Évaluation des performances de descripteurs pour le suivi d'objets

M. MIKRAM, R. MEGRET, Y. BERTHOUMIEU

Laboratoire IMS, département LAPS, UMR 5218 CNRS, Université Bordeaux 1 – ENSEIRB – ENSCPB, Talence, France.

{mikram, megret, berthoumieu}@laps.ims-bordeaux.fr

Résumé – Dans cet article nous présentons une nouvelle approche pour l'évaluation quantitative de la performance des modèles d'apparence formés d'un descripteur d'objet et d'une mesure de similarité dans le contexte du suivi d'objet. L'évaluation est menée en tirant partie de l'existence de vérités-terrains issues de benchmarks pour le suivi d'objet, qui sont utilisées d'une façon originale. L'approche proposée est une extension au contexte spécifique de la vidéo des méthodes d'évaluation de modèles d'apparence utilisées en recherche d'images par le contenu. Les mesures utilisées prennent en effet en compte de la dimension temporelle, en quantifiant la capacité d'un modèle d'apparence à rester discriminant au cours du temps. Cette approche est illustrée par des expérimentations sur des vidéos naturelles.

Abstract – In this paper, a new framework is presented for the quantitative evaluation of the performance of appearance models composed of an object descriptor and a similarity measure in the context of object tracking. The evaluation takes advantage of existing ground-truths from object tracking benchmarks that are used in an original way. The proposed approach is an extension of the methods dedicated to the performance evaluation of appearance models to the specific context of object tracking. Indeed, the presented framework takes into account the temporal dimension, by measuring the capacity of such a model to remain discriminative over time. This approach is illustrated by experiments on natural video data.

1. Introduction

La mise en œuvre de l'évaluation d'un système de suivi d'objets [1] est un problème difficile du fait de la complexité du système lui-même, qui est constitué d'au moins trois composants de base qui sont :

- le modèle d'apparence qui décrit ce à quoi un objet doit ressembler dans une image,
- l'algorithme d'optimisation, qui tente d'estimer la position de l'objet en optimisant la correspondance entre l'apparence actuelle et le modèle d'apparence,
- les contraintes spatio-temporelles, qui donnent un a priori sur la position de l'objet en fonction du suivi passé.

Différentes méthodes pour la mesure des performances de systèmes de suivi ont déjà été proposées [2, 3, 4, 5]. Chacune de ces méthodes évalue les performances grâce à un certain nombre de mesures sur la qualité de la localisation estimée par le système. Ces mesures se fondent sur un corpus vidéo [6] auquel est associée une vérité terrain qui capture l'interprétation vraie de la scène en terme d'objets à suivre. Une telle évaluation prend en compte uniquement la réponse fournie par le système, ce qui correspond à une approche de type "boîte noire". Ce type d'évaluation, même si elle offre une quantification utile des performances, cantonne la mesure à un niveau global et ne permet pas de caractériser les performances intrinsèques des différents éléments composant le système.

Dans cet article, nous proposons de compléter le paradigme standard "boîte noire" en nous focalisant sur le modèle d'apparence. Ce choix s'explique par le fait que le modèle d'apparence représente le lien par lequel le système extrait l'information de position à partir des données brutes. A ce titre, il constitue un élément essentiel du système dont nous désirons évaluer la performance indépendamment des autres éléments. Nous nous limitons dans cette étude à un modèle d'apparence formé d'un descripteur, qui associe un vecteur de description à une zone de l'image, et d'une similarité, qui quantifie le degré de vraisemblance du descripteur courant par rapport à un descripteur de référence. Ce modèle est à la base de nombreux algorithmes utilisés pour le suivi [1] qui sont basés sur un « noyau » spatial définissant la zone de l'objet. Bien que la notion de performance d'un descripteur soit mise en avant, il sera entendu qu'il s'agit de la performance d'un couple descripteur/similarité. En effet, un même descripteur peut produire des performances très différentes selon le type de similarité à laquelle il est associé.

