CLASSIC : un projet d'équipe

Convex Learning through Aggregation, Supervised Statistical Inference and Classification

Olivier Catoni Vincent Rivoirard Gilles Stoltz

Département de Mathématiques et Applications
Ecole Normale Supérieure
45 rue d'Ulm 75320 Paris Cedex 05

4 juin 2009 - Comité des projets
INRIA Rocquencourt

Membres fondateurs

Chercheurs seniors

- Olivier Catoni, DR CNRS, porteur du projet ;
- Gilles Stoltz, CR CNRS, professeur affilié à HEC ;
- Vincent Rivoirard, Mdc U. Paris Sud et ENS ;
- Gérard Biau, prof. U. P6 et ENS ;

Doctorant

- Sébastien Gerchinowitz (thèse entreprise en septembre 2008 sous la direction de Gilles Stoltz).

Membres fondateurs

Chercheurs seniors

- Olivier Catoni, DR CNRS, porteur du projet ;
- Gilles Stoltz, CR CNRS, professeur affilié à HEC ;
- Vincent Rivoirard, Mdc U. Paris Sud et ENS ;
- Gérard Biau, prof. U. P6 et ENS ;

Doctorant

- Sébastien Gerchinowitz (thèse entreprise en septembre 2008 sous la direction de Gilles Stoltz).

Membres fondateurs

Chercheurs seniors

- Olivier Catoni, DR CNRS, porteur du projet ;
- Gilles Stoltz, CR CNRS, professeur affilié à HEC ;
- Vincent Rivoirard, Mdc U. Paris Sud et ENS ;
- Gérard Biau, prof. U. P6 et ENS ;

Doctorant

- Sébastien Gerchinowitz (thèse entreprise en septembre 2008 sous la direction de Gilles Stoltz).

Membres fondateurs

Chercheurs seniors

- Olivier Catoni, DR CNRS, porteur du projet ;
- Gilles Stoltz, CR CNRS, professeur affilié à HEC ;
- Vincent Rivoirard, Mdc U. Paris Sud et ENS ;
- Gérard Biau, prof. U. P6 et ENS ;

Doctorant

- Sébastien Gerchinowitz (thèse entreprise en septembre 2008 sous la direction de Gilles Stoltz).

Membres fondateurs

Chercheurs seniors

- Olivier Catoni, DR CNRS, porteur du projet ;
- Gilles Stoltz, CR CNRS, professeur affilié à HEC ;
- Vincent Rivoirard, Mdc U. Paris Sud et ENS ;
- Gérard Biau, prof. U. P6 et ENS ;

Doctorant

- Sébastien Gerchinowitz (thèse entreprise en septembre 2008 sous la direction de Gilles Stoltz).

Membres fondateurs

Chercheurs seniors

- Olivier Catoni, DR CNRS, porteur du projet ;
- Gilles Stoltz, CR CNRS, professeur affilié à HEC ;
- Vincent Rivoirard, Mdc U. Paris Sud et ENS ;
- Gérard Biau, prof. U. P6 et ENS ;

Doctorant

- Sébastien Gerchinowitz (thèse entreprise en septembre 2008 sous la direction de Gilles Stoltz).

Membres fondateurs

Chercheurs seniors

- Olivier Catoni, DR CNRS, porteur du projet ;
- Gilles Stoltz, CR CNRS, professeur affilié à HEC ;
- Vincent Rivoirard, Mdc U. Paris Sud et ENS ;
- Gérard Biau, prof. U. P6 et ENS ;

Doctorant

- Sébastien Gerchinowitz (thèse entreprise en septembre 2008 sous la direction de Gilles Stoltz).

Thèmes principaux

Apprentissage supervisé

Entrées

Échantillon statistique $(X_i)_{i=1}^N$

$\longrightarrow$

Sorties

correspondantes $(Y_i)_{i=1}^N$

trouver un estimateur de régression $\widehat{f}$ qui minimise,

au sein d'un modèle $\mathcal{F}$,

un coût moyen :

$$\mathbb{E}\left[\ell\left(\widehat{f}(X), Y\right)\right] \simeq \inf_{f\in\mathcal{F}} \mathbb{E}\left[\ell\left(f(X), Y\right)\right]$$

Thèmes principaux

Prédiction séquentielle

Entrées

Échantillon statistique $(X_i)_{i=1}^N$

$\longrightarrow$

Sortie

X_{N+1}

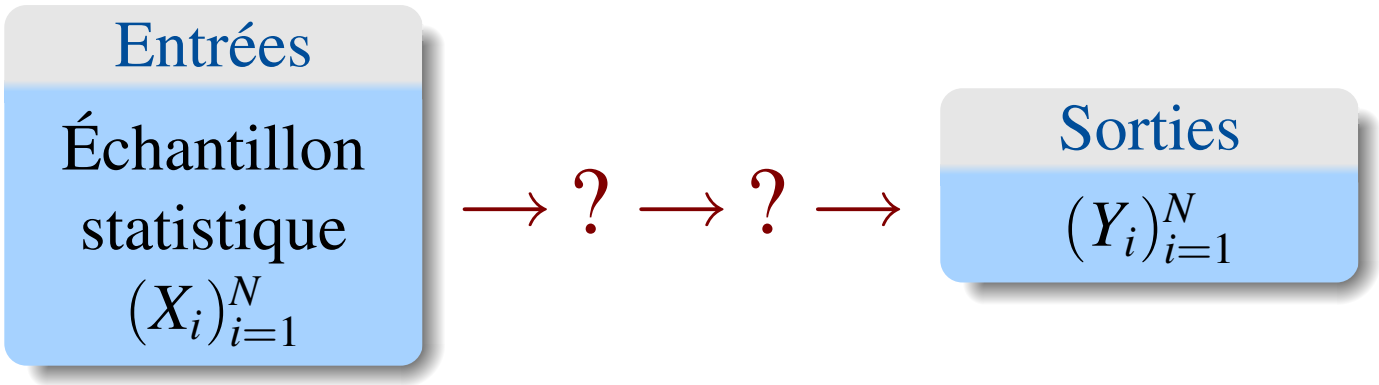
trouver un prédicteur $\widehat{f}$ qui minimise $\mathbb{E}\left[\ell\left(\widehat{f}(X_1,\ldots,X_N),X_{N+1}\right)\right]$

risque cumulé : minimiser $\mathbb{E}\left[\sum_{i=1}^N \ell\left(\widehat{f}(X_1,\ldots,X_{i-1}),X_i\right)\right]$

suites individuelles : minimiser $\sum_{i=1}^N \ell\left(\widehat{f}(X_1,\ldots,X_{i-1}),X_i\right)$.

