière, les représentations BOV ou FV permettent de prendre en compte la composition globale ("a-t-on un mélange de contours nets et de contours flous ?", "y a-t-il une couleur dominante, ou un mélange de couleurs dans cette image ?"). Enfin, en utilisant une pyramide spatiale (comme décrit dans [14]), nous pouvons modéliser la composition de l'image (comme la règle des tiers).
Nous considérons aussi le descripteur GIST [20], utilisé au départ pour la catégorisation de scènes, puisqu'il devrait également capturer la composition des images. Ci-dessous, nous décrivons rapidement ces descripteurs, plus de détails peuvent être trouvés dans [20, 4, 21].

GIST. Oliva et Torralba [20] ont introduit le descripteur GIST comme un descripteur de scène de faible dimension. Un ensemble de dimensions perceptuelles, qui représente la structure globale d'une scène est estimé en utilisant l'information spectrale et une localisation grossière. En pratique, l'image est partitionnée en une grille régulière 4x4, et un descripteur de gradient de dimension 20 est calculé pour chaque région, et pour chaque canal de couleur. Leur concaténation produit un vecteur de dimension 960.

Sac-de-mots (BOV). Dans la représentation BOV [28, 4], une image est décrite comme un histogramme d'occurrence de descripteurs locaux quantifiés. Plus précisément, un ensemble non ordonné de patchs locaux est d'abord extrait et chaque patch est représenté, par exemple en utilisant le descripteur SIFT [16]. Un vocabulaire visuel, est construit en utilisant un algorithme de partitionnement (*clustering*) sur un grand nombre de descripteurs. L'ensemble des représentations locales extraites d'une image donnée est ensuite transformé en une représentation unique, un histogramme de longueur fixe, en comptant le nombre de représentations locales associées à chaque mot visuel. BOV a été appliqué avec succès à la recherche d'images [28] et à la classification [4].
Dans ce travail, nous suivons [9] et utilisons un modèle de mélange gaussien (GMM) pour modéliser la distribution des représentations locales, en d'autres termes, nous avons un vocabulaire visuel probabiliste. Le mélange gaussien fournit un moyen de gérer l'incertitude des associations, chaque description locale est associée avec une certaine probabilité à chaque mot visuel. Pour finir, nous appliquons la racine carré aux histogrammes, comme suggéré par [23, 31], ce qui correspond à un encodage explicite des données pour un noyau de Bhattacharyya. Nous avons vérifié expérimentalement une amélioration des résultats.
Alors que dans sa formulation initiale, BOV ne contient pas d'information géométrique, Lazebnik *et al.* [14] proposent d'inclure un modèle spatial grossier (appelé pyramide spatiale), en divisant hiérarchiquement l'image en un ensemble de régions, en calculant un histogramme BOV par région, puis en concaténant ces représentations en un seul vecteur .

Vecteurs de Fisher (FV). FV [11, 21] permet d'étendre BOV en allant au delà des simples comptages (statistiques d'ordre 0) et en encodant des statistiques d'ordre supérieur (1er et 2ème ordre) sur la distribution des descripteurs de patchs associés à chaque mot visuel. Il s'est avéré être performant sur plusieurs applications comme la classification [21, 23] et la recherche d'images [22]. Cela peut être expliqué par le fait qu'il traite les images comme des distributions continues, alors que BOV les traite comme des distributions discrètes. Un avantage majeur par rapport à BOV est l'obtention de signatures d'images discriminantes de grande dimension, même avec un vocabulaire visuel de taille modeste, et donc à très faible coût en CPU. De plus, [3] a montré que parmi les encodages de BOV récemment proposés, FV donne les meilleurs résultats.
Le FV $\mathcal{G}_\lambda^X$ caractérise un échantillon $X = \{x_t, t = 1 \dots T\}$ par sa déviation d'une distribution u_λ (de paramètres λ) :

$$\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X. \tag{1}$$

G_λ^X est le gradient de la log-vraisemblance selon λ :

$$G_\lambda^X = \frac{1}{T} \nabla_\lambda \log u_\lambda(X). \tag{2}$$

L_λ est la décomposition de Cholesky de l'inverse de la matrice d'information de Fisher F_λ de u_λ, soit $F_\lambda^{-1} = L'_\lambda L_\lambda$ où par définition :

$$F_\lambda = E_{x \sim u_\lambda} \left[\nabla_\lambda \log u_\lambda(x) \nabla_\lambda \log u_\lambda(x)' \right]. \tag{3}$$

