

L'idée d'utiliser des points d'intérêt locaux remonte aux travaux de Hans Moravec en 1981 sur la recherche de correspondance entre images stéréoscopiques<sup>5,6</sup> et aux améliorations apportées en 1988 par Harris et Stephens<sup>5,7</sup>. À la suite de quoi, différents types de caractéristiques sont proposés<sup>1</sup>, tels que des segments de lignes<sup>8</sup>, des groupements d'arêtes<sup>9,10</sup> ou de zones<sup>11</sup>. En 1992, l'intérêt de ce type de détecteur est confirmé par les travaux de Harris<sup>12</sup> et le descripteur de coins qu'il propose, améliorant la plupart des défauts du détecteur de Moravec, va connaître un important succès. Il faut cependant attendre 1997 et le travail précurseur de Cordelia Schmid et Roger Mohr pour établir l'importance, dans le domaine de la vision, des caractéristiques locales invariantes appliquées au problème général de détection et de recherche de correspondance<sup>5,13</sup>. Si le descripteur qu'ils développent (et qui s'appuie sur le détecteur de coins de Harris) apporte l'invariance à la rotation, il reste cependant sensible aux changements d'échelle, d'angle d'observation et d'exposition<sup>14</sup>. Lowe comblera grandement ces défauts avec son descripteur SIFT<sup>15</sup>.



Exemple de caractéristiques SIFT. Chaque point-clé est représenté par une flèche dont la direction correspond à la direction principale associée et la norme au facteur d'échelle (illustration : *Fantasia*, par Eugène Delacroix).

Lowe comblera grandement ces défauts avec son descripteur SIFT<sup>15</sup>.

## Généralités

---

La méthode proposée par Lowe comprend deux parties<sup>15</sup> (chacune ayant fait l'objet de recherches plus ou moins indépendantes par la suite) :

- un algorithme de détection de caractéristiques et de calcul de descripteurs ;
- un algorithme de mise en correspondance proprement dit.

De ces deux aspects, le premier est sans doute celui qui a le plus assuré la popularité de la méthode<sup>16</sup>, à tel point que le sigle SIFT fait plus souvent référence aux « descripteurs SIFT » qu'à la méthodologie globale. Il s'agit tout d'abord de détecter sur l'image des zones circulaires « intéressantes », centrées autour d'un *point-clé* et de rayon déterminé appelé *facteur d'échelle*. Celles-ci sont caractérisées par leur unité visuelle et correspondent en général à des éléments distincts sur l'image. Sur chacune d'elles, on détermine une orientation intrinsèque qui sert de base à la construction d'un histogramme des orientations locales des contours, habilement pondéré, seuillé et normalisé pour plus de stabilité. C'est cet histogramme qui sous la forme d'un vecteur à 128 dimensions (ou valeurs) constitue le descripteur SIFT du point-clé, et l'ensemble des descripteurs d'une image établissent ainsi une véritable signature numérique du contenu de celle-ci.

Ces descripteurs présentent l'avantage d'être invariants à l'orientation et à la résolution de l'image, et peu sensibles à son exposition, à sa netteté ainsi qu'au point de vue 3D. Ils possèdent ainsi des propriétés similaires à celles des neurones du Cortex visuel primaire qui permettent la détection d'objet en intégrant les composantes de base comme les formes, la couleur ainsi que le mouvement<sup>17</sup>. Par exemple, ils décriront de façon très semblable les détails d'une image originale et ceux d'une image retouchée par l'application d'une rotation, d'un recadrage ou d'un lissage, par une correction de l'exposition, par occultation partielle ou encore par l'insertion dans une image plus grande.

Une fois le calcul des descripteurs effectué, l'algorithme de mise en correspondance recherche les zones de l'image qui contiennent des éléments visuellement similaires à ceux d'une bibliothèque d'images de référence, c'est-à-dire des descripteurs numériquement proches. Ceci ne peut fonctionner que parce que les descripteurs SIFT sont à la fois robustes (aux principales transformations affines et aux changements d'exposition ou de perspective) et discriminants (deux objets différents ont une grande probabilité d'avoir des descripteurs différents).