Thèmes principaux

Apprentissage non supervisé



Quand les données sont compliquées
(images, ADN, signal de parole, ...)
des représentations intermédiaires
adaptées aux entrées
mais pas nécessairement guidées par les sorties
peuvent être nécessaires.

Thèmes principaux

Estimation de la densité :

une approche duale de la représentation

$$\inf_{\theta\in\Theta} -\mathbb{E}\{\log[p_\theta(X)]\} = \inf_{\theta\in\Theta} \mathcal{K}(\mathbb{P}, P_\theta) + \text{ cte.}$$

quand P_θ est loin de $\mathbb{P}$, on peut l'interpréter comme un code idéal, et donc une représentation implicite des entrées, au lieu d'y voir un estimateur de $\mathbb{P}$.

Modèles de mélange

$$p_\rho = \int \rho(d\theta) p_\theta, \text{ où } \rho \in \mathcal{F} \subset \mathcal{M}_+^1(\Theta).$$

Définit une loi jointe sur le couple (θ, x) et un changement possible de représentation de x par $\arg\max_\theta \frac{d\rho}{d\pi}(\theta) p_\theta(x)$ où π est une mesure de référence sur Θ.

Thèmes principaux

Bandits manchots

A chaque instant, on observe le gain pour une seule valeur du paramètre (un seul bras), que l'on peut choisir (problème de décision).

$$\frac{1}{N}\sum_{i=1}^{N} \int y \mathbb{P}_{\widehat{f}(Y_1,\dots,Y_{i-1})}(dy) - \max_{f\in\mathcal{F}} \int \mathbb{P}_f(dy),$$

$$\text{où } Y_i \sim \mathbb{P}_{\widehat{f}(Y_1,\dots,Y_{i-1})}$$

et $\mathcal{F}$ est un ensemble d'actions (bras) possibles.

Objectifs et méthodes

Comment traiter

des données complexes ?
images, signaux, séquences génomiques, ...

- entrées de grande dimension ;
- très structurées (interactions à plusieurs échelles, problèmes de segmentation et d'invariance).

Réponse :

Inégalités oracle non-asymptotiques :

$$\mathbb{E}\left\{\ell\left[Y,\widehat{f}(X)\right]\right\} \leq \inf_{f\in\mathcal{F}}\left\{\mathbb{E}\left\{\ell\left[Y,f(X)\right]\right\}+R(f,N)\right\}$$

(en espérance par rapport aux observations ou avec un niveau de confiance $1-\epsilon$ fixé.)

Objectifs et méthodes

Théorie du processus empirique

Contrôler $\inf_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} f(W_i) - \inf_{f\in\mathcal{F}} \mathbb{E}\big[f(W)\big]$,

où $W, W_1, \ldots, W_N$ est une suite de variables i.i.d..

Dépend de la taille de N, de la « taille » du modèle $\mathcal{F}$, de la structure des covariances, et plus précisément de $\mathbb{E}\big\{\big[f(W) - f^*(W)\big]^2\big\}$ où $f^* \in \arg\min \mathbb{E}\big[f(W)\big]$.

Objectifs et méthodes

Théorie du processus empirique

Contrôler $\inf_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} f(W_i) - \inf_{f\in\mathcal{F}} \mathbb{E}\big[f(W)\big]$,

où $W, W_1, \ldots, W_N$ est une suite de variables i.i.d..

Dépend de la taille de N, de la « taille » du modèle $\mathcal{F}$, de la structure des covariances, et plus précisément de $\mathbb{E}\big\{\big[f(W) - f^*(W)\big]^2\big\}$ où $f^* \in \arg\min \mathbb{E}\big[f(W)\big]$.

Objectifs et méthodes

Théorie du processus empirique

Contrôler $\inf_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} f(W_i) - \inf_{f\in\mathcal{F}} \mathbb{E}\big[f(W)\big]$,

où $W, W_1, \ldots, W_N$ est une suite de variables i.i.d..

Dépend de la taille de N, de la « taille » du modèle $\mathcal{F}$, de la structure des covariances, et plus précisément de $\mathbb{E}\big\{\big[f(W)-f^*(W)\big]^2\big\}$ où $f^* \in \arg\min \mathbb{E}\big[f(W)\big]$.

Objectifs et méthodes

Théorie du processus empirique

Contrôler $\inf_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} f(W_i) - \inf_{f\in\mathcal{F}} \mathbb{E}\big[f(W)\big]$,

où $W, W_1, \ldots, W_N$ est une suite de variables i.i.d..

Dépend de la taille de N, de la « taille » du modèle $\mathcal{F}$, de la structure des covariances, et plus précisément de $\mathbb{E}\big\{\big[f(W) - f^*(W)\big]^2\big\}$ où $f^* \in \arg\min \mathbb{E}\big[f(W)\big]$.

Objectifs et méthodes

- Il existe des réponses génériques à l'aide d'inégalités de concentration de la mesure, et de mesures de la taille ou complexité du modèle $\mathcal{F}$ — dimension de Vapnik, complexité de Rademacher, théorèmes PAC-Bayésiens, entre autres ;
- Il existe aussi des algorithmes efficaces, SVN, pénalisation ℓ_1 (calibrée), mélange factorisé d'arbres de décision, permettant de réaliser une minimisation structurelle du risque (c'est-à-dire d'être adaptativement optimal dans une famille de sous-modèles de $\mathcal{F}$).