Dans notre cas, X est l'ensemble des T descripteurs locaux extraits d'une image, et $u_\lambda = \sum_{i=1}^{N} w_i u_i$ est le mélange gaussien qui modélise les descripteurs locaux (le vocabulaire visuel probabiliste).
Comme montré dans [23], appliquer la racine carré, et une normalisation L2 au vecteur de Fisher améliore grandement ses résultats de classification. De plus, un modèle géométrique peut être ajouté de la même façon que pour BOV, en utilisant une pyramide spatiale.

4 Les bases d'images

Les bases d'images qui sont utilisées pour cette étude proviennent de sites web communautaires de partage

d'images, comme *photo.net* ou *DPChallenge*[1]. Ces sites regroupent un grand nombre de photographes, amateurs ou professionnels, qui partagent, regardent, et évaluent les photos en ligne. Ces photographes se mettent également d'accord sur les règles d'annotation. De telles annotations peuvent prendre la forme d'une étiquette ("j'aime", "je n'aime pas") ou d'une échelle de valeur numérique (un score). À partir de ces annotations, les images peuvent être étiquettées comme étant visuellement plaisantes ou non. Cette vérité terrain permet une évaluation quantitative des différentes méthodes.
Notre évaluation est basée sur les deux bases décrites ci-dessous, qui sont publiquement disponibles.

La base Photo.net. *Photo.net* est un réseau communautaire dont les utilisateurs peuvent évaluer les photographies à l'aide d'un score entre 1 (*déplaisante*) et 7 (*Magnifique*). Les administrateurs du site donnent les conseils suivants : "Les raisons pour un score proche de 7 : a) l'image est visuellement agréable, b) elle attire l'attention, c) sa composition est intéressante, d) elle a une bonne utilisation des couleurs, e) (si photo-journalisme) elle raconte une histoire, a de l'humour, a de l'impact, f) (si sport) elle représente l'apogée de l'action, un moment de lutte de l'athlète". La figure 1 montre des exemples de photos de grande qualité, ainsi que leur score et le nombre de votes.
Nous avons utilisé deux ensembles d'images dérivés de *photo.net* que nous appelons Photo.net (PN) et Photo.net découpée (PNc). PN a été introduite par [25] et utilisée par [25, 24]. Elle contient 3581 images. Cependant, seules les adresses (URL) des images originales sont fournies. Comme un certain nombre d'images ont été supprimées du site, seules 3118 images sont disponibles. Après inspection visuelle de PN, nous avons remarqué un lien entre les images qui obtiennent une bonne note, et la présence d'un cadre décoratif ajouté manuellement par les photographes. Nous avons ainsi observé que 30% des images possèdent un tel cadre. Comme cela introduit un biais dans la base, nous avons supprimé manuellement ces cadres, et avons créé une deuxième base d'images : PNc (*PN cropped*).

La base CUHK. Notre deuxième base d'images est tirée du site web *DPChallenge.com*. Nous utilisons la collection créée par [12], où 60000 photos ont été extraites, parmi celles notées par au moins 100 utilisateurs. La moyenne des votes sur les photos est utilisée comme vérité terrain, et les images dont la note se situe parmi les 10% meilleures, et les 10% moins bonnes ont été associées respectivement aux classes "bonne" et "mauvaise". 12000 images sont ainsi sélectionnées. Ke *et al*. [12] ont observé le même biais sur les cadres que celui observé pour PN, et ont enlevé tous les cadres de leur base.