## Détection des points-clés et calcul du descripteur SIFT

---

La première étape de l'algorithme est la détection des points d'intérêt, dits *points-clés*. Un point-clé  $(x, y, \sigma)$  est défini d'une part par ses coordonnées sur l'image ( $x$  et  $y$ ) et d'autre part par son facteur d'échelle caractéristique ( $\sigma$ ). En toute rigueur, il s'agit d'une zone d'intérêt circulaire, le rayon de la zone étant proportionnel au facteur d'échelle. Il s'ensuit une étape de reconvergence et de filtrage qui permet d'améliorer la précision sur la localisation des points-clés et d'en éliminer un certain nombre jugés non pertinents. Chaque point-clé restant est ensuite associé à une orientation intrinsèque, c'est-à-dire ne dépendant que du contenu local de l'image autour du point clé, au facteur d'échelle considéré. Elle permet d'assurer l'invariance de la méthode à la rotation et est utilisée comme référence dans le calcul du descripteur, qui constitue la dernière étape de ce processus<sup>18</sup>.

## Détection d'extrema dans l'espace des échelles

La détection s'effectue dans un espace discret que l'on appelle *espace des échelles* (*scale space*) qui comporte trois dimensions : les coordonnées cartésiennes  $x$  et  $y$  et le facteur d'échelle  $\sigma$ . On appelle *gradient* de facteur d'échelle  $\sigma$  (noté  $L$ ) le résultat de la convolution d'une image  $I$  par un filtre gaussien  $G$  de paramètre  $\sigma$ , soit<sup>19</sup> :

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$



Pyramide de gradients : 3 octaves de 6 gradients.

Cette convolution a pour effet de lisser l'image originale  $I$  de telle sorte que les détails trop petits, c'est-à-dire de rayon inférieur à  $\sigma$ <sup>note 1</sup>, sont estompés. Par conséquent, la détection des objets de dimension approximativement égale à  $\sigma$  se fait en étudiant l'image appelée différences de gaussiennes (en anglais *difference of gaussians*, DoG) définie comme suit :

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma),$$

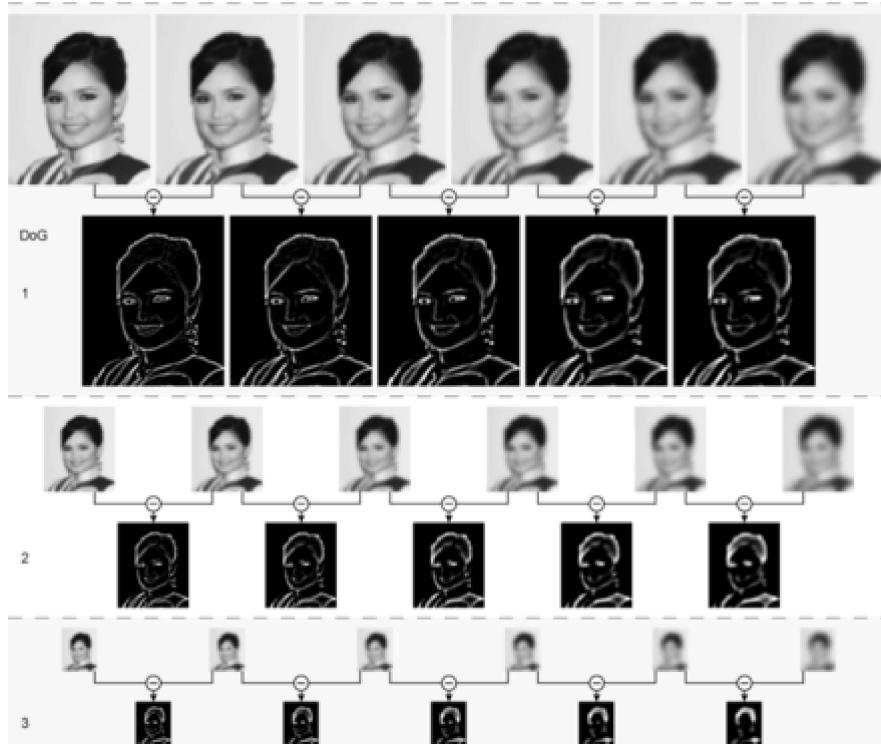
où  $k$  est un paramètre fixe de l'algorithme qui dépend de la finesse de la discréétisation de l'espace des échelles voulue<sup>19</sup>.

