Un moyen d'identification simple : les empreintes digitales(*)

Hervé Lehning(**)

Le positivisme scientifique n'a pas que des défauts. Depuis la fin du XIX^e siècle, la police fait place aux preuves de nature scientifique. La première d'entre elles est fondée sur les empreintes digitales. Est-elle fiable et comment faire pour consulter les millions d'empreintes dont dispose la police ?

La première énigme policière résolue en comparant des empreintes digitales recueillies sur la scène d'un crime avec celles contenues dans un fichier préétabli date de 1892. L'histoire se déroule en Argentine et peut être retrouvée par ceux qu'elle intéresse sur l'Internet. Au préalable, Sir Francis Galton – cousin de Charles Darwin – (voir l'annexe I) avait établi l'unicité et la permanence de ces figures cutanées (voir l'annexe II). On attendra dix ans pour que la même histoire se produise en France où l'inventeur de la méthode se nomme Alphonse Bertillon.

Fiche signalétique de Henri Léon Scheffer,
premier criminel à être identifié par ses empreintes digitales par la police française.

Premier fichier français

Dans le premier fichier français d'Alphonse Bertillon, chaque doigt se voyait attribuer un chiffre entre 0 et 8 et donc chaque individu une signature de dix chiffres comme par exemple 13267 – 23486. Ce système permettait d'accélérer la recherche dans ce fichier en ne retenant que les fiches ayant une signature convenable. Depuis 1896, les détenus parisiens étaient ainsi tous fichés. En 1902, alors qu'ils sont

(*) Cet article fait suite à celui de Rémi Belloeil.

(**) lehning@noos.fr

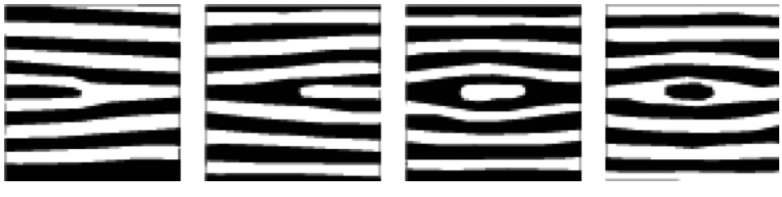
appelés sur la scène d'un crime, les policiers trouvent comme seules traces quatre empreintes digitales sur une vitrine en verre (un pouce sur la face extérieure, un index, un majeur et un annulaire sur la face intérieure). Alphonse Bertillon les agrandit quatre fois, puis les compare avec celles de son fichier constitué par celles des détenus parisiens. En quelques jours, il trouve qu'elles coïncident avec celle d'un malfaiteur – Henri Léon Scheffer – condamné pour vol et abus de confiance. Arrêté à Marseille six jours plus tard, il passe aux aveux. Cette affaire Scheffer constitue le véritable acte de naissance de la police scientifique en France.

Comment chercher dans un gros fichier ?

Bertillon a réussi à retrouver un criminel grâce à son fichier car celui-ci était de taille modeste. De nos jours, le fichier de la police contient les empreintes de plus d'un million et demi d'individus. Comme chacun en possède dix, cela donne quinze millions d'empreintes ! Comment chercher dans un tel fichier ? À la main, c'est impossible. La recherche doit être confiée à un ordinateur. Quel algorithme doit-il suivre pour ce faire ? L'idée la plus simple est de tenter de superposer l'empreinte trouvée avec celles du fichier en commençant par la première jusqu'à la dernière. Le problème est beaucoup plus compliqué qu'il ne paraît à première vue : l'empreinte trouvée est forcément partielle et déplacée (décalée ou tournée). Même si l'on imagine une orientation préalable des empreintes trouvées, il est nécessaire de les comparer avec l'empreinte du fichier et des transformations de celle-ci par certains déplacements. Sans entrer dans les détails, il est facile de voir que cette voie est beaucoup trop lourde pour être efficace.