5 Validation expérimentale

Descripteurs de référence. Nous avons implémenté les 56 descripteurs esthétiques de [25], les 7 descripteurs de [12] et les 5 descripteurs de [17]. Nous sommes parvenus à reproduire les résultats de [25] et [12] en utilisant les mêmes protocoles expérimentaux et les mêmes bases d'images. Cependant, nous n'avons pas réussi à répliquer les résultats annoncés par [17]. En effet, alors que [17] rapporte une réduction de l'erreur de classification de 80% par rapport à [25], notre implémentation de [17] nous amène à des résultats significativement moins bons que ceux de [25]. Par conséquent, nous n'incluons pas les descripteurs de [17] dans nos descripteurs esthétiques de référence.

Descripteurs génériques basés sur le contenu. Le descripteur GIST est calculé à l'aide du code mis à disposition par les auteurs de [20]. Pour BOV et FV, des patchs de taille 32x32 sont extraits selon une grille régulière, sur 5 échelles différentes. Nous utilisons deux types de descripteurs locaux pour représenter les patchs : SIFT [16] et un descripteur couleur (COL) [23]. SIFT divise chaque patch en une grille 4x4, et calcule un histogramme de l'orientation du gradient sur chaque sous-partie de la grille. De la même manière, le descripteur COL que nous utilisons divise le patch en une grille 4x4 et calcule des simples statistiques (moyenne et écart-type) sur chaque canal, pour chaque sous-partie de la grille. SIFT produit un vecteur de 128 dimensions et COL de 96 dimensions. Tout deux sont réduits à 64 dimensions, à l'aide d'une analyse en composante principale. Le vocabulaire visuel (GMM), est appris à l'aide d'un algorithme EM standard. Nous utilisons 1024 gaussiennes pour BOV, et 256 pour FV.
Pour la pyramide spatiale, nous suivons la stratégie de découpage adoptée par le système vainqueur de la compétition PASCAL VOC 2008 [8]. Nous extrayons 8 vecteurs par image : un pour toute l'image, 3 pour les régions haute, centrale et basse, et 4 pour chaque quart de l'image. Le but étant de capturer des informations sur la composition.

Classification. Nous apprenons un séparateur à vaste marge (SVM) linéaire, à l'aide d'un algorithme de gradient stochastique et d'une fonction de perte *hinge*, dans la formulation primale [2]. Pour chaque descripteur global, BOV ou FV, le descripteur local peut être SIFT ou le descripteur COL. La fusion entre SIFT et COL est calculée en moyennant les scores obtenus par le classifieur de chaque représentation (basée sur SIFT ou sur COL).

5.1 Protocole d'évaluation

PN. Suivant [25], pionnier dans l'utilisation de la base *photo.net*, nous calculons pour chaque image i la moyenne de tous les scores disponibles $q_{av}(i)$. Une analyse statistique de ces scores est disponible dans [6], et il apparaît que, pour *photo.net*, les votes sont biaisés en faveur des valeurs positives et que le consensus parmi les utilisateurs est faible. De plus, la valeur 5.0 semble être la valeur moyenne des votes. Suivant [6], nous définissons deux seuils $\theta_1 = 5 + \delta/2$ et $\theta_2 = 5 - \delta/2$. Les images sont ensuite annotées comme "bonne" si $q_{av}(i) \geq \theta_1$ et comme "mauvaise" si $q_{av}(i) \leq \theta_2$. δ permet de créer un intervalle entre les images de haute et de faible qualité, puisque les

[1] http ://www.photo.net et http ://www.dpchallenge.com

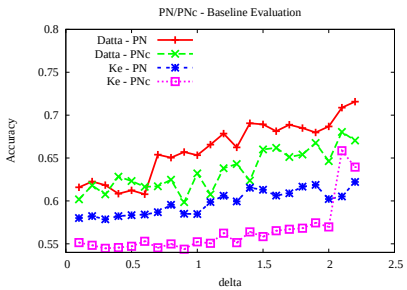

FIG. 2 – Précision pour les descripteurs Datta *et al*. [25] et Ke *et al*. [12], évaluée sur PN et PNc (sans les cadres).