Exemple (collaboration avec Willow)

$$\mathcal{F} = \left\{ f_\theta = \sum_{j=1}^{d} \theta_j \phi_j ; \theta \in \Theta \subset \mathbb{R}^d \right\}, \Theta \text{ convexe fermé}$$

$$\widehat{f} \in \arg\min_{f_\theta \in \mathcal{F}} r(f_\theta) + \lambda \|\theta\|^2, \quad r(f) = \sum_{i=1}^{N} \left[Y_i - f(X_i) \right]^2,$$

$$\widetilde{f} \in \arg\min_{f_\theta \in \mathcal{F}} R(f_\theta) + \lambda \|\theta\|^2, \quad R(f) = \mathbb{E}\left[r(f) \right],$$

$$D = \sum_{j=1}^{d} \frac{\nu_j}{\nu_j + \lambda} \mathbb{1}(\nu_j > 0) \leq d, \text{ où } Q = \mathbb{E}\left[\phi(X) \phi(X)^T \right],$$

Exemple (collaboration avec Willow)

$$\text{Supposons } \mathbb{E}\left[\|\phi(X)\|^4\right] < +\infty,$$

$$\text{et } \mathbb{E}\left\{\|\phi(X)\|^2\left[\widetilde{f}(X)-Y\right]^2\right\} < +\infty.$$

Pour tout $\epsilon > 0$, il existe N_ϵ t.q. pour tout $N > N_\epsilon$ avec proba $1-\epsilon$

$$R(\widehat{f}) + \lambda\|\widehat{\theta}\|^2 \leq R(\widetilde{f})$$

$$+\lambda\|\widetilde{\theta}\|^2 + \text{ess sup}\mathbb{E}\left\{\left[\widetilde{f}(X)-Y\right]^2 | X\right\}\frac{30D + 1000\log(3/\epsilon)}{N}.$$

Exemple (collaboration avec Willow)

$$\text{Supposons} \quad \sup_{f_1,f_2\in\mathcal{F},x\in\mathcal{X}} |f_1(x)-f_2(x)| \leq H,$$

$$\text{ess sup}\mathbb{E}\left\{ \left[Y-\widetilde{f}(X)\right]^2 | X \right\} \leq \sigma^2 < +\infty.$$

Avec probabilité $1-\epsilon$, pour un estimateur modifié, (dans le cas $\lambda = 0$),

$$R(\widehat{f}) - R(\widetilde{f}) \leq (2\sigma + H)^2 \frac{8,3d + 12,5\log(2/\epsilon)}{N}.$$

Pourquoi une équipe INRIA ?

- Vis à vis du DMA : promouvoir le lien entre statistique et algorithmique auprès des élèves mathématiciens de l'ENS en affichant une structure dépendant d'un organisme dont l'interface mathématique/informatique est la vocation ;
- Vis à vis de l'INRIA : alimenter les nombreux projets INRIA traitant d'apprentissage statistique en proposant
  - de nouveaux résultats mathématiques donnant des garanties sur le comportement de modèles et de méthodes efficaces en pratique ;
  - de nouveaux modèles et de nouvelles méthodes issus d'une réflexion mathématique sur le contrôle du risque.

L'apprentissage statistique dans les projets INRIA

projets partenaires

- Willow, (Paris) projet imagerie et apprentissage avec lequel nous avons des collaborations, des projets en communs ainsi qu'une feuille de route pour le futur ;
- CLIME, (Paris) projet consacré à l'analyse et à la prédiction de données environnementales. Collaboration concernant l'agrégation d'experts dans le cadre des suites individuelles ;
- SequeL, (Lille), projet dédié principalement à l'apprentissage séquentiel, collaboration concernant les bandits et les algorithmes de décision/prédiction ;
- Select, (Saclay) projet dédié à la sélection de modèles, séminaire commun d'apprentissage statistique bimensuel à l'ENS ;

L'apprentissage statistique dans les projets INRIA

projets reliés au thème

- TAO (Saclay) optimisation et apprentissage. Optimisation stochastique, apprentissage statistique, processus de décision et de contrôle (dont les problèmes de bandit). Olivier Teytaud (TAO) et Gilles Stoltz sont membres de l'ANR dirigé par Remis Munos (Sequel) sur le sujet.
- IMEDIA (Paris) reconnaissance des formes appliquée à l'indexation d'images par le contenu. Modèles hiérarchiques et stratégies « coarse to fine » apparentées à la statistique séquentielle. Agrégation convexe d'arbres de décisions.
- LEAR, (Grenoble) interprétation de scènes et reconnaissance des formes, méthodes de noyaux ;

L'apprentissage statistique dans les projets INRIA

projets reliés au thème

- Ariana (Sophia Antipolis) modèles probabilistes et variationnels en imagerie : champs de Markov, ondelettes, géométrie stochastique. Application aux images aériennes et satellites ;
- MISTIS (Grenoble) modèles de mélange et modèles de Markov cachés. Indexation d'images par le contenu avec LEAR ;
- AVIZ (Saclay) visualisation d'images, clustering, sélection de variables. Classification de grandes bases de données par SVN ;

L'apprentissage statistique dans les projets INRIA

projets reliés au thème

- SYMBIOSE (Rennes) étude du génome à grande échelle ;
- TEXMEX (Rennes) indexation multimédia ; descripteurs d'images, analyse statistique de textes, fouille de données supervisée, arbres de décision et SVM ;
- MOSTRARE (Lille) fouille de données sur internet et structures d'arbres.

Conclusion : l'analyse statistique de données multimédia (textes, images, vidéos) est fortement représentée à l'INRIA. La conception et l'étude de méthodes statistiques adaptées aux données complexes, thème central de l'équipe CLASSIC, viendrait compléter de façon cohérente cet effort.

Applications : qualité de l'air

Agrégation de prédicteurs séquentiels

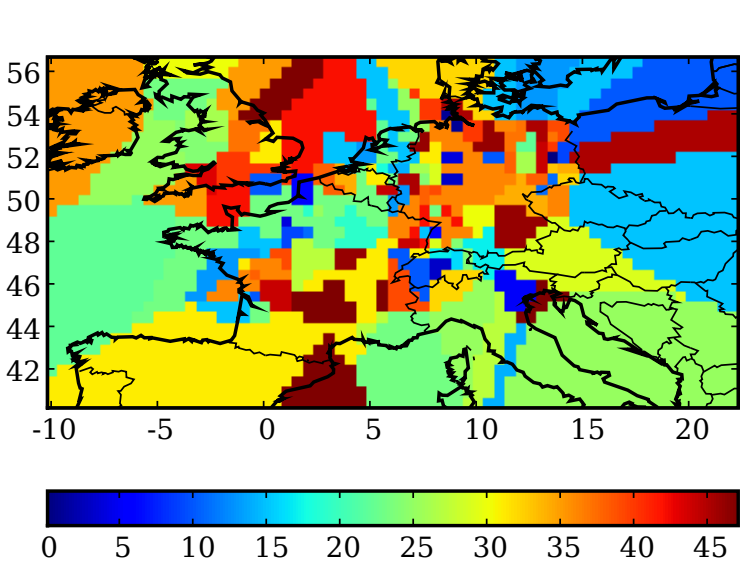
A partir de N prédicteurs, on peut former un modèle en considérant soit leurs combinaisons convexes, soit leurs combinaisons linéaires. On dispose d'algorithmes assurant que la perte cumulée est presque optimale pour toute séquence de données.