Dans cette image ne persistent plus que les objets observables dans des facteurs d'échelle qui varient entre  $\sigma$  et  $k\sigma$ . De ce fait, un point-clé candidat  $(x, y, \sigma)$  est défini comme un point où un extremum du DoG est atteint par rapport à ses voisins immédiats, c'est-à-dire sur l'ensemble contenant 26 autres points défini par :

$$\{D(x + \delta_x, y + \delta_y, s\sigma), \delta_x \in \{-1, 0, 1\}, \delta_y \in \{-1, 0, 1\}, s \in \{k^{-1}, 1, k\}\}$$

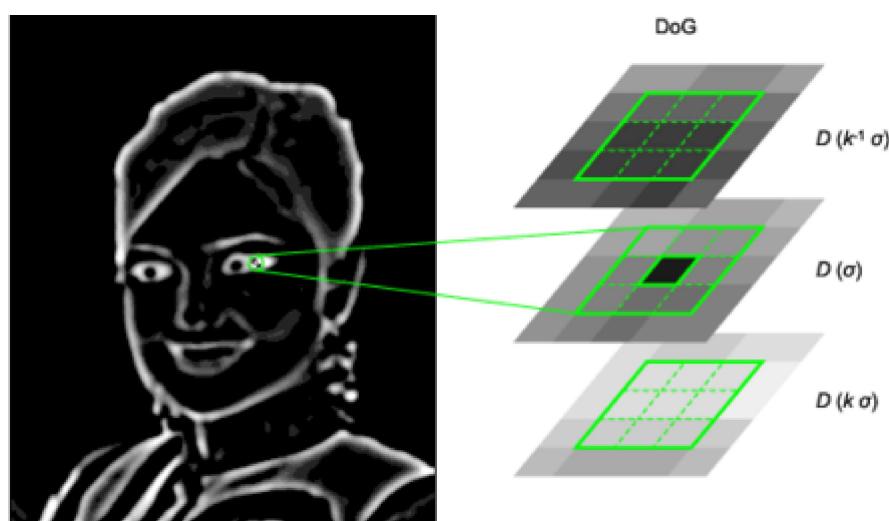
L'utilisation d'une pyramide est préconisée pour optimiser le temps de calcul des images floutées à un grand nombre d'échelles différentes. La base de la pyramide est en général l'image originale et un niveau donné – on parle d'*octave* par analogie avec la musique – est obtenu à partir du précédent en divisant la résolution de l'image par 2, ce qui revient à doubler le facteur d'échelle. Au sein d'une même octave, le nombre de convolées à calculer est constant. Le facteur fixe  $k$  dans les formules ci-

dessus est calculé pour qu'au final, l'espace discréte des facteurs d'échelles considérés corresponde à une progression géométrique  $\{\sigma_0, k\sigma_0, k^2\sigma_0, \dots\}$ , avec à chaque changement d'octave une valeur  $k^p\sigma_0$  qui devient égale à une quantité de la forme  $2^t\sigma_0$ . Ce détail – la progression géométrique des facteurs d'échelle – est important pour que les valeurs des DoG à différentes échelles soient comparables entre elles et il évite, observe Lowe, d'avoir à utiliser un facteur de normalisation dans leur calcul<sup>20</sup>.



Construction de la pyramide de différences de gaussiens (DoG) à partir de la pyramide de gradients.

L'étape de détection des points-clés candidats décrite ci-dessus est une variante de l'une des méthodes de *blob detection* (détection de zones) développée par Lindeberg, qui utilise le laplacien normalisé par le facteur d'échelle<sup>21</sup> au lieu des DoG. Ces derniers peuvent être considérés comme une approximation des laplaciens et présentent l'avantage d'autoriser l'utilisation d'une technique pyramidale<sup>22</sup>.



Exemple de détection d'extremums dans l'espace des échelles.

## Localisation précise de points clés

L'étape de détection d'extremums<sup>20</sup> produit en général un grand nombre de points-clés candidats, dont certains sont instables ; de plus, leur localisation, en particulier aux échelles les plus grandes (autrement dit dans les octaves supérieures de la pyramide où la résolution est plus faible) reste approximative. De ce fait, des traitements supplémentaires sont appliqués, pour un objectif double : d'une part, reconverger la position des points pour améliorer la précision sur  $\mathbf{x}$ ,  $\mathbf{y}$  et  $\sigma$  ; d'autre part, éliminer les points de faible contraste ou situés sur des arêtes de contour à faible courbure et donc susceptibles de « glisser » facilement.