L'évaluation des performances de descripteurs a été abondamment étudiée dans le contexte de la recherche d'images par le contenu [7, 8]. Il s'agit de retrouver une

classe d'images représentant un objet particulier, un type d'objet ou un type de scène dans une base de données, en comparant les descripteurs calculés sur chaque image. La problématique qui nous intéresse est proche de l'approche de recherche d'images par le contenu tout en possédant une spécificité fondamentale : le suivi consiste à localiser l'objet et à le distinguer du fond dans une succession temporelle d'images au lieu d'identifier une classe d'images dans un ensemble non ordonné d'images.

Après avoir exposé les principes généraux de notre approche d'évaluation dans la section 2, nous détaillerons dans la section 3 les mesures qualitatives et quantitatives proposées pour évaluer la performance des modèles d'apparence. Des résultats expérimentaux permettront d'illustrer comment le cadre présenté peut être utilisé pour comparer des modèles d'apparence.

2. Principe général

2.1 Modélisation d'un système de suivi

Pour un objet n, la vérité terrain est représentée par $\mathbf{b}^{*}_{n,t}$ la boîte englobante de celui-ci dans l'image I_t. Pour une boîte candidate $\mathbf{b}_{n,t}$, il est possible de définir une mesure d'erreur notée e entre la boîte candidate et la vérité terrain :

$$e_{n,t} = e(\mathbf{b}_{n,t}, \mathbf{b}^{*}_{n,t}) \qquad (1)$$

La méthodologie standard de type "boîte noire" [2-5] consiste à prendre $\mathbf{b}_{n,t}$ en sortie du système de suivi. La mesure d'erreur e sert alors de base à la définition de métriques de performance. La méthodologie proposée dans cet article utilise un paradigme différent, qui met le modèle d'apparence au premier plan.

Le modèle d'apparence M est représenté par un couple descripteur/similarité. À toute boîte englobante $\mathbf{b}^{i}_{n,t}$ est associé un descripteur $v^{M,i}_{n,t}$ calculé sur cette boîte. En particulier, toute boîte $\mathbf{b}^{*}_{n,tref}$ de la vérité terrain donne lieu à un descripteur de référence $v^{M*}_{n,tref}$ calculé sur l'image *tref*. Il est alors possible de qualifier la vraisemblance d'une boîte englobante au sens du descripteur par une mesure de similarité notée :

$$s^{M,i}_{n,t} = s(v^{M,i}_{n,t}, v^{M*}_{n,tref}) \qquad (2)$$

Dans le cadre de notre approche, l'évaluation quantifie la capacité d'un modèle d'apparence à rester discriminant, c'est-à-dire à distinguer les positions correctes des positions incorrectes, malgré l'écart temporel *t-tref* entre l'image de référence et l'image courante.

2.2 Mise en place d'un corpus d'évaluation

Afin de définir concrètement ce qui constitue une position correcte et une position incorrecte, un corpus de données est construit à partir des informations de vérité-terrain utilisées dans des benchmarks annoté à la main, tels que PETS [2-4], ou générés de façon semi automatique [5]. Les séquences vidéo utilisées dans notre évaluation et leurs vérité-terrain associées sont issues du projet CAVIAR [6].

Pour chaque objet n et chaque instant t, une base de données est constituée, à partir d'éléments $(I_t, \boldsymbol{b}_{n,t}, v_{n,t})$ associant un descripteur à l'image et à la boîte englobante sur lequel il a été calculé. Chaque élément appartient à l'une de deux classes suivantes définies par rapport à l'objet n :

- Une classe des "cibles" associée à l'objet, qui contient des éléments de toutes les images où l'objet apparaît ayant une position acceptable $\mathbf{b}_{n,t} \in B^{in}_{n,t}$.
- Une classe des "distracteurs" associée au fond qui contient des éléments ayant une position incorrecte $\mathbf{b}_{n,t} \in B^{out}_{n,t}$.