Qualité de l'air (avec CLIME, depuis juillet 2005)

On veut prédire, jour après jour, les hauteurs des pics d'ozone du lendemain, en combinant 48 experts, solutions d'EDP.

Applications : qualité de l'air

Les figures ci-dessous montrent que tous les experts sont utiles et apportent de l'information.



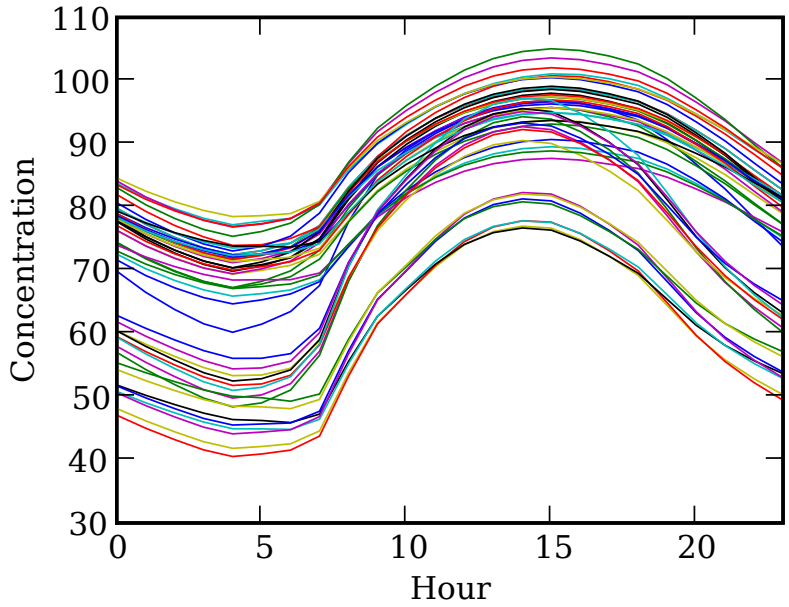


FIG.: A gauche : Coloration de l'Europe en fonction de l'indice du meilleur expert local. A droite : Profils moyens de prédiction sur une journée (moyennes spatiales et temporelles, en $\mu g/m^3$).

Applications : qualité de l'air

Performances (pertes quadratiques moyennes) sur les données

EG	EG esc.	Ridge	Ridge esc.
21.47	21.31	20.77	19.45

vs.

Moyenne	M. fondamental	M. convexe	M. linéaire
24.41	22.43	21.45	19.24

Si l'on regarde comment les algorithmes prédisent, on voit qu'ils ne se contentent pas de mettre un poids constant très grand au meilleur expert global : les poids varient rapidement et souvent.

Applications : consommation électrique

En collaboration avec EDF R&D, depuis février 2009

Le système de prédiction Eventail nécessite 700 paramètres. (Prédiction de la consommation du jour à venir heure par heure). Certains jeux de paramètres sont bons pour les week-ends, d'autres pour l'été, d'autres pour les vacances, etc.
On construit alors des experts qui seront inactifs à certains pas de temps en fonction des paramètres externes. A chaque tour, on agrège ainsi un nombre variable d'experts.

Problèmes :

- Construire de bons experts (tirer parti de leur spécialisation)
- Adapter les algorithmes d'agrégation à ce cadre

Applications : analyse d'images

Un banc d'essai pour l'apprentissage non supervisé

L'indexation automatique d'images pose des problèmes de représentation, en particulier en ce qui concerne

- la segmentation des zones d'intérêt ;
- la prise en compte des invariants (changements d'échelle, de point de vue, d'illumination ...).

Applications : analyse d'images

Représentation adaptative invariante par transformations projectives

- première étape (validée) : utilisation d'une représentation implicite par des codes idéaux (modèle de mélange de probabilités) pour la détection multi-échelles des contours orientés.
- deuxième étape (en projet) : modèles probabilistes de configurations de contours invariants par transformations projectives. Apprentissage non supervisé de configurations typiques sur des bases de données d'exemples.