### Amélioration de la précision par interpolation des coordonnées

Visant à augmenter de façon significative la stabilité et la qualité de la mise en correspondance ultérieure<sup>23</sup>, cette étape, qui est une amélioration de l'algorithme original, s'effectue dans l'espace des échelles à trois dimensions, où  $D(\mathbf{x}, \mathbf{y}, \sigma)$ , qui n'est connu que pour des valeurs discrètes de  $\mathbf{x}$ ,  $\mathbf{y}$  et  $\sigma$ , doit être interpolé.

Cette interpolation s'obtient par un développement de Taylor à l'ordre 2 de la fonction différence de gaussiennes  $D(\mathbf{x}, \mathbf{y}, \sigma)$ , en prenant comme origine les coordonnées du point-clé candidat<sup>23</sup>. Ce développement s'écrit comme suit :

$$D(\mathbf{x}) = D + \frac{\partial D}{\partial \mathbf{x}}^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2} \mathbf{x}$$

où  $D$  et ses dérivées sont évaluées au point-clé candidat et où  $\mathbf{x} = (\mathbf{x}, \mathbf{y}, \sigma)^T$  est un delta par rapport à ce point. Les dérivées sont estimées par différences finies à partir des points voisins connus de façon exacte. La position précise de l'extremum  $\hat{\mathbf{x}}$  est déterminée en résolvant l'équation annulant la dérivée de cette fonction par rapport à  $\mathbf{x}$  ; on trouve ainsi<sup>23</sup> :

$$\hat{\mathbf{x}} = -\frac{\partial^2 D}{\partial \mathbf{x}^2}^{-1} \frac{\partial D}{\partial \mathbf{x}}$$

Un delta  $\hat{\mathbf{x}}$  supérieur à 0,5 dans l'une des trois dimensions signifie que le point considéré est plus proche d'un des voisins dans l'espace des échelles discret. Dans ce cas, le point-clé candidat est mis à jour et l'interpolation est réalisée à partir des nouvelles coordonnées. Sinon, le delta est ajouté au point candidat initial qui gagne ainsi en précision<sup>24</sup>.



Après la détection des extrema dans l'espace des échelles (leurs positions sont indiquées sur l'image du haut), l'algorithme élimine les points de faible contraste (les points restants apparaissent sur l'image du milieu), puis les points situés sur les arêtes. Les points restants sont indiqués sur l'image du bas.

Un algorithme de reconvergence similaire a été proposé dans l'implémentation temps-réel basée sur les pyramides hybrides de Lindeberg et Bretzner<sup>25</sup>.

### Élimination des points-clés de faible contraste

La valeur de  $D(\mathbf{x})$  aux coordonnées précises  $\hat{\mathbf{x}}$  du point-clé peut être calculée à partir du développement de Taylor de cette fonction, et constitue donc un extremum local. Un seuillage absolu sur cette valeur permet d'éliminer les points instables, à faible contraste<sup>[23, note 2](#)</sup>.

## Élimination des points situés sur les arêtes

Les points situés sur les arêtes (ou contours) doivent être éliminés car la fonction DoG y prend des valeurs élevées, ce qui peut donner naissance à des extrema locaux instables, très sensibles au bruit : si l'image devait subir un changement numérique même imperceptible, de tels points-clés peuvent se retrouver déplacés ailleurs sur la même arête, ou même simplement disparaître<sup>[26](#)</sup>.

Un point candidat à éliminer, si l'on considère les deux directions principales à sa position, est caractérisé par le fait que sa courbure principale le long du contour sur lequel il est positionné est très élevée par rapport à sa courbure dans la direction orthogonale<sup>[26](#)</sup>. La courbure principale est représentée par les valeurs propres de la matrice hessienne  $\mathbf{H}$  :

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}$$

Les dérivées doivent être évaluées aux coordonnées du point d'intérêt  $(x, y, \sigma)$  dans l'espace des échelles. Les valeurs propres de  $\mathbf{H}$  sont proportionnelles aux courbures principales de  $D$ , dont seul le rapport  $r$  est intéressant. La trace de  $\mathbf{H}$  représente la somme de ces valeurs, le déterminant son produit. Par conséquent, en adoptant un seuil  $r_{th}$  sur le ratio des courbures ( $r_{th} = 10$  dans la méthode originale de Lowe<sup>[2](#)</sup>), un point-clé candidat va être retenu, selon le critère adopté par Lowe, si<sup>[26](#)</sup> :

$$R = \frac{\text{tr}(\mathbf{H})^2}{\det(\mathbf{H})} = \frac{(r+1)^2}{r} < \frac{(r_{th}+1)^2}{r_{th}}$$

La vérification de ce critère est rapide, ne nécessitant qu'une dizaine d'opérations flottantes seulement. Lorsque ce critère n'est pas vérifié, le point est considéré comme localisé le long d'une arête et il est par conséquent rejeté.

