Un jeu, des images, des clics et du texte : collecte implicite de données visuelles et sémantiques

Axel Carlier, Vincent Charvillat

Un jeu, des images, des clics et du texte : collecte implicite de données visuelles et sémantiques

Axel Carlier [1] et Vincent Charvillat[1]

[1]Université de Toulouse, IRIT-ENSEEIHT, 2 rue Camichel, 31000 Toulouse
axel.carlier@enseeiht.fr vincent.charvillat@enseeiht.fr

Résumé
Nous décrivons un corpus de données visuelles et sémantiques collectées à partir d'un jeu. Nous avons conçu ce jeu pour deux joueurs qui coopèrent à distance sur le Web. Les données collectées sont directement utilisables pour résoudre des problèmes de vision (par exemple des problèmes de détection, de segmentation, d'étiquetage sémantique). La collecte est toutefois implicite au sens où le jeu n'a pas pour but explicite de détecter ou segmenter des objets présents dans une image. Le corpus inclut plus de 3,250 jeux basés sur 104 images et comporte des annotations textuelles et des données spatiales (clics, relations spatiales). Dans cet article, nous expliquons pourquoi et comment utiliser ces données pour différentes applications visant la compréhension des images. Nous montrons surtout qu'elles sont suffisamment riches pour superviser, dans un sens à définir, une analyse sémantique globale du contenu visuel. Le corpus est rendu accessible aux chercheurs.

Mots clé : Jeu GWAP, Corpus de données, Sémantique, Détection, Segmentation

1. Introduction

L'analyse sémantique d'images au sens le plus ambitieux est parfois nommée *holistic image understanding* ou *image parsing* dans la littérature anglo-saxonne. Il s'agit d'un des problèmes centraux en vision par ordinateur et sans doute aussi d'un des plus ardus. Il s'agit d'associer une étiquette sémantique à chaque pixel pour assurer l'interprétation complète d'une image. Cela est si compliqué que les chercheurs préfèrent souvent décomposer le problème en plusieurs sous-problèmes comme la classification d'images, la détection d'objets ou la segmentation. Ces sous-problèmes, bien que très fouillés ces dernières années, restent ouverts. A titre d'exemple et malgré plus de vingt ans de travaux dans ce domaine, le problème de la segmentation d'un objet d'intérêt n'est pas encore résolu par des approches générales et automatiques.

Les recherches dans ce domaine se scindent en deux catégories principales :

– *Les approches par apprentissage artificiel* : la résolution de problèmes de segmentation s'appuie dans ce cas sur des bases d'apprentissage de grande taille qui regroupent des vérités terrains, c'est-à-dire des masques de segmentation considérés comme exacts pour différents objets. A partir de ces données d'apprentissage, des prédicteurs sont appris pour segmenter, par inférence, de nouvelles images. Bien qu'efficaces pour certains objets, ces approches se heurtent à des difficultés pour différentes catégories d'objets. Par exemple, le vainqueur de la dernière compétition *PASCAL VOC Challenge, 2012* pour la segmentation obtient des scores modestes dans le cas des bicyclettes, des chaises, des tables ou autres canapés.
– *Les approches interactives de la segmentation* : on admet dans ce cas que seule une segmentation semi-automatique est atteignable. Ce qui revient à dire que le *gap sémantique* est délicat à combler sans intervention humaine. Placer un (ou des) utilisateur(s) dans le processus permet d'initialiser, de superviser, de corriger des algorithmes de segmentation au travers d'interfaces adaptées.

Dans les deux cas précédents, notons bien qu'une supervision humaine est présente, ou bien dans la constitution de la vérité terrain ou bien au travers d'une segmentation interactive. La question du recrutement de ces humains, en particulier lorsqu'ils servent d'experts, est ouverte. Les approches dites de *crowdsourcing* donnent une réponse au travers de plateformes comme *amazon mechanical turk* ou *microworkers*. Selon Luis Von Ahn qui est parmi les premiers à avoir parlé d'*human computation*, une alternative est possible au travers de jeux appelés GWAP *Games With A Purpose*. Alors qu'un paiement récompense les humains travaillant sur *amazon mechanical turk*, la motivation pour participer à un jeu est ou devrait être naturelle. Par principe, un jeu GWAP doit simultanément être amusant et pouvoir contribuer à la ré-