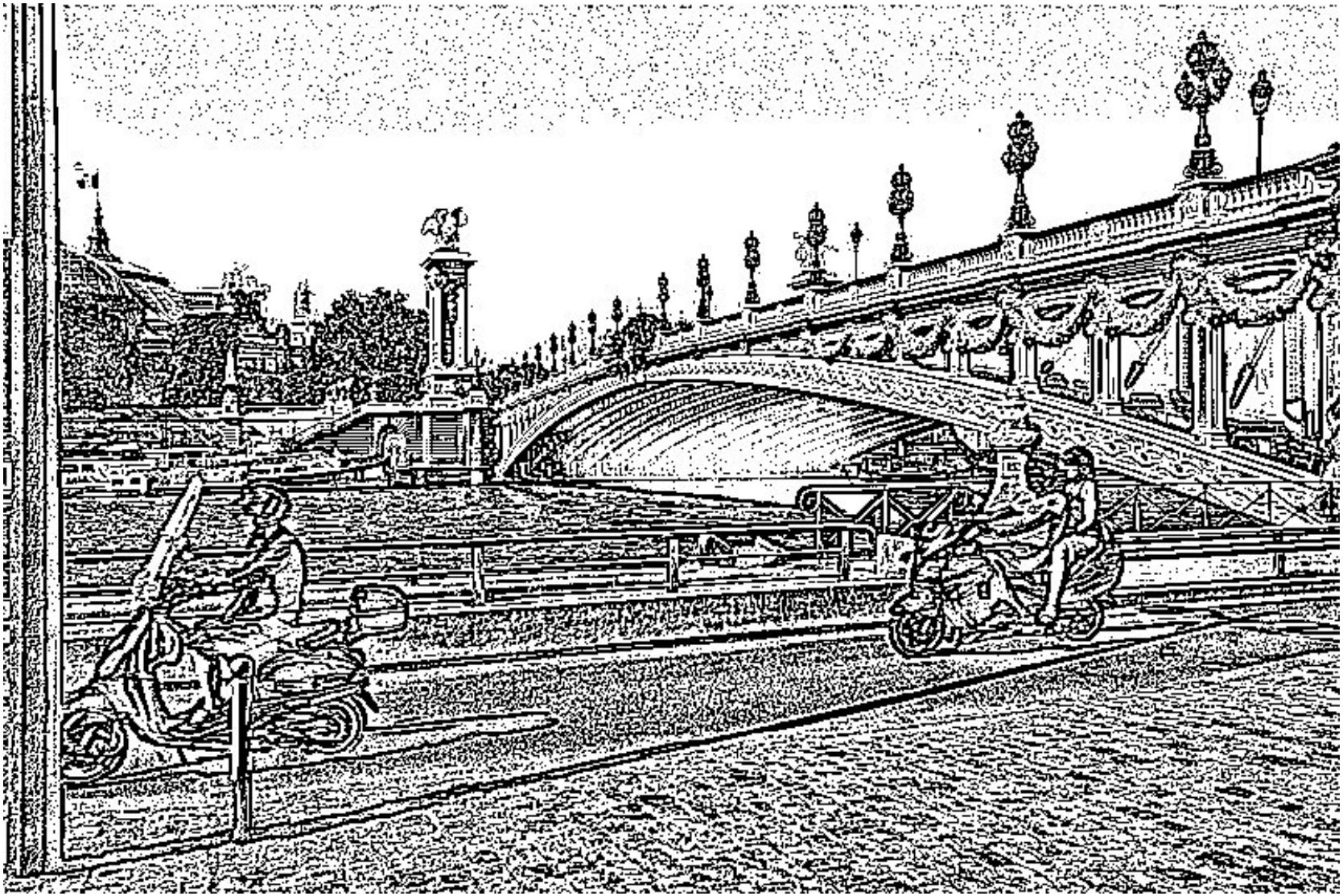
Applications : analyse d'images

Applications : analyse d'images

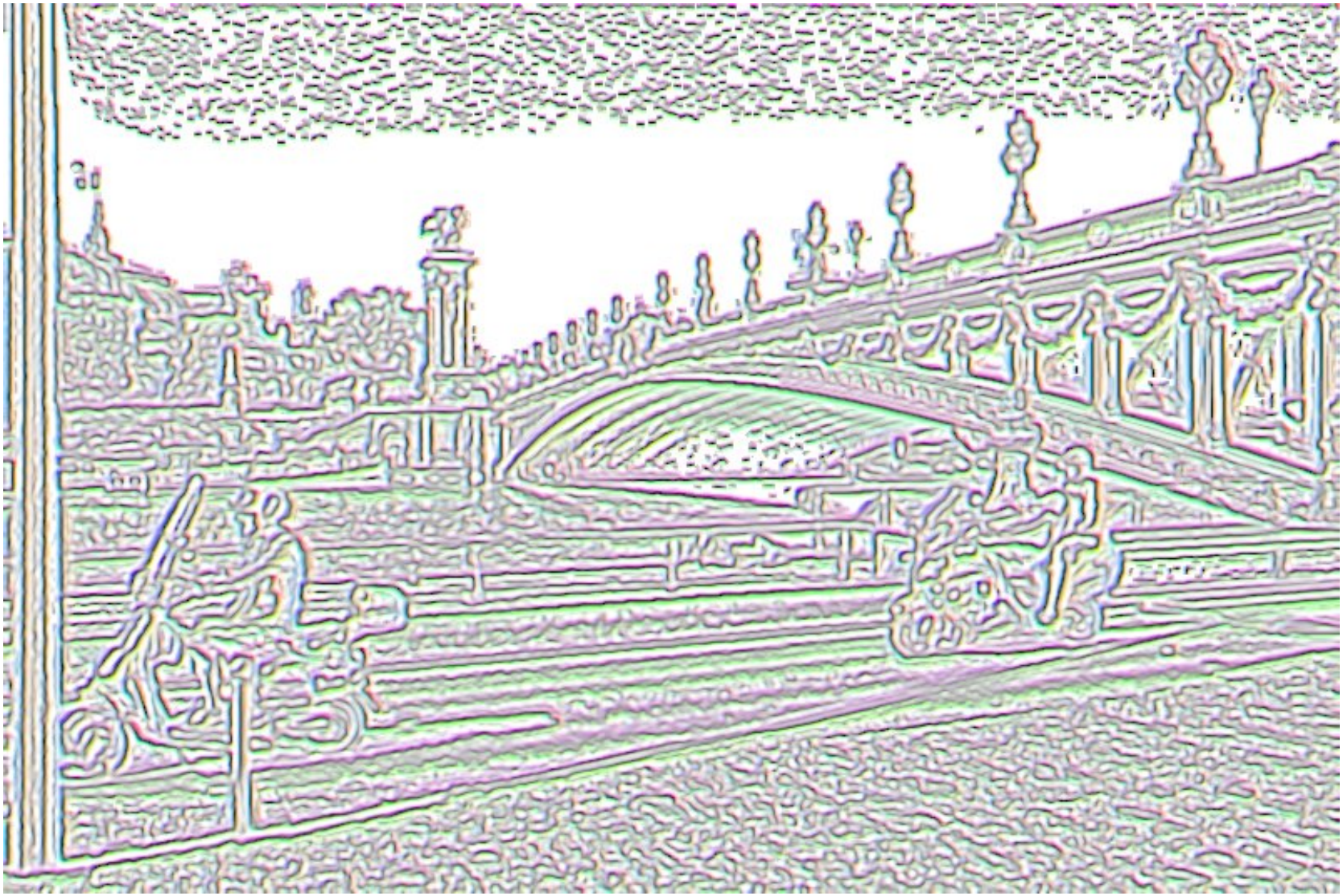
Applications : analyse d'images