Cette étape est inspirée de la technique de détection de points d'intérêt par l'opérateur de Harris ; pour le seuillage, une matrice hessienne est utilisée au lieu de la matrice des moments d'ordre 2 (tenseur)<sup>[26](#)</sup>.

## Assignation d'orientation

L'étape d'assignation d'orientation consiste à attribuer à chaque point-clé une ou plusieurs orientations déterminées localement sur l'image à partir de la direction des gradients dans un voisinage autour du point. Dans la mesure où les descripteurs sont calculés relativement à ces orientations, cette étape est essentielle pour garantir l'invariance de ceux-ci à la rotation : les mêmes descripteurs doivent pouvoir être obtenus à partir d'une même image, quelle qu'en soit l'orientation<sup>[14](#)</sup>.

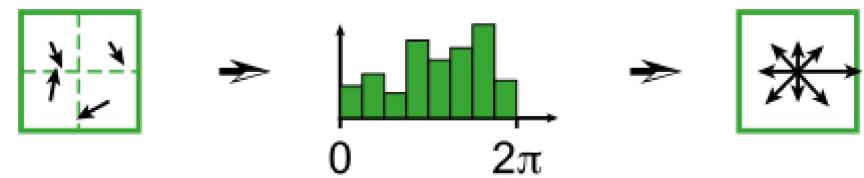
Pour un point-clé donné  $(x_0, y_0, \sigma_0)$ , le calcul s'effectue sur  $L(x, y, \sigma_0)$ , à savoir le gradient de la pyramide dont le paramètre est le plus proche du facteur d'échelle du point. De cette façon, le calcul est également invariant à l'échelle. À chaque position dans un voisinage du point-clé, on estime le gradient par différences finies symétriques, puis son amplitude (c.-à-d. sa norme)  $m(x, y)$ , et son orientation  $\theta(x, y)$ <sup>[14](#)</sup> :

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2}$$

$$\theta(x, y) = \tan^{-1} \left( \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \right)$$

$\forall (x, y)$  dans un voisinage de  $(x_0, y_0)$ .

Un histogramme des orientations sur le voisinage est réalisé avec 36 intervalles, couvrant chacun 10 degrés d'angle. Chaque voisin est doublement pondéré dans le calcul de l'histogramme : d'une part, par son amplitude  $m(x, y)$ ; d'autre part, par une fenêtre circulaire gaussienne de paramètre égal à 1,5 fois le facteur d'échelle  $\sigma_0$  du point-clé. Les pics dans cet histogramme correspondent aux orientations dominantes. Toutes les orientations permettant d'atteindre au moins 80 % de la valeur maximale sont prises en considération, ce qui provoque si nécessaire la création de points-clés supplémentaires ne différant que par leur orientation principale<sup>14</sup>.



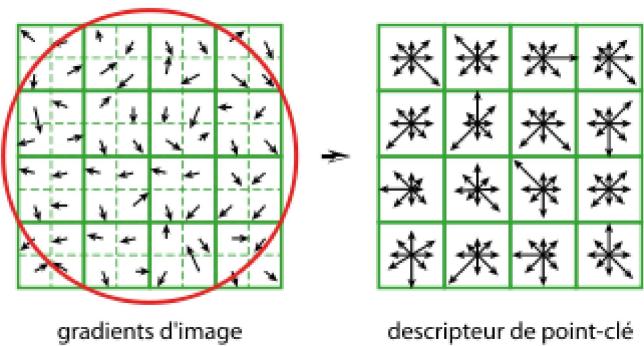
Construction de l'histogramme des orientations.

À l'issue de cette étape, un point-clé est donc défini par quatre paramètres  $(x, y, \sigma, \theta)$ . Il est à noter qu'il est parfaitement possible qu'il y ait sur une même image plusieurs points-clés qui ne diffèrent que par un seul de ces quatre paramètres (le facteur d'échelle ou l'orientation, par exemple).