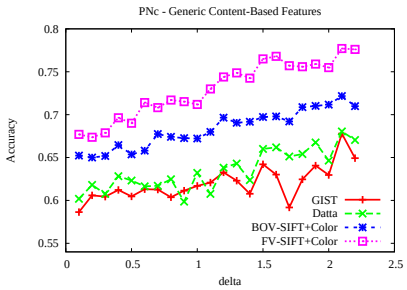

FIG. 3 – Précision sur PNc en utilisant GIST, la combinaison (SIFT+COL) pour les descripteurs BOV et FV.

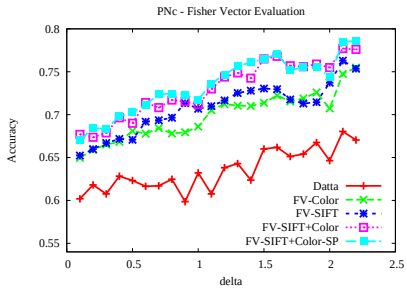

FIG. 4 – Différents FV sur PNc. SIFT et couleur, individuellement, fusionnés, avec la pyramide spatiale.

images dans cet intervalle sont susceptibles de représenter du bruit dans le processus d'annotation. De manière similaire à [25] nous faisons varier δ. Il est évident qu'augmenter δ induit une simplification de la tâche de classification. Comme suggéré dans [25], le score de classification rapporté est obtenu par validation croisée, et par calcul d'une précision moyenne.

CUHK. Comme défini par [12], nous utilisons la moitié des images (6000) pour l'entraînement, et l'autre moitié (6000) pour le test. Nous n'avons pas accès aux votes des utilisateurs. Nous nous appuyons donc sur les étiquettes binaires ("bonne" et "mauvaise") fournies par les auteurs.

5.2 Évaluation quantitative

Descripteurs de référence. La figure 2 donne la précision de classification en fonction du seuil δ pour les descripteurs de référence Datta *et al*.[25] et Ke *et al*.[12]. Afin d'évaluer le biais introduit par les cadres décoratifs, nous réalisons cette évaluation sur les 3118 images originales de la base PN, et sur les mêmes images manuellement recadrées. En moyenne, le biais se chiffre à environ 2% de précision pour chaque descripteur. Nous observons que les descripteurs de Datta surpassent ceux de Ke. Nous avons également essayé de fusionner les descripteurs de Datta et de Ke, mais aucune amélioration significative n'a été obtenue. Les résultats précédemment publiés sur [25] et [6] ne prennent pas en compte le biais possiblement induit par les cadres. Pour notre étude, nous utilisons comme référence la base d'images sans les cadres : PNc, et comme descripteurs de référence ceux proposés par [25] (appelés "Datta").

Descripteurs génériques proposés. La figure 3 présente les résultats obtenus avec les 3 descripteurs basés sur le contenu : GIST, BOV et FV (Pour BOV et FV nous utilisons la fusion des représentations basées sur SIFT et sur COL). Nous pouvons tirer les conclusions suivantes.

Premièrement, bien qu'étant le moins performant des descripteurs génériques, le descripteur GIST nous permet d'obtenir des résultats comparables à Datta. Ceci montre que de simples descripteurs, conçus pour un usage général et non pour une évaluation esthétique, peuvent être aussi performants que des descripteurs conçus, à la main, spécifiquement pour ce problème.

Deuxièmement, FV et BOV surpassent les descripteurs GIST et Datta, ce qui montre que la distribution des descripteurs locaux a un fort pouvoir discriminant pour l'évaluation de la qualité d'une image. En classification d'images, le vocabulaire visuel est souvent perçu comme une représentation intermédiaire qui comble le *fossé sémantique* entre les descripteurs locaux bas niveau et les concepts sémantiques haut niveau. Il semble également aider à combler le *fossé esthétique* entre les descripteurs bas niveau et la perception haut niveau.

Troisièmement, FV est le descripteur qui obtient les meilleures performances. Ceci a déjà été observé pour plusieurs tâches sémantiques, comme la catégorisation d'objets [21] et la recherche d'images [22].