Applications : analyse d'images

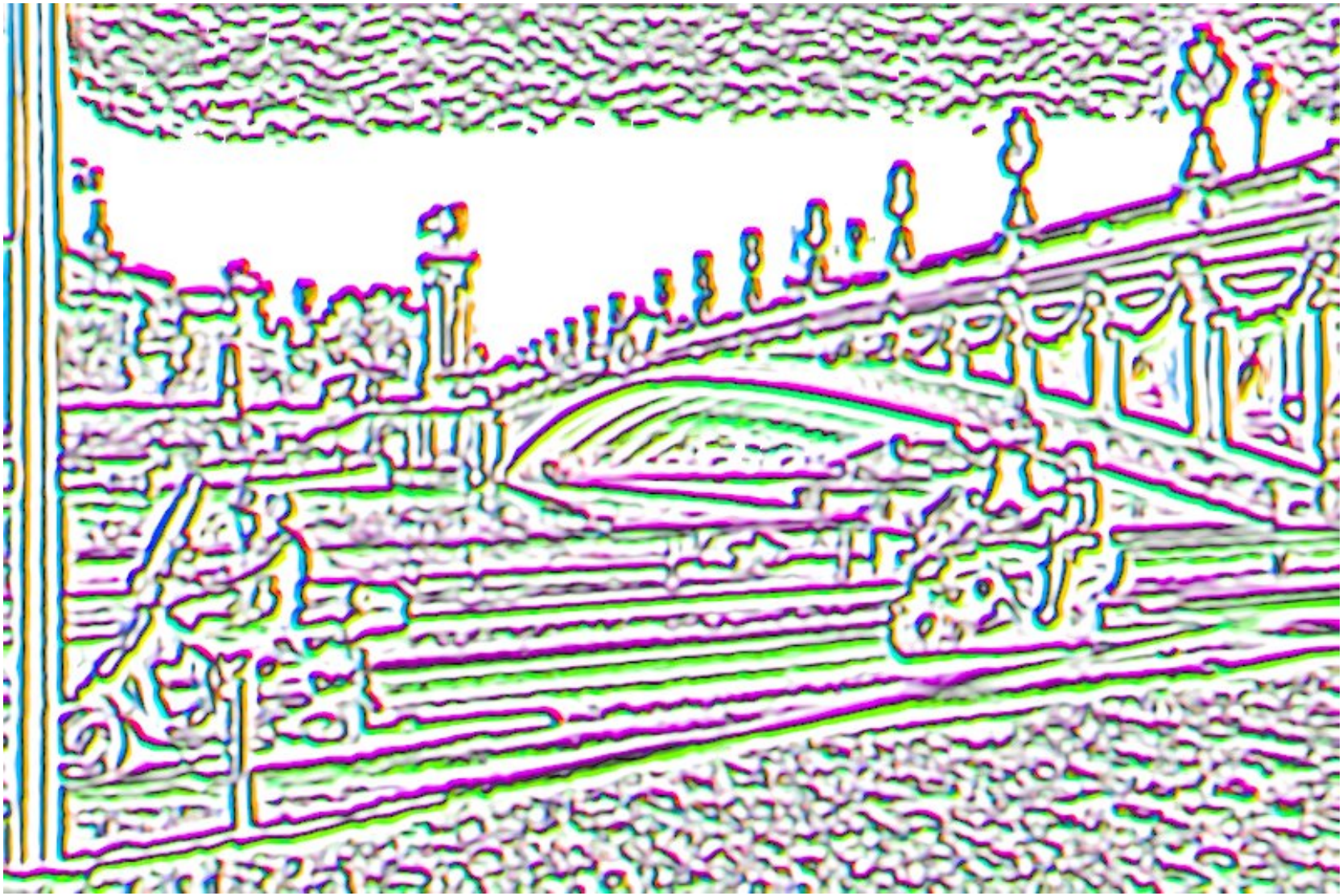
Applications : analyse d'images



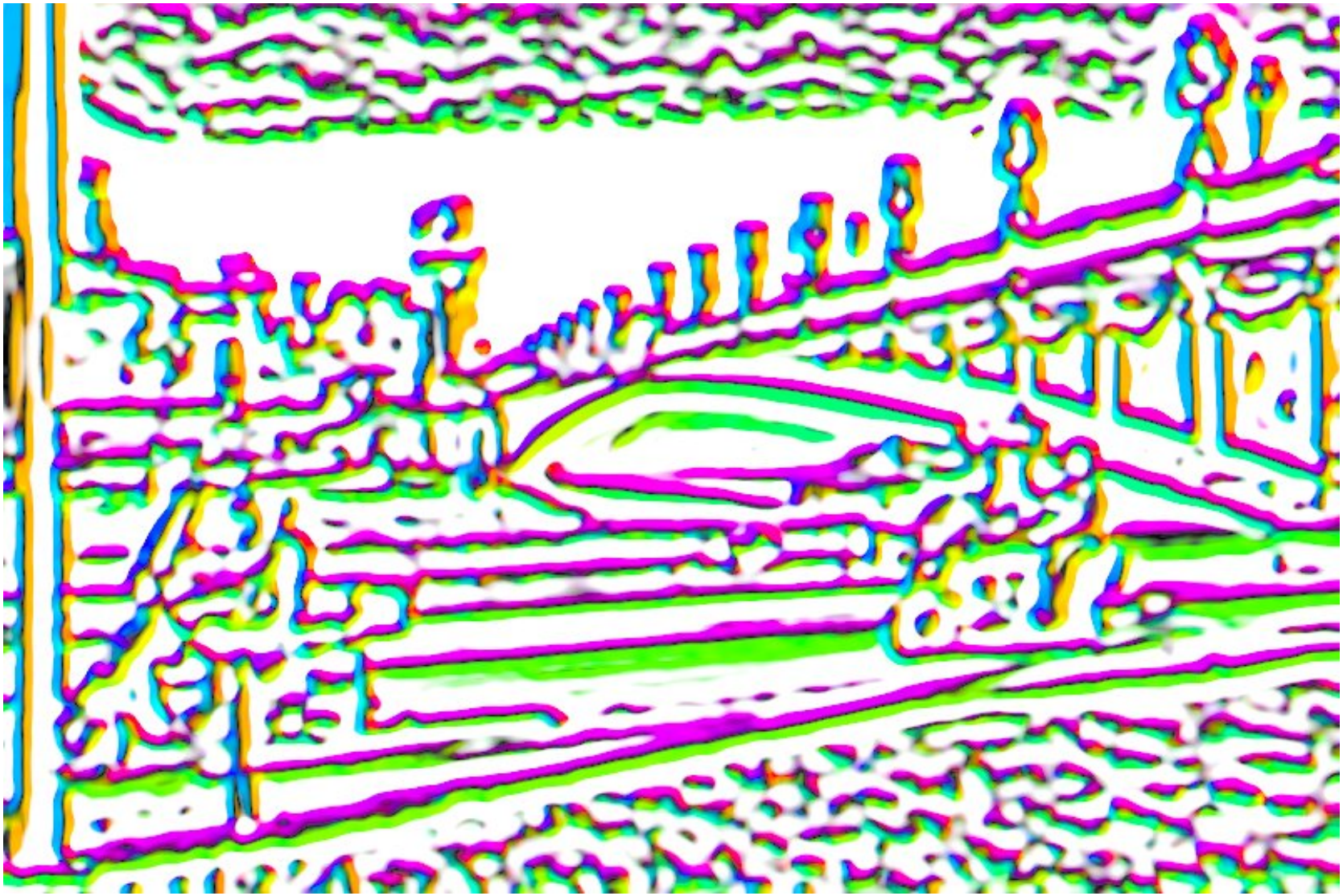