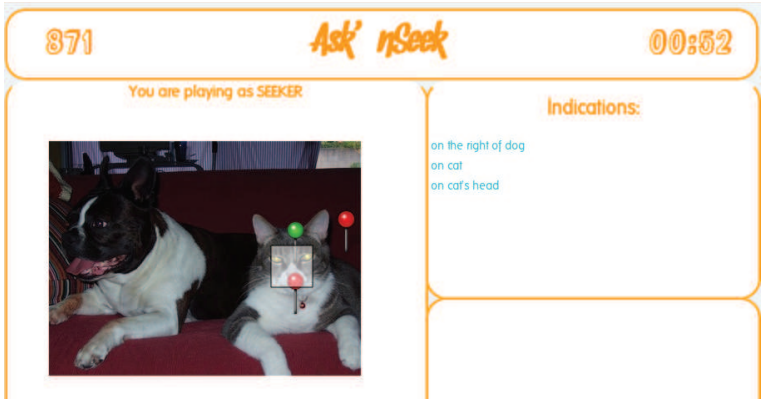

Figure 1: *Ecran du jeu de l'enquêteur qui demande des indices relatifs au chien, au chat et finalement à la tête du chat.*

solution d'un problème. Un jeu GWAP parmi les plus populaires est le jeu ESP [VAD04] qui réunit deux joueurs à distance à travers une application web. Les joueurs gagnent des points lorsqu'ils annotent les mêmes images avec des étiquettes similaires. Le scénario permet, statistiquement, d'annoter globalement les images utilisées par ESP.

Dans cet article, nous suivons la même voie en utilisant un jeu que nous avons proposé il y a deux ans [CMC12, SCGiN*13]. Cet article dresse un bilan des données que nous avons collectées après 3,250 jeux basés sur 104 images. Dans les paragraphes suivants, nous montrons le très fort potentiel du corpus de données visuelles et textuelles obtenu. Ce corpus sera publiquement accessible.

2. Corpus de données existants

De nombreux corpus de données sont disponibles pour évaluer les algorithmes de segmentation en vision par ordinateur. Parmi eux, deux jeux de données sont particulièrement populaires : le corpus de Berkeley *Berkeley Segmentation Dataset* [MFTM01] et le corpus du réseau d'excellence PASCAL *PASCAL VOC Dataset* [EVGW*10].

Celui issu de Berkeley (BSDS300 en résumé) est composé de 300 images auxquelles sont associées des segmentations manuelles opérées par des experts. Il y a plusieurs (typiquement entre 5 et 10) segmentations par image et chaque segmentation, pour une image donnée, est différente et donc complémentaire des autres. Ensemble ces segmentations fournissent une bon approximation des différentes interprétations possibles d'une même image. Notons que ce corpus a été étendu à 500 images. Le corpus PASCAL n'a pas uniquement été créé pour la segmentation mais aussi pour la détection d'objets, la classification d'actions, les détections de parties du corps etc. En 2012, à la compétition PASCAL, on pouvait dénombrer dans le corpus 11530 images associées à 27450 annotations liées à des régions d'intérêt (ROI) et 6929 segmentations. Le corpus *MSRA dataset* [LYS*11] est aussi reconnu dans le domaine de la détection d'objets. Ce corpus intègre 20840 images associées à une vérité terrain constituée de boîtes englobantes indiquant la localisation des objets les plus saillants dans chaque image.

Ces trois corpus sont intéressants car ils intègrent une grande quantité de données et une vérité terrain pour différents problèmes. Le corpus que nous présentons dans ce travail est de nature différente. Il intègre des images et des traces collectées au travers d'un jeu. Ces traces ne correspondent pas directement à des solutions aux problèmes de détection ou de segmentation auxquels on s'intéresse. Les traces collectées sont plus proches en nature des gribouillages (ou *scribbles*) utilisés pour apporter une supervision partielle des algorithmes de segmentation interactive.

La segmentation interactive a été beaucoup étudiée dans les dernières années et plusieurs corpus ont été proposés pour évaluer ces algorithmes. Le corpus *Grabcut* [RKB04] est bâti sur 50 images (dont 20 proviennent du corpus de Berkeley) et sur trois types de masques pour la vérité terrain. Gulshan et al [GRC*10] ont augmenté ce corpus avec des images de PASCAL pour atteindre un total de 151 images avec vérité terrain. McGuinness et al [MO10] proposent quant à eux un corpus de 96 images extraites de Berkeley auxquelles ils ajoutent des masques déterminés manuellement en guise de vérité terrain. Ce corpus est utilisé comme *benchmark* pour comparer différents algorithmes de segmentation interactive. Bien que plus proches de notre travail, ces corpus n'intègrent pas de traces utilisateurs comme notre ensemble de données. Une telle collecte a, par contre, été proposée dans d'autres travaux autour des GWAP.