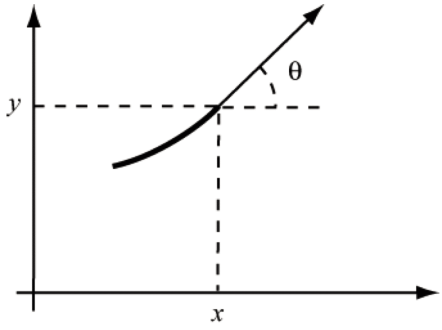
Signature d'une empreinte

Une idée simple est d'oublier la majorité des caractéristiques d'une empreinte pour se concentrer sur les points les plus typiques que l'on nomme *minuties*. Il s'agit de bifurcations, d'îles, de nœuds, de lignes qui disparaissent, etc. On peut en relever seize types dont les quatre plus importantes sont : terminaison, bifurcation, île et lac:



Principales minuties : terminaison, bifurcation, île et lac.

La description de chaque minutie demande en moyenne seize octets. Pourquoi ? Oublions pour un instant le problème de la définition d'un repère intrinsèque à l'empreinte (voir l'annexe III). Une minutie est caractérisée par un type (terminaison, bifurcation, île ou lac) et trois ou quatre nombres réels suivant son type. Dans les deux premiers cas, trois nombres suffisent : ses coordonnées et un angle (voir la figure ci-dessous), dans les deux autres, un de plus est nécessaire.



Repérage d'une terminaison par trois nombres.

Le type demande un seul octet pour être stocké et chacun des nombres, quatre. Il s'agit de la norme IEEE des nombres réels en simple précision. En tout, il faut donc treize octets pour stocker une terminaison ou une bifurcation, dix-sept pour une île ou un lac.

La question s'alourdit quand on remarque que, sur une empreinte digitale, on peut repérer un grand nombre de minuties comme le montre la figure *sur cette empreinte*.

Sur cette empreinte, quelques minuties ont été soulignées par un cercle.
Selon la loi française, douze minuties sont nécessaires
pour caractériser une empreinte digitale.

Comment automatiser leur repérage puis leur stockage afin d'obtenir une signature de l'empreinte ? Pour cela, elle est d'abord stockée sous un format de photographie numérique (comme le format JPEG) puis filtrée de façon à ce que les lignes aient toutes la même épaisseur (un pixel). On obtient ainsi une image squelettique en noir

et blanc. Les minuties sont alors extraites. À ce stade, on en détecte normalement une centaine. Comme chaque minutie demande en moyenne seize octets pour être décrite, ce nombre est beaucoup trop grand. Arrivés à ce niveau, on ne conserve en fait que les quinze plus fiables ce qui donne une signature de 240 octets. Pour ce faire, parmi les minuties proches l'une de l'autre, on ne garde que la plus marquée. Les minuties retenues sont alors classées dans l'ordre des nombres qui les représentent, le tout est regroupé dans un nombre de 240 octets ; L'important est que les algorithmes utilisés permettent ainsi de n'associer à chaque empreinte digitale qu'une seule et même signature. En revanche, il est possible que plusieurs empreintes possèdent la même. On trouvera une étude de ces minuties ainsi que leurs modélisations dans *les empreintes digitales* [4]. Cette propriété peut avoir des conséquences fâcheuses si aucune vérification humaine n'est faite comme le montre le cas de Shirley McKie (voir l'annexe II).