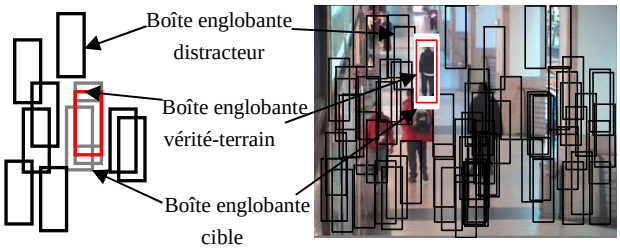

FIG. 1 : conception d'une base de données de boîtes englobantes pour l'objet *n=1* de la séquence *1*. Les boîtes cibles sont translatées d'une petite distance de la boîte vérité-terrain. Les boîtes distracteurs ne recouvrent pas la boîte vérité-terrain.

La décision pour inclure un item dans la classe des cibles ou des distracteurs dépend d'un seuil sur l'erreur de position. Ainsi la classe des cibles correspond aux boîtes englobantes qui ont une erreur autorisée de quelques pixels par rapport à la vérité terrain. D'autre part, la classe des distracteurs forme un échantillonnage de boîtes qui ne chevauche pas l'objet d'intérêt (voir figure 1).

3. Evaluation des performances

3.1 Mesures brutes du pouvoir de discrimination

L'évaluation est fondée sur la définition d'un critère de discrimination : étant donné un objet requête n dont le modèle $v^{M*}_{n,tref}$ est estimé à l'instant *tref*, le modèle est discriminant à l'instant t si les descripteurs cibles $v^{M,i}_{n,t}$ calculés en $\mathbf{b}^{i}_{n,t} \in B^{in}_{n,t}$ sont plus similaires à $v^{M*}_{n,tref}$ que les descripteurs distracteurs $v^{M,j}_{n,t}$ calculés en $\mathbf{b}^{j}_{n,t} \in B^{out}_{n,t}$.

Après classement de tous les descripteurs à l'instant t par ordre décroissant de similarité, on considère les rangs respectifs de la cible la plus similaire, de la cible la moins similaire, et du distracteur le plus similaire. L'utilisation d'une distance $d_{n,t,tref}$ au lieu d'une similarité est possible, il faut alors classer les descripteurs par ordre croissant de distance. On notera $d^{in}_{n,t,tref}$ la distance entre le modèle et la cible la plus similaire.

Pour un couple d'instants $(tref, t)$ donné, le pouvoir de discrimination $c^M_{n,tref,t}$ d'un modèle d'apparence M pour un objet n est quantifié en trois catégories :

- Totalement discriminant ($c^M_{n,tref,t}$ =2) lorsque toutes les cibles sont mieux classées que les distracteurs.
- Discriminant ou partiellement discriminant ($c^M_{n,tref,t}$ =1) si l'une des cibles est moins similaire qu'un distracteur.
- Non discriminant ($c^M_{n,tref,t}$ =0) lorsque le descripteur le plus similaire est un distracteur.

Il est important de noter que les valeurs de similarité ou de distance associées à différents modèles d'apparences ne sont pas manipulées et comparées directement, mais seulement à travers la capacité à discriminer entre cibles et distracteurs. Ainsi des types de similarités et de descripteurs différents peuvent être comparés.