Von Ahn et al ont par exemple publié un jeu de données contenant 100,000 images et des annotations provenant du jeu GWAP nommé ESP [VAD04]. Ce corpus est pratique pour tester des requêtes et des algorithmes de recherche d'images dans un contexte d'indexation textuelle. Il n'est pas naturellement adapté à l'évaluation d'algorithmes de détection ou de segmentation. Russell et al [RTMF08] ont aussi rendu public le corpus *LabelMe* provenant du jeu du même nom. Ce corpus intègre des tracés de polygones détourant les objets d'intérêt. Ce détourage est trop grossier pour évaluer des algorithmes de segmentation. A la différence du jeu que nous défendons, les joueurs ont clairement conscience de détourer ou segmenter les objets d'intérêt. La conception de notre jeu est telle que les joueurs n'ont pas conscience que leurs actions (mises en commun) contribuent indirectement à l'identification des objets visibles dans une image.

3. Le jeu

Le jeu que nous utilisons est un jeu à 2 joueurs déployé sur Internet. Il a été initialement introduit par Carlier et al. dans [CMC12]. Le jeu est non symétrique : le premier joueur appelé maître commence par cacher une cible (ou région cachée) sur une image, puis doit aider le second joueur appelé enquêteur (ou *seeker* dans la version anglaise) à la découvrir. Sur son écran, l'enquêteur voit la même image que le maître mais ne connait évidemment pas la position de la cible. Son but est de la trouver en cliquant en son sein.

Pour ce faire, l'enquêteur peut demander des indications au maître. Plus spécifiquement, l'enquêteur peut taper des nom d'objets qui apparaissent à l'écran c'est-à-dire dans l'image, puis demander au maître de localiser la cible par rapport à ces objets. Le maître peut répondre de 7 manières différentes. La cible peut se situer "à gauche", "à droite", "au dessus" ou "en dessous" de l'objet cité. Elle peut également être située sur l'objet, ou partiellement sur l'objet. Enfin, il arrive qu'il ne soit pas possible de catégoriser la position de la cible par rapport à l'objet avec l'une de ces 6 possibilités : dans ce cas la cible ne peut pas être reliée à l'objet. Grâce à ces indications, l'enquêteur va peu à peu réduire le champ des positions possibles pour la cible.

Ce mécanisme est illustré sur la figure 1. Dans la partie jouée sur cet exemple, l'enquêteur a commencé par demander au maître un indice lié au chien. Le maître a répondu que la cible était située à droite du chien. L'enquêteur a obtenu le droit de cliquer sur l'image (logiquement à droite du chien), mais a manqué la cible. Il a donc eu le droit de demander un second indice, et après avoir tapé le mot "chat", il a reçu l'indication que la cible était située "sur" le chat. Après avoir à nouveau manqué la cible, il a finalement obtenu l'information que la cible était située "sur la tête du chat" et a réussi à cliquer dans la région cachée, déclenchant du même temps l'apparition de la cible sur son image et la fin de la partie.

Comme nous allons l'expliquer avec plus de détails, les informations collectées sont directement utilisables pour résoudre des problèmes de vision (par exemple des problèmes de détection ou de segmentation). Pour autant, la collecte est implicite au sens où le jeu n'a pas pour but de détecter ou segmenter des objets présents dans une image.

4. Nature des données

Les données collectées via notre jeu sont produites par les interactions entre deux joueurs ayant les rôles de maître et d'enquêteur. Il est important de comprendre que les données produites par ces deux rôles sont distinctes et sont complémentaires en vue de la compréhension des images. L'enquêteur fournit deux types d'informations : des informations textuelles (lorsqu'il demande des indices) et des informations spatiales quand il clique sur l'image en espérant atteindre la région cachée. Le maître fournit deux autres types d'informations. Le premier type correspond à la position de la région qu'il sélectionne pour être découverte durant le jeu (dont on rappelle qu'il répond à un scénario de coopération). Le second type relève des relations spatiales qu'ils fournit entre les objets/indices demandés par l'enquêteur et la région cachée. Puisque l'enquêteur prend en compte ces informations avant de cliquer, ces relations spatiales sont aussi révélatrices de relations entre les objets et les clics produits. Nous détaillons maintenant les données collectées.