Structure du fichier

Le fichier d'empreintes digitales comprend trois parties : celui des signatures, celui des empreintes proprement dites et celui des biographies des individus ainsi recensés. Chaque signature renvoie à une ou plusieurs empreintes qui elles-mêmes renvoient chacune à un individu. La recherche se fait dans le fichier des signatures. En moyenne, vu la taille du fichier, on trouve ainsi quelques dizaines d'empreintes possibles. Le reste de l'analyse se fait de façon traditionnelle c'est-à-dire par un spécialiste humain qui seul peut établir la concordance des empreintes. Dans cette analyse, l'ordinateur ne sert donc qu'à un tri préalable.
Revenons à cette étape informatique. Étant donnée une empreinte digitale trouvée sur la scène d'un crime ou ailleurs, on calcule sa signature de la façon décrite précédemment et on la recherche dans le fichier. Il s'agit ainsi de trouver un mot de 240 octets dans un fichier de plusieurs millions d'articles. S'il n'est pas structuré, la seule méthode possible consiste à essayer toutes les fiches les unes après les autres. On ne peut pas même s'arrêter à la première réponse car deux empreintes distinctes peuvent avoir la même signature. Quinze millions de signatures doivent donc être essayées. Pour accélérer cette recherche, une idée simple est de trier le fichier comme l'est un dictionnaire. La méthode dichotomique permet alors de limiter les essais à 24 ! Pourquoi ? Tout simplement parce que chaque essai permet de diviser par deux le nombre de fiches possibles (voir l'annexe IV).

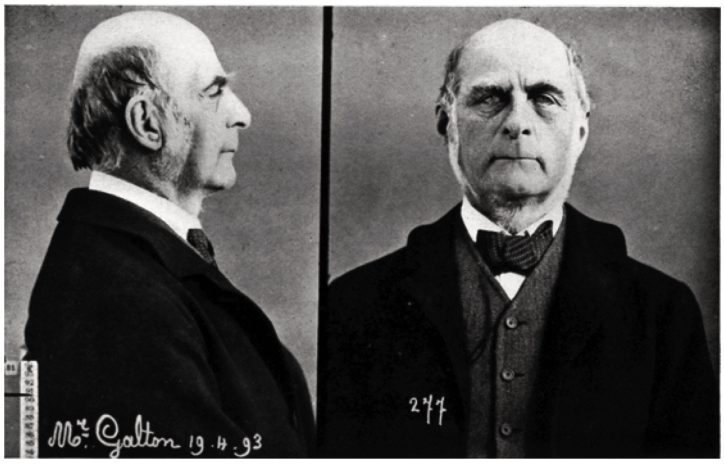
Mise à jour du fichier

Il reste une difficulté : celle de la mise à jour du fichier. Il est très facile d'ajouter de nouvelles fiches à un fichier non trié : il suffit de les mettre à la queue. Il est plus difficile d'en ajouter à un fichier trié. Pour cela, il faut d'abord trouver la place où insérer la nouvelle fiche, puis décaler les suivantes pour placer la nouvelle. Cela donne plusieurs millions de décalages nécessaires. Il est tout aussi long de supprimer une fiche. Pour tourner la difficulté, on a mis au point plusieurs façons de structurer les fichiers. Pour la plupart, elles utilisent la notion d'arbre et donc de graphe (voir les arbres de recherche dans [2], pages 165 – 168).

Problème d'identification

La même technique est utilisée pour identifier automatiquement les personnes ayant le droit d'accéder à certaines zones ou à certains services. Elle permet de remplacer avantageusement la méthode des mots de passe. En effet, ceux-ci peuvent être obtenus par la menace ou l'astuce. Voler un doigt reste possible c'est pourquoi un détecteur contrôle également la circulation sanguine. Dans cette application de la méthode, le fichier est réduit mais la même technique des signatures est employée. On utilise normalement les quinze minuties les plus pertinentes, le risque d'erreur est alors très faible.

Annexe I : Francis Galton



Sir Francis Galton, 1822 – 1911, en 1893. Un peu d'humour dans un monde policé.

Outre ses travaux fondateurs sur les empreintes digitales, Francis Galton est connu pour ses travaux sur l'intelligence humaine. D'autre part, ses idées ont fortement influencé le développement des statistiques. Son introduction du terme « eugénisme » lui valut de voir son nom ensuite très mal associé puisque, pour Adolf Hitler, cette idée justifiait le massacre des juifs et des tziganes entre autres. Inutile de préciser que cette association est tout à fait abusive.