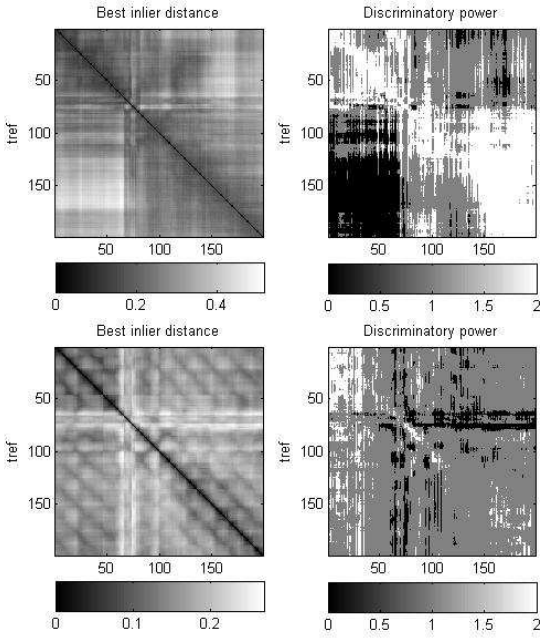

FIG. 2 : la meilleure distance $d^{in}_{n,tref,t}$ (à gauche), et le pouvoir de discrimination $c^M_{n,tref,t}$ (à droite) utilisant le modèle *MHG* (en haut) ou *MTG* (en bas), pour l'objet de la figure 1.

Les mesures précédentes sont définies pour chaque objet *n* et chaque image *t*, pour un modèle calculé sur une image de référence *tref*. Elles peuvent être représentées sous la forme de matrices où chaque ligne correspond à une image de référence *tref*, et chaque colonne à une image *t*.

Afin d'illustrer ces mesures, nous utiliserons les modèles d'apparence suivants, nommés selon la nomenclature suivante : H pour Histogramme ou T pour Template, G pour niveaux de gris ou C pour couleur. Les modèles *MHG* et *MHC* correspondent à un histogramme respectivement en niveaux de gris (*256*) et en couleur RVB (*6×6×6*), calculé sur le contenu de la boîte englobante, et associé à une distance de Matusita. Les modèles *MTG* et *MTC* correspondent à une imagette respectivement en niveaux de gris et en couleur RVB obtenue en redimensionnant le contenu de la boîte englobante à une taille de *20×20* pixels, et à laquelle associée à une distance Euclidienne.

Les représentations matricielles de la distance de la meilleure cible $d^{in}_{n,tref,t}$ et du pouvoir de discrimination $c^M_{n,tref,t}$ sont illustrées pour les modèles *MHG* et *MTG* sur la figure 2. On peut remarquer que la diagonale correspond à la recherche d'un objet dans la même image que celle sur laquelle le modèle a été calculé. En s'éloignant de la diagonale, la distance temporelle |*t-tref*| entre l'image courante et l'image de référence s'accroît (figure 2 à gauche), ce qui augmente la chance que le modèle soit moins discriminant, à cause d'un changement d'apparence de l'objet au cours du temps. Ainsi un modèle est toujours au moins partiellement discriminants à proximité immédiate de la diagonale, mais peut devenir non discriminant lorsqu'un changement temporel de son apparence rend un distracteur plus similaire à la référence que les cibles. C'est notamment le cas (figure 2 à droite) pour le modèle *MHG* pour *tref<70* et *t>70*, alors que le modèle *MTG* reste discriminant dans cette situation.

3.2 Mesures quantitatives

Plusieurs mesures quantitatives peuvent être extraites des matrices précédentes. En premier lieu, une mesure du pouvoir de discrimination global peut être associée à un descripteur pour chaque objet n en calculant la proportion de couples (*tref,t*) pour lesquels le modèle est discriminant:

$$D^M_n = \frac{\#\left\{(tref,t)/\, c^M_{tref,t} \geq 1\right\}}{\#\{t\}\#\{tref\}} \qquad (3)$$

où $\#\{tref\}=\#\{t\}$ représente le nombre d'images dans lesquelles l'objet *n* apparaît.