4.1. Informations textuelles

Les données textuelles (ou *tags*) saisis par l'enquêteur sont de formes variées. Comme la durée du jeu est limitée, les joueurs se contentent généralement de textes courts. La plupart des données sont réduites à un mot faisant référence à un objet présent dans l'image ou à une partie d'un tel objet. La partie droite de figure 2 illustre cela avec mes mots "tree", "plant" et "tiger" (mot) et "face", "tail", ou "tongue" qui sont des méronymes de "tiger". L'analyse (statistique) de ces données permet un étiquetage sémantique de l'image.

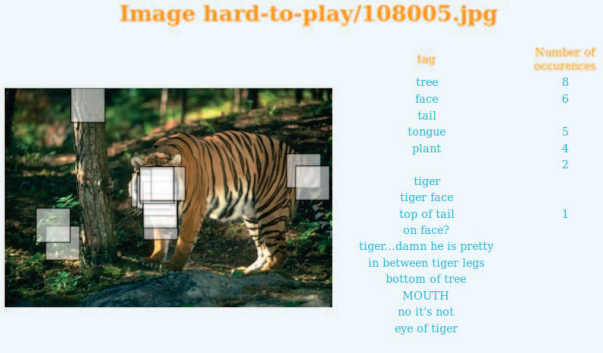

Figure 2: *Illustration compacte des jeux joués sur une image. A gauche, les régions cachées et à droite les textes et leur nombre d'occurrences.*

Il arrive aussi que les textes soient des groupes de mots qui peuvent être très informatifs en connectant un objet et une de ses parties ("eye of tiger", "tiger face") ou plus compliqués à interpréter ("in between tiger legs"). L'analyse de ces traces peut conduire à la détermination d'une hiérarchie d'objets au sens sémantique/ontologique.

Dans le cas des instances multiples d'objets, les traces collectées sont riches et complexes. Les textes permettent aux joueurs de distinguer différents objets de même catégorie.

Dans la figure 3, le texte "soldier in the foreground on the left" suggère qu'il y a des soldats à l'avant et à l'arrière plan. Nous avons observé, dans le corpus, une intéressante corrélation entre la longueur des textes et la présence d'instances multiples d'objets : pour gérer les ambiguïtés les textes se complexifient.

De manière évidente enfin, les données textuelles présentes dans notre corpus sont adaptées à la catégorisation d'images et à l'étude des co-occurrences sémantiques d'objets dans certaines catégories d'images.

4.2. Régions cachées

Les régions cachées sont les cibles positionnées par le maître du jeu. La figure 4 présente, dans la ligne centrale, toutes les régions cachées sur deux images du corpus par les joueurs maîtres. Il est intéressant de noter que les positions

soldier in the foreground on
the left
the left soldiers back
soldier on the right
blue berret
man on left
man to the right

Figure 3: *Une image, en haut, où des instances multiples conduisent à des textes visant à les distinguer, en bas.*

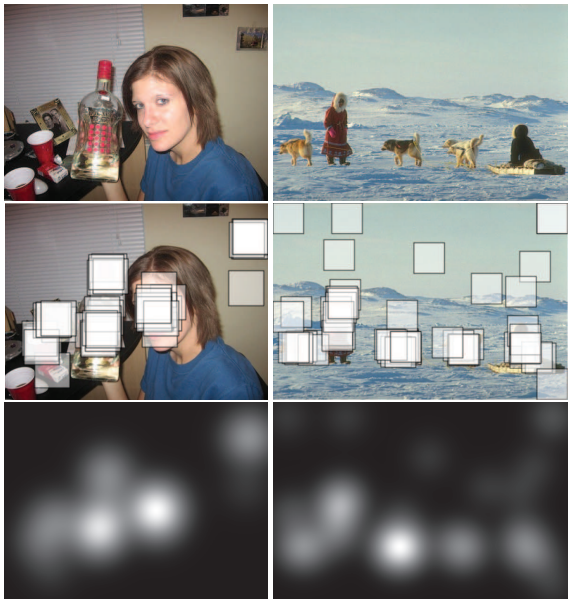

Figure 4: *Cartes de saillance issues des régions cachées par les joueurs pour deux images (chacune sur une colonne).*