Nous nous intéressons maintenant plus particulièrement à FV, et les résultats sont présentés dans la figure 4. Nous observons les propriétés suivantes.

Tout d'abord, les deux classifieurs entraînés avec les descripteurs SIFT et COL ont des performances équivalentes. La combinaison de ces deux représentations améliore les résultats.

Nous observons également que la pyramide spatiale a un impact très limité. Deux explications sont possibles. Le partitionnement choisi peut être inapproprié à cette tâche (d'autres choix de régions pourraient être testés). Ou bien, la pyramide augmentant significativement la dimension du descripteur d'images, la base d'images utilisée peut être trop petite pour entraîner de manière fiable un classifieur avec ces descripteurs.

Résultats pour CUHK. Nous complétons cette évaluation par des expériences sur la base d'images CUHK, dont les résultats sont présentés dans le tableau 1.

Premièrement, le descripteur GIST présente, pour cette base d'images, des résultats significativement moins bons que tous les autres descripteurs. Deuxièmement, les deux descripteurs de référence (Datta et Ke) donnent des résultats similaires. Encore une fois, FV présente les meilleurs résultats. Pour BOV et FV, le descripteur COL obtient de meilleures performances que SIFT, et cette différence est encore plus marquée pour FV. Leur fusion (SIFT+COL)

Descripteurs	GIST	Datta	Ke	BOV-SIFT-SP	BOV-Color-SP	FV-SIFT-SP	FV-Color-SP	FV-Fusion-SP
Précision	67.96	75.85	76.53	81.36	81.86	82.83	91.32	92.25

TAB. 1 – Performances sur la base CUHK pour les descripteurs de Datta, Ke, BOV et FV.

	FV-SIFT-SP	FV-Color-SP
RGB	82.83	91.32
rgb	86.65	87.00
$r_s g_s b_s$	83.05	85.03
HSV	83.85	89.62
Lab	78.88	90.38

TAB. 2 – Performances sur la base CUHK pour les descripteurs SIFT et Color en fonction de l'espace de couleur.

nous permet d'obtenir une légère amélioration.

Au vu des bons résultats du descripteur COL, nous complétons cette étude par d'autres descripteurs calculés sur des espaces de couleur alternatifs. Nous nous sommes inspirés de [30] pour sélectionner les espaces à tester. Nous considérons l'espace original R,G,B (résultats présentés précédemment), ainsi que deux variantes de celui-ci : *RGB normalisé* (r, g, b) et *RGB sphérique* (r_s, g_s, b_s). Ces deux espaces permettent de rendre les images invariantes à certaines modifications locales de l'intensité lumineuse (comme des ombres). Ils sont obtenus par $(r, g, b) = (\frac{R}{R+G+B}, \frac{G}{R+G+B}, \frac{B}{R+G+B})$ et $(r_s, g_s, b_s) = (\frac{R-\mu_R}{\sigma_R}, \frac{G-\mu_G}{\sigma_G}, \frac{B-\mu_B}{\sigma_B})$ respectivement, avec μ_i la moyenne du canal *i* sur l'image et σ_i son écart-type.

Nous avons également considéré les espaces HSV et CIE Lab*. L'espace HSV contient des informations sur la saturation, qui est une propriété technique utile pour l'évaluation esthétique des images. L'espace CIE Lab* a été conçu pour capturer les similarités perçues par la vision humaine. Le descripteur SIFT est calculé sur la luminance, dans RGB, ou après transformation dans les espaces normalisé ou sphérique. Dans le cas de HSV et de CIE Lab*, nous utilisons respectivement les canaux V et L. Le descripteur COL concatène les statistiques locales calculées sur chaque canal du descripteur, quelque soit la représentation choisie. Les résultats sont rapportés dans le tableau 2 et les observations sont les suivantes.