Annexe II : Probabilité d'identité des empreintes

Bien qu'il eut des précurseurs, Francis Galton (1822 – 1911) fut le premier à établir l'unicité et la permanence de ces figures cutanées. De plus, il proposa de les classifier suivant leurs formes : boucles, volutes ou arcs. En dénombrant ces formes, il arrivait de façon assez obscure à 2^{36} configurations d'empreintes possibles. En les supposant équiprobables, cela donne une chance sur 64 milliards que deux doigts distincts aient les mêmes empreintes. Depuis Francis Galton, d'autres méthodes de calcul ont été proposées. Ainsi, Christophe Champod [1] propose un modèle probabiliste où

plusieurs facteurs sont supposés indépendants. Par une étude statistique, il estime la probabilité de chaque facteur. La multiplication des probabilités de ces facteurs permet de conclure. L'étude de Christophe Champod conforte les règles d'identification édictées par Edmond Locard, directeur du laboratoire de police scientifique de Lyon, en 1914 : absence de discordance, existence de 12 points concordants.

Shirley McKie, victime d'une confusion d'empreintes digitales.

Ceci dit, ces notions ne doivent pas aboutir à un quelconque automatisme. Par exemple, en 1997, Shirley McKie, enquêtrice de la police écossaise, fut accusée d'un meurtre parce que ses empreintes digitales avaient été « identifiées » sur la scène d'un crime. En fait, elles n'étaient que quasiment identiques à celles du véritable meurtrier (voir the price of innocence [3]).

Annexe III : Définition d'un repère intrinsèque à une empreinte.

Pour être identifiable quelque soit l'orientation de l'empreinte, il est important que les coordonnées de chaque minutie soient calculées dans un repère lié à l'empreinte et non à sa position. Cela demande de définir le centre du repère et l'orientation de l'axe des abscisses. Pour le centre, le problème est relativement simple. Il suffit de chercher le point de convergence des circonvolutions au centre du doigt. Pour cela, un algorithme consiste à définir des lignes orthogonales à celles-ci et à trouver leur point de concours. Il nous suffit alors d'un autre point lié à l'empreinte pour définir un repère intrinsèque. Par exemple, on peut choisir le centre de gravité des quinze minuties retenues.

Annexe IV : Recherche dichotomique

Imaginons que nous cherchions une fiche parmi 100 fiches triées dans l'ordre alphabétique. Nous ne le savons pas à l'avance bien sûr mais la fiche cherchée « brigand » se trouve être la 10^{e} du fichier. Nous essayons d'abord la 50^{e}, « escroc ».

La fiche cherchée se trouve précéder celle-ci. Nous essayons alors la 25e puis la 12e, la 6e, 9e, la 11e et nous trouvons enfin la 10e. Ainsi, en sept essais, nous avons trouvé la fiche cherchée. Remarquez qu'il en sera toujours ainsi. Cela tient au fait que $2^7 = 128 > 100$ puisqu'à chaque fois, nous réduisons par deux le nombre de fiches à essayer. Par rapport à la méthode consistant à essayer toutes les fiches, le gain est considérable : sept fiches à tester au lieu de cent ! Si nous cherchons dans un fichier de quinze millions de fiches, le gain est encore plus impressionnant puisque le nombre d'essais est le plus petit nombre n tel que $2^n > 15.10^6$ c'est-à-dire 24.

Références

[1] Christophe Champod, *Reconnaissance automatique et analyse statistique des minuties sur les empreintes digitales*, Université de Lausanne, 1995.

[2] Hervé Lehning, *Questions de maths sympas pour M. et Mme Toutlemonde*, Ixelles éditions, 2011.

[3] Iain McKie et Michael Russel, Shirley McKie : *The price of innocence*, Birlinn, 2007.

[4] Véronique Messéant, Patrick Nizou et Nathalie Villain, *Les empreintes digitales*, mémoire de master de didactique, Université Paris VII, 2006.