de ces carrés ne sont pas aléatoires mais souvent placées sur les objets les plus importants, les plus discriminants et dans un sens, les plus saillants. Cela peut être expliqué par la nature coopérative du jeu : pour vite gagner des points, le maître du jeu doit placer la région cachée à un emplacement qui sera facilement identifié par l'enquêteur et cela grâce à quelques indices seulement. Au travers du corpus, les positions des régions cachées sont précurseurs de nouvelles cartes de saillance que nous illustrons en mélangeant des gaussiennes centrées sur les positions en question en bas de la figure 4. Ce résultat prometteur mérite une étude ultérieure et une comparaison avec d'autres mécanismes attentionnels que nous estimons possible grâce à la mise à disposition de notre corpus de données.

4.3. Clics et relations spatiales

Tous les clics des enquêteurs sont associés à un *tag* et à des relations spatiales (*above*, *below*,*on the left of*,*on the right of*,*on*,*partially on*).

Au dessus, En dessous, A gauche, A droite. La figure 5 illustre comment nous pouvons utiliser ces clics pour détecter des objets. Les points rouges sont les clics collectés au dessus du chat alors que les points bleus sont les clics collectés à droite du chat. Les lignes correspondent aux boîtes englobantes que nous pouvons utiliser pour détecter le chat si nous faisons entière confiance aux traces présentes dans le corpus (lignes en pointillés) ou si nous choisissons de résister aux erreurs potentielles (lignes pleines). Dans ce dernier cas, l'utilisation d'estimateurs de position robustes aux donnés aberrantes est nécessaire ; le plus simple d'entre eux est la médiane.

Sur. La figure 6 montre tous les clics de type "on" (sur un objet) obtenus pour une image. Il est clair que la distribution de ces points peut permettre de segmenter (ou de contraindre la segmentation de) certains des objets présents (hut, man, trees, stick). L'image en dessous montre les points de type "on" obtenus sur des parties (head, skirt, butt, feet) d'un objet qui les inclut (man).

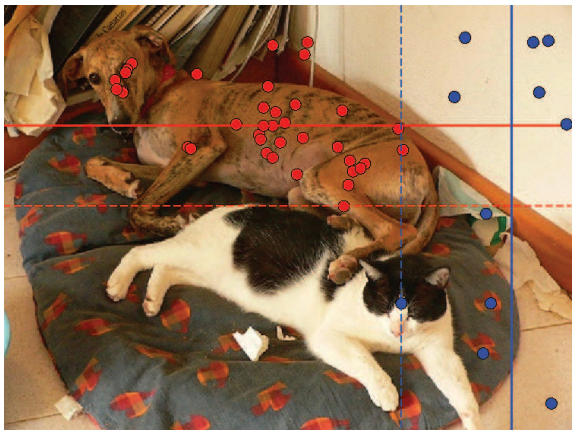

Figure 5: *Clics au dessus (en rouge) et à droits (en bleu) du chat. Le* tag *'cat' étant avec 'dog' le plus fréquemment collecté pour cette image.*

Partiellement sur.. La figure 7 montre tous les clics situés "partiellement sur" l'objet "urn". Dans la version actuellement en ligne du jeu cette relation, assez mal comprise de certains joueurs, est remplacée par "sur le bord de". La figure montre cependant que les "partially on points" sont situés majoritairement au voisinage des contours de l'urne. Ce constat est vérifié sur beaucoup d'images du corpus. Ces données sont une source intéressante d'information pour la segmentation en conjonction avec les points cliqués "sur" un objet et ceux situés en dehors de l'objet (clics "above", "below", "left" et "right").

Figure 6: *Les clics sur des objets ou* 'on clicks'.

Les traces collectées via notre jeu sont à la fois variées et complémentaires les unes par rapport aux autres. Nous pouvons les utiliser pour la segmentation d'un objet d'intérêt présent dans une image mais aussi, en généralisant à plusieurs objets, pour la compréhension globale de la scène représentée par une image. La figure 6 illustre ce potentiel vis-à-vis d'un objectif d' **image parsing**. L'originalité du corpus est aussi de reposer sur des relations spatiales nombreuses qui structurent les informations textuelles, les *tags* et les clics. Des relations spatiales entre objets, entre instances multiples de mêmes objets ou entre parties d'objets peuvent être extraites du corpus.