Afin de caractériser numériquement la capacité d'un modèle d'apparence à rester discriminant au cours du temps, le pouvoir de discrimination peut être exprimé en fonction de l'écart temporel Δt, en analysant les diagonales secondaires des matrices.

$$D^M_n(\Delta t) = \frac{\#\left\{(tref,t)/\, c^M_{tref,t} \geq 1 \text{ et } t-tref = \Delta t\right\}}{\#\left\{(tref,t)/\, t-tref = \Delta t\right\}} \qquad (4)$$

Comme nous l'avons noté pour les modèles *MHG* et *MTG*, deux modèles peuvent ne pas avoir les mêmes modes de défaillance, les valeurs de $(tref,t)$ qui correspondent à une situation de non discrimination étant différentes dans les deux cas. En reprenant à la figure 3 (gauche) la mesure $D^M_n(\Delta t)$ sur l'exemple des figures 1 et 2, on retrouve l'effet de la présence d'un distracteur pour *t<70* sur le modèle *MHG*, mais pas sur *MTG*, qui semble par contre plus sensible à des variations périodiques de l'apparence (dues au mouvement des jambes). L'étude d'un autre objet de la séquence (figure 3 à droite) permet de révéler une autre situation, où *MHG* est toujours meilleur que *MTG*.

Afin de résumer les différents cas, et de déterminer pour un couple de modèles d'apparence M_1 et M_2 s'ils échouent dans les mêmes situations, ou s'ils montrent des comportements complémentaires, il est intéressant de quantifier la proportion de couples (*t,tref*) pour lesquelles

l'un des modèles est discriminant alors que l'autre ne l'est pas. Le pouvoir de discrimination comparatif pour que le modèle M1 soit supérieur au modèle M2 est défini par :

$$D_n^{M_1>M_2} = \frac{\#\left\{(tref,t)/\, c_{tref,t}^{M_1} \geq 1 \text{ et } c_{tref,t}^{M_2} = 0\right\}}{\#\{t\}\#\{tref\}} \quad (5)$$

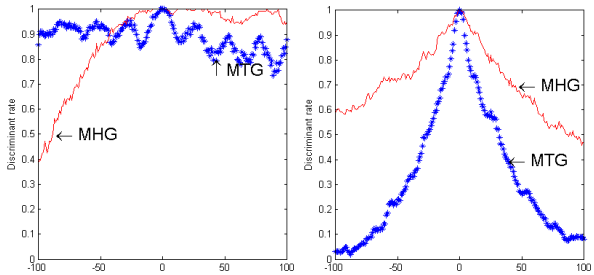

FIG. 3 : pouvoir de discrimination en fonction de l'écart temporel $D_n^M(\Delta t)$ pour les modèles *MHG* et *MTG* pour l'objet de la fig. 1 (à gauche) et un autre objet (à droite).

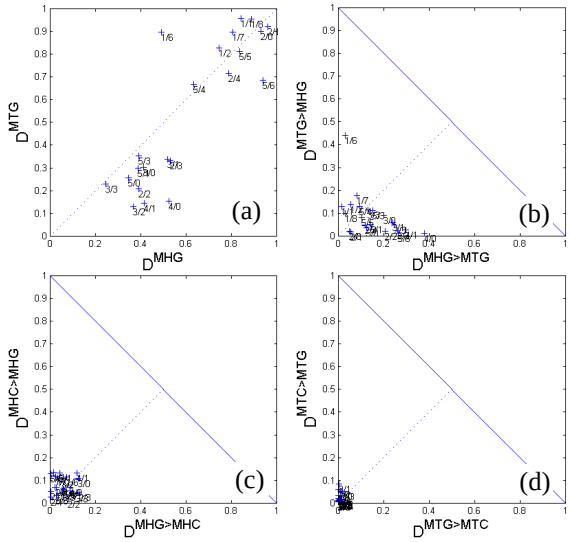

FIG. 4 : comparaison de paires *M1/M2* de modèles d'apparence : *MHG/MTG* (a,b), *MHG/MHC* (c), *MTG/MTC* (d), selon le pouvoir de discrimination global (D^{M1}, D^{M2}) (a) et le pouvoir de discrimination comparatif ($D^{M1>M2}$, $D^{M2>M1}$) (b,c,d). Chaque point est étiqueté par n°-séquence/n°-objet.