Applications : analyse d'images



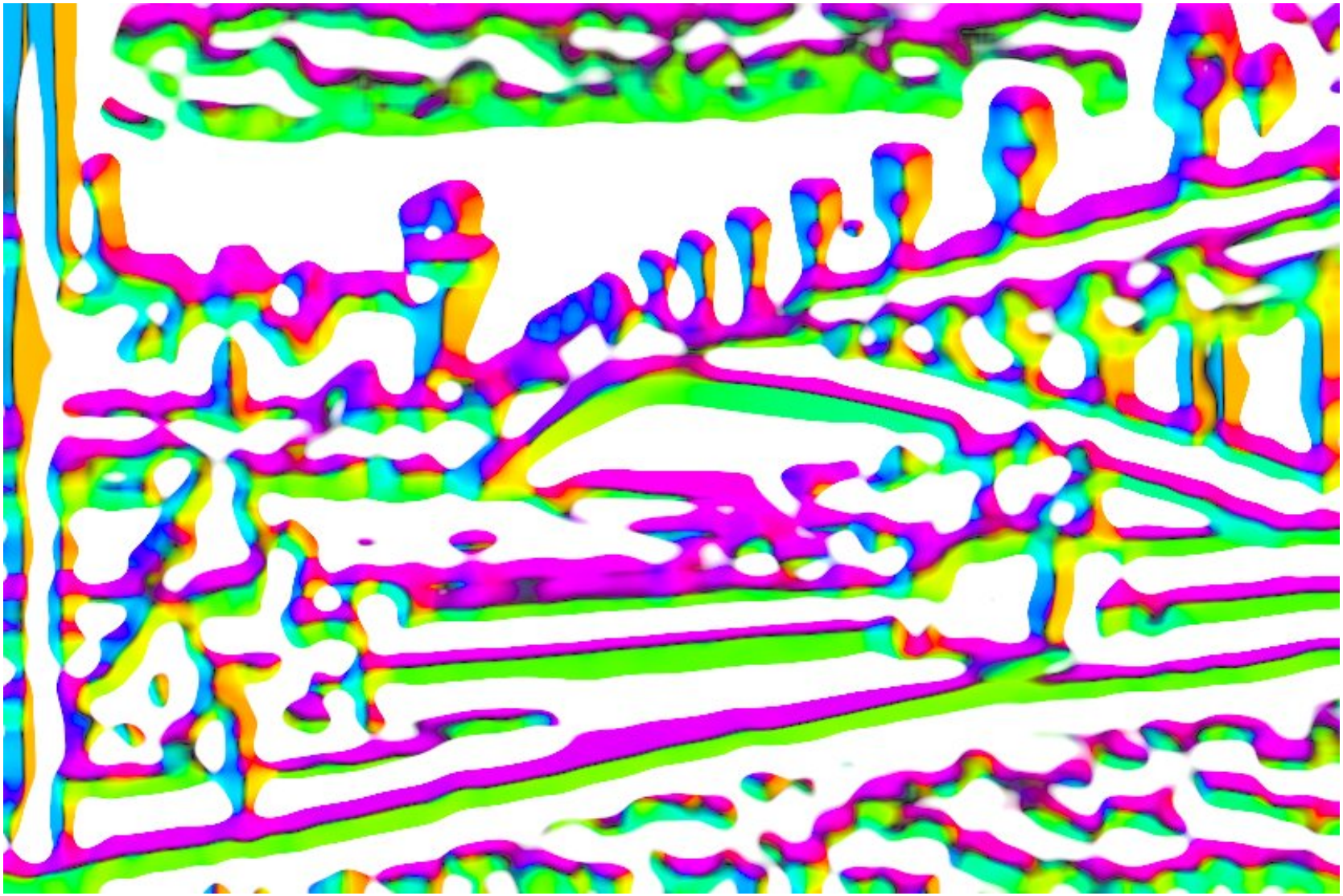
Applications : analyse d'images



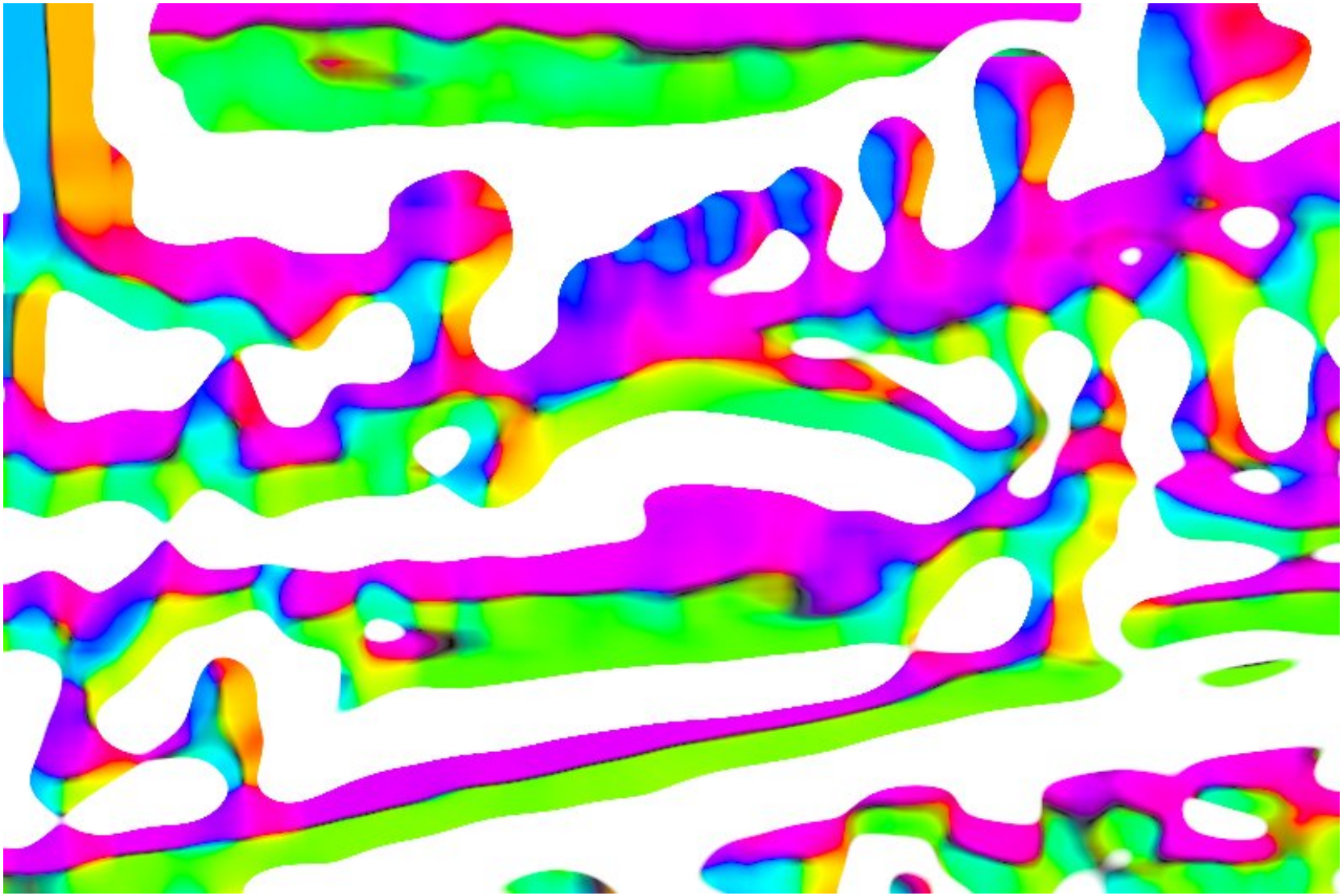
Applications : analyse d'images