## Descripteur de point-clé

Une fois les points-clés, associés à des facteurs d'échelles et à des orientations, détectés et leur invariance aux changements d'échelles et aux rotations assurée, arrive l'étape de calcul des vecteurs descripteurs, traduisant numériquement chacun de ces points-clés. À cette occasion, des traitements supplémentaires vont permettre d'assurer un surcroît de pouvoir discriminant en rendant les descripteurs invariants à d'autres transformations telles que la luminosité, le changement de point de vue 3D, etc. Cette étape est réalisée sur l'image lissée avec le paramètre de facteur d'échelle le plus proche de celui du point-clé considéré<sup>27</sup>.

Autour de ce point, on commence par modifier le système de coordonnées local pour garantir l'invariance à la rotation, en utilisant une rotation d'angle égal à l'orientation du point-clé, mais de sens opposé. On considère ensuite, toujours autour du point-clé, une région de  $16 \times 16$  pixels, subdivisée en  $4 \times 4$  zones de  $4 \times 4$  pixels chacune. Sur chaque zone est calculé un histogramme des orientations comportant 8 intervalles. En chaque point de la zone, l'orientation et l'amplitude du gradient sont calculés comme précédemment. L'orientation détermine l'intervalle à incrémenter dans l'histogramme, ce qui se fait avec une double pondération – par l'amplitude et par une fenêtre gaussienne centrée sur le point clé, de paramètre égal à 1,5 fois le facteur d'échelle du point-clé<sup>28</sup>.



Construction d'un descripteur SIFT.

Ensuite, les 16 histogrammes à 8 intervalles chacun sont concaténés et normalisés. Dans le but de diminuer la sensibilité du descripteur aux changements de luminosité, les valeurs sont plafonnées à 0,2 et l'histogramme est de nouveau normalisé, pour finalement fournir le descripteur SIFT du point-clé, de dimension 128<sup>29</sup>.

Cette dimension peut paraître bien élevée, mais la plupart des descripteurs de dimension inférieure proposés dans la littérature présentent de moins bonnes performances dans les tâches de mise en correspondance<sup>29</sup> pour un gain en coût de calculs bien modéré, en particulier quand la technique *Best-Bin-First* (BBF) est utilisée pour trouver le plus proche voisin. Par ailleurs, des descripteurs de plus grande dimension permettraient probablement d'améliorer les résultats, mais les gains escomptés seraient dans les faits assez limités, alors qu'à l'inverse augmenterait sensiblement le risque de sensibilité à la distorsion ou à l'occultation. Il a également été démontré que la précision de recherche de correspondance de points dépasse 50 % dans les cas de changement de point de vue supérieur à 50 degrés, ce qui permet d'affirmer que les descripteurs SIFT sont invariants aux transformations affines modérées. Le pouvoir discriminant des descripteurs SIFT a pour sa part été évalué sur différentes tailles de bases de données de points-clés ; il en ressort que la précision de mise en correspondance est très marginalement impactée par l'augmentation de la taille de la base de données, ce qui constitue une bonne confirmation du pouvoir discriminant des descripteurs SIFT<sup>30</sup>.

## Utilisation pour la recherche d'objets dans des images

---

La problématique de base pour laquelle la méthode SIFT a été conçue est la suivante : peut-on trouver dans une image donnée (dite *image question* ou *image suspecte*), des objets déjà présents dans une collection d'images de référence pré-établie ?

Dans la méthode originale de David Lowe<sup>1,2</sup>, les points-clés et les descripteurs SIFT sont tout d'abord extraits des images de référence et stockés dans une sorte de base de données. Un objet est identifié dans l'image question en effectuant une comparaison de ses descripteurs à ceux des images de référence disponibles en base de données, fondée simplement sur la distance euclidienne. Parmi toutes les correspondances ainsi établies, des sous-ensembles (*clusters*) sont identifiés, au sein desquels la mise en correspondance est cohérente du point de vue des positions des points, des facteurs d'échelle et des orientations. Les clusters contenant au moins trois correspondances ponctuelles sont conservés. Dans chacun d'eux, on modélise la transformation permettant de passer de l'image question à l'image de référence, et on élimine les correspondances aberrantes par simple vérification de ce modèle. Enfin, Lowe applique un modèle probabiliste pour confirmer que la détection d'une correspondance d'objets entre l'image question et l'une des images de référence n'est pas due au hasard, basé sur l'idée que si de nombreux points n'ont pas pu être mis en correspondance c'est que l'on a peut-être affaire à un faux positif<sup>31</sup>.

À chacune de ces étapes, Lowe<sup>2</sup> propose une approche efficace présentée succinctement dans le tableau ci-dessous et dont les principes sont détaillés dans les paragraphes suivants.