Quand $(D_n^{M_1>M_2}, D_n^{M_2>M_1}) \approx (0,0)$, les deux modèles d'apparence ont le même comportement, et échouent dans les mêmes situations. Quand $D_n^{M_1>M_2} \approx 0$ et $D_n^{M_2>M_1}$ est élevé, le modèle M_2 est meilleur que le modèle M_1. Quand les deux valeurs sont élevées, les deux modèles sont complémentaires, et échouent dans différentes situations.

La comparaison de deux modèles peut ainsi utiliser les mesures D_n^M et $D_n^{M1>M2}$, en affichant ces mesures pour un nombre important de situations (une situation étant définie comme l'étude du suivi d'un objet au sein d'une séquence). Dans les figures 4-a et 4-b, il est ainsi montré notamment que les modèles *MHG* et *MTG* ne réussissent pas systématiquement dans les même situations, l'un ou l'autre des modèles étant mieux adapté en fonction des situations. Cette représentation permet par exemple d'étudier l'apport de la couleur dans les cas étudiés : les modèles par imagette *MTG* et *MTC* ont quasiment les même situations de réussite (figure 4-c : points concentrés autour de l'origine), alors que les approches par histogramme *MHG* et *MHC* semblent réussir dans des cas plus complémentaires (figure 4-d : concentration moins forte des points autour de l'origine).

4. Conclusion

Ce papier a présenté une approche originale pour l'évaluation des performances des modèles d'apparence composés d'un descripteur et d'une mesure de similarité pour le suivi. Elle étend les approches d'évaluation de descripteurs pour la recherche d'images par le contenu au contexte du suivi d'objets en prenant en compte spécifiquement l'aspect temporel des vidéos, au travers de la conception d'une structure spécifique du corpus d'évaluation et de la proposition de nouvelles mesures de performance. Cette approche est complémentaire des benchmarks de suivi d'objets au sens où elle se focalise sur l'étude du modèle d'apparence, au lieu d'étudier un système « boîte noire » complet, tout en réutilisant les bases de vérité-terrain existantes. L'accent a ici été mis sur la présentation du cadre, qui a été illustré par quelques exemples montrant l'intérêt des mesures proposées. Le champ expérimental sera à l'avenir étendu en incluant plus de vidéos et de modèles d'apparence.

Références

[1] A. Yilmaz, O. Javed et M. Shah, "Object Tracking: A Survey,'' *ACM Journal of Computing Surveys*, Vol. 38, No. 4, 2006.

[2] S.M Schneiders, T. Jager, H.S. Loos et W. Niem, "Performance Evaluation of a Real Time Video Surveillance Systems," *VS-PETS 2005*, Beijing, pp. 15-16.

[3] L.M. Brown, A.W. Senior, Y.L Tian, J. Connell et A. Hampapur, C-F. Shu, H. Merkl et M. Lu "Performance Evaluation of Surveillance Systems under Varying Conditions," *PETS 2005*, Breckenridge, Colorado, pp.1-8.

[4] F. Bashir et F. Porikli, "Performance Evaluation of Object Detection and Tracking Systems,'' *PETS 2006*, New-York, pp. 7-14.

[5] J. Black, T. Elis et P. Rosin, "A novel method for video tracking performance evaluation,'' *VS-PETS 2003*, Nice, pp. 125–132.

[6] CAVIAR: EU funded project, IST 2001 37540, URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/ (2004).

[7] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet et T. Pun, "Performance Evaluation in Content-based Image Retrieval: Overview and Proposals," *Pattern Recognition Letters*, Vol. 22, No. 5, pp. 593–601, 2001.

[8] T. Deselaers, D. Keysers et H. Ney, "Features for image retrieval: A quantitative comparison," In *DAGM'04: 26th Pattern Recognition Symposium*, Tübingen, Vol. 3175 LNCS, pp. 228–236, 2004.