Le choix de l'espace pour le calcul du descripteur SIFT influe beaucoup sur les résultats. Le choix de l'espace CIE Lab* semble peu judicieux, alors que l'espace normalisé (r, g, b) améliore d'environ 4% par rapport au SIFT standard sur RGB. Cette normalisation des couleurs de l'image semble être bénéfique à l'extraction de l'information de texture.

Pour COL, les espaces alternatifs sont peu convainquants. L'espace CIE Lab* et l'espace initial RGB donnent des résultats équivalents et sont les descripteurs les plus appropriés pour cette base.

6 Discussion

Les bons résultats obtenus par notre méthode, qui utilise des descripteurs d'image génériques, prouvent que la stratégie adoptée est capable de capturer les statistiques utiles à l'évaluation esthétique des photographies.

Dans cette partie, nous proposons une analyse qualitative des résultats, et nous essayons d'expliquer les raisons de ce succès. Tirer des conclusions à partir d'un examen visuel de nos résultats est une tâche difficile. En effet, les bases d'images sont trop grandes pour visualiser chaque image individuellement et les annotations binaires sont disponibles sans plus d'explication. Cependant, nous souhaitons discuter de quelques tendances qui ont émergé de cette analyse qualitative. Pour commenter ces résultats, nous regardons les pratiques et techniques photographiques connues pour produire des images esthétiques. Nous nous concentrons sur la base CUHK puisqu'elle contient un plus grand nombre d'images et plus de votes par image en comparaison avec PN. Cependant, nos observations sont également valables pour PN. Nous étudions plus particulièrement deux descripteurs parmi les plus performants : FV-Color-SP et FV-SIFT-SP. Pour chacun, la figure 6 montre les 24 images ayant reçu les plus hauts et les plus bas scores. Les étiquettes issues de la vérité terrain sont affichées sous forme de cadres colorés (vert pour les images étiquetées comme "bonne" et rouge pour les autres).

Pour le descripteur SIFT, l'observation la plus évidente est que l'information de flou est particulièrement bien détectée. Toutes les images avec un score de classification élevé sont très nettes avec des contours contrastés. De la même manière, toutes les images floues reçoivent des scores faibles, ainsi que les images de faible résolution, pour lesquelles les contours ne peuvent pas être extraits facilement (voir les premiers faux négatifs, figure 5). Les scores restent élevés pour les photos avec une premier plan très net et un arrière plan flou. A l'inverse, les images encombrées sont classées comme "mauvaise". Finalement, les images à grande gamme dynamique (HDR) obtiennent également un score élevé. Nous pensons que toutes ces propriétés esthétiques sont apprises par le descripteur SIFT qui capture des informations de forme, de texture et de contour, mais aussi l'illumination de la scène. Si nous regardons les erreurs de classification pour le descripteur SIFT sur la figure 5, le premier faux positif correspond à une image dont la qualité bas niveau est bonne, mais pour laquelle le sujet peut être considéré comme inintéressant.

Nous regardons maintenant les résultats obtenus par le descripteur Couleur. Les propriétés chromatiques des images sont, comme attendues, celles qui différencient le plus les images de haute et de basse qualité. De nombreuses images de coucher de soleil, en général très populaires, font partie des mieux notées. Généralement, ces images ont une couleur dominante, ou des couleurs complémentaires (rouge et vert, bleu et jaune). En contraste, les images ayant trop de

FIG. 5 – (haut) images négatives avec un score élevé, *i.e*. faux positifs, (bas) images positives avec un score faible *i.e*. faux négatifs, pour SIFT-FV-SP et Col-FV-SP.

couleurs différentes ou des couleurs tristes, reçoivent des scores plus faibles. Mais le descripteur Couleur apprend également ce qui rend une image noir et blanc ou sépia visuellement attirante, et un bon nombre d'entre elles obtiennent un score élevé.

Même si l'influence est plus faible que pour SIFT, le descripteur Couleur capture également des informations de flou. Les images nettes et bien contrastées obtiennent des scores plus élevés que les images floues. Le descripteur Couleur est basé sur des statistiques sur les couleurs, mais il inclut également une information géométrique locale (à l'aide de la grille 4x4 qui couvre le patch). Cela signifie qu'il peut extraire (dans une moindre mesure que SIFT) des informations liées aux contours, au contraste et à la netteté. Cette dualité pourrait expliquer qu'il surpasse SIFT.