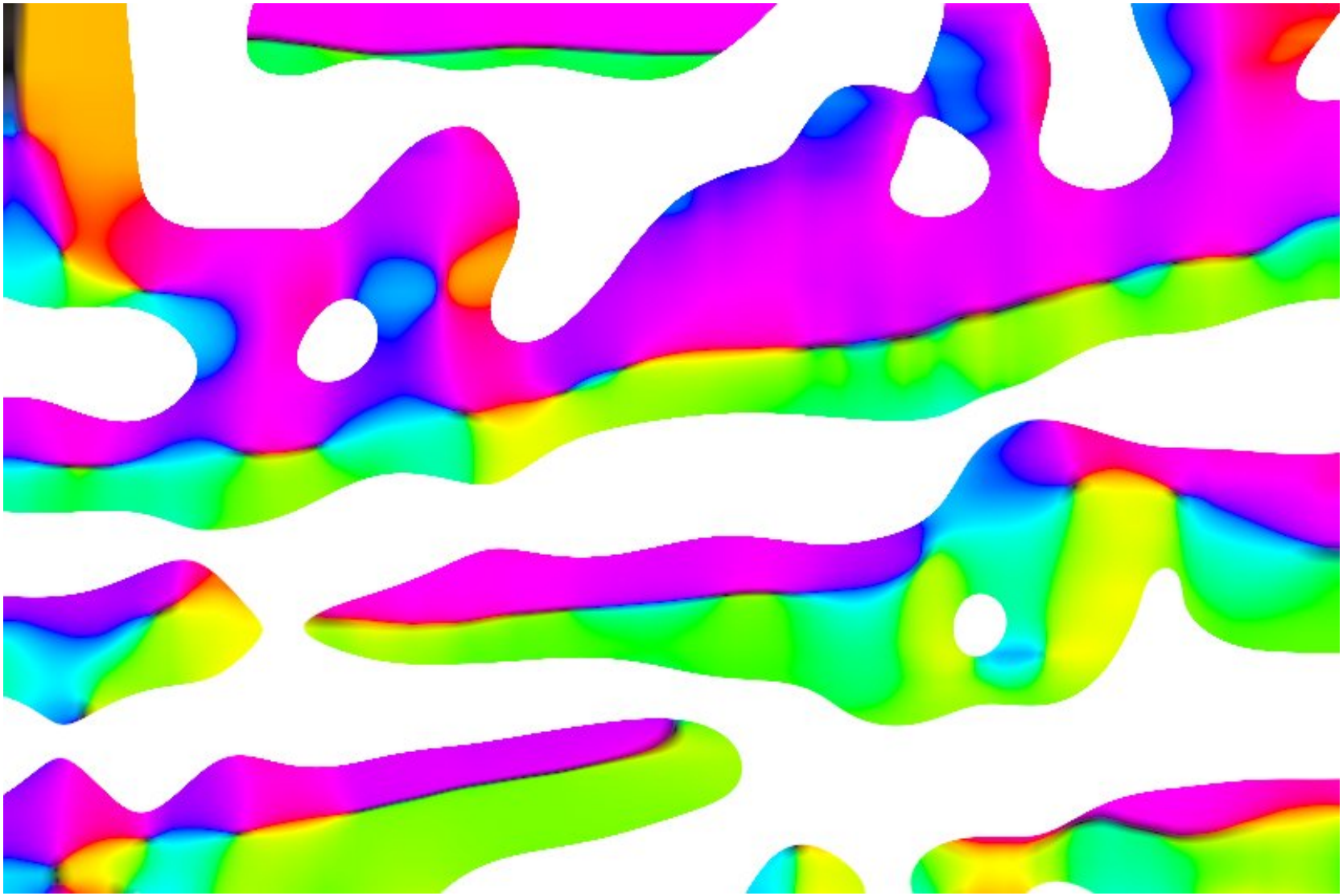
Applications : analyse d'images



Applications : analyse d'images



Applications : analyse d'images



Applications : analyse d'images