Figure 7: *En vert, tous les points qui ont été cliqués comme étant "partiellement sur" l'urne.*

5. Collecte des données

Dans ce paragraphe nous expliquons comment les traces ont été collectées pour former le corpus présenté. D'abord nous avons choisi des images déjà sélectionnées par McGuinness et al [MO10] pour leur propre corpus. Nous sommes partis de cet ensemble car il avait déjà été utilisé pour comparer différentes approches de segmentation interactive. En plus des 96 images de McGuinness et al. nous avons aussi sélectionné 8 images de PASCAL pour être jouées intensément. Les traces sur ces images peuvent être utilisées pour déterminer combien de jeux sont suffisants pour obtenir des résultats satisfaisants étant donné un problème à résoudre (détection, segmentation etc.).

Le corpus intègre des images de difficultés différentes vis-à-vis du jeu. Des images trop simples (intégrant un seul objet centré de grande taille par rapport à celle de la région à cacher) se révèlent difficiles à utiliser durant le jeu par manque d'indices visuels par rapports auxquels l'enquêteur peut guider sa recherche. Inversement, des images plus complexes intégrant plusieurs objets se révèlent plus simples à manier au travers du jeu. Lorsqu'un nouveau joueur se connecte à la plateforme de jeu, des images faciles à jouer lui sont d'abord proposées.

Dans tous les cas, les joueurs débutants doivent suivre un tutoriel vidéo et deux tutoriels interactifs les initiant aux rôles de maître ou d'enquêteur respectivement. Aucune explication n'est donnée quant à l'intérêt du jeu vis-à-vis de la recherche scientifique et de la compréhension semi-automatique d'images. Par contre, des recommandations sont fournies pour rappeler le caractère coopératif du jeu, pour indiquer que des *tags* simples permettent de bien gérer la limite de temps etc. Comme le jeu n'est pas célèbre, nous avons lancé plusieurs campagnes auprès d'étudiants et

de réseaux sociaux. Les campagnes ont impliqué plus de 50 joueurs qui se sont réunis à des heures précises pour faciliter l'appariement aléatoire de joueurs en ligne.

6. Le corpus en pratique

Caractéristiques du corpus : Les données des milliers de jeux joués ont été stockés dans une base de données MySQL. Nous rendons publique la partie scientifiquement pertinente de cette base de données. La base de données est constituée des tables suivantes :

- *User*. Cette table contient les *login* et les *password* cryptés des joueurs. Ces informations ne seront pas rendues publiques. Du point de vue scientifique, les seules informations intéressantes de cette table sont l'âge et le genre des joueurs.
- *Game*. Pour chaque jeu, nous stockons l'image qui a été jouée, les identificateurs (*ID*) des joueurs formant une paire maître-enquêteur, le score final, le temps restant à la fin du jeu. On peut ainsi simplement déduire de ce dernier attribut le temps effectif du jeu puisque le chronomètre interrompt le jeu après 120 secondes.
- *Region*. Pour chaque jeu, nous stockons la région cachée par le maître. La région est identifiée par ses coordonnées centrales avec, de plus, sa hauteur et sa largeur. Les figures 2 et 11 montrent toutes les régions cachées sur des images particulières.
- *Indication*. Cette table est la plus intéressante puisqu'elle contient les textes (*tags*) soumis par les enquêteurs, les relations données en retour par les maîtres et les coordonnées des clics qui s'en suivent. Cet ensemble forme une indication. Chaque ligne de la table du même nom consiste en une indication pour un jeu donné.

Interface : Nous fournissons une interface web pour visualiser les traces. Cette interface permet, pour chaque image, d'afficher rapidement les positions des régions cachées durant l'ensemble des jeux concernant cette image. En même temps, les informations textuelles utilisées pour décrire le contenu de l'image sont aussi restituées. Une capture de cette interface est visible dans la figure 2. L'interface peut aussi être utilisée pour rejouer les jeux : l'image jouée et la région cachée sont affichées avec la séquence d'indications ponctuées par des clics de l'enquêteur. La figure 8 illustre cette fonctionnalité.

Numbers : Un total de 3250 jeux constitue le résultat des deux campagnes de jeux. Parmi ces jeux, 3010 se sont terminés. Seuls 114 jeux ne se sont pas terminés avec avec la découverte de la région (96% des jeux sont gagnants avec un score, lié au temps, plus ou moins important). La durée moyenne d'un jeu se terminant par un succès est de 29 secondes. Cette durée