Comme observé dans la figure 4, la pyramide spatiale n'apporte pas d'amélioration significative. Nous pouvons cependant noter que, autant pour SIFT que pour Couleur, FV donne un score élevé aux images avec un agencement artistique. Les images appliquant la règle des tiers, montrant un sujet non centré, ou présentant de fortes lignes directrices semblent obtenir un score de classification élevé. Cela peut provenir d'une corrélation entre les propriétés d'agencement et les autres aspects esthétiques.

Toutes ces observations concordent avec notre hypothèse initiale : *les descripteurs d'images génériques sont capables d'apprendre efficacement et de manière implicite ce que les descripteurs de référence essayent d'apprendre de manière explicite*. Pour cette raison, nous croyons fermement qu'ils sont adaptés à l'évaluation esthétique des images. Dans nos expériences, nous avons considéré les deux plus grandes bases d'images disponibles publiquement pour de la catégorisation esthétique. Les descripteurs que nous considérons sont de très grande dimension, nous pensons donc qu'une évaluation à grande échelle permettrait d'améliorer significativement nos résultats, et de proposer une étude plus poussée. Les prochains travaux porteront donc sur la mise en place d'une grande base pour l'évaluation esthétique des images.

Références

[1] S. Bhattacharya, R. Sukthankar, and M. Shah. A coherent framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM MM*, 2010.

[2] Léon Bottou. Sgd. http ://leon.bottou.org/ projects/sgd.

[3] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details : an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV SLCV Workshop*, 2004.

[5] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik. Image quality assessment based on a degradation model. *Image Processing*, 9(11) :363–650, 2000.

[6] R. Datta, D. Joshi, J. Li, and J. Wang. Algorithmic inferencing of aesthetics and emotion in natural images : An exposition. In *ICIP*, 2008.

[7] R. Datta and J. Ze Wang. Acquine : aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In *MIR*, 2010.

[8] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, 2008.

[9] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation. Technical report, University of Southampton, 2005.

[10] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009.

[11] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999.

[12] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

[13] Kodak. *How to take good pictures : a photo guide*. Random House Inc, 1982.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[15] C. Li, A. Gallagher, A. Loui, and T. Chen. Aesthetic visual quality assessment of consumer photos with faces. In *ICIP*, 2010.

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[17] Y. Luo and X. Tang. Photo and video quality evaluation : Focusing on the subject. In *ECCV*, 2008.

[18] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011.

[19] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. *ICIP*, 2010.

[20] A. Oliva and A. Torralba. Modeling the shape of the scene : a holistic representation of the spatial envelope. *IJCV*, 2001.

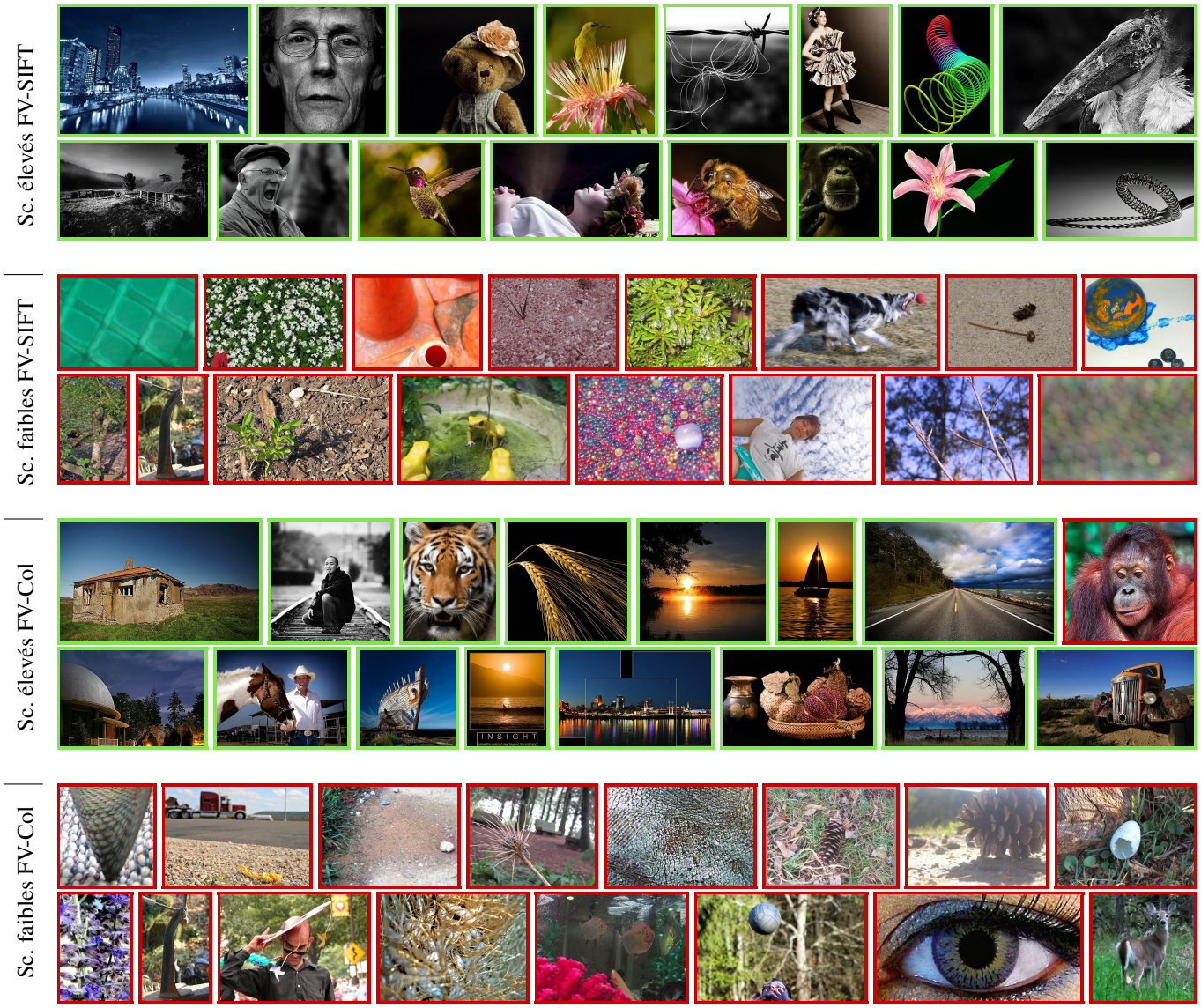

FIG. 6 – Pour la base CUHK, pour FV-SIFT-SP et FV-Color-SP, images les mieux (sc. élevé) et les moins (sc. faibles) bien notées sont affichées. Le cadre de couleur représente la vérité terrain (vert pour “bonne” et rouge pour “mauvaise”).

[21] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[22] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.

[23] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[24] J. Li R. Datta and J. Z. Wang. Learning the consensus on visual quality for next-generation image management. In *ACM MM*, 2007.

[25] J. Li R. Datta, D. Joshi and J. Ze Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.

[26] H.R. Sheikh, A.C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics : Jpeg2000. *Image Processing*, 14(11) :1918–1927, 2005.

[27] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing*, 15(11) :3440–3451, 2006.

[28] J. Sivic and A. Zisserman. Video Google : A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[29] H. Tong. Blur detection for digital images using wavelet transform. In *In Proceedings of IEEE International Conference on Multimedia and Expo*, pages 17–20, 2004.

[30] K. E. A. van de Sande ans T. Gevers and C. G. M. Snoek. Evaluation of color descriptors for object and scene recogition. In *CVPR*, 2008.

[31] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010.

[32] H. R. Sheikh Z. Wang and A. C. Bovik. *The Handbook of video databases : design and applications*, chapter 41. CRC press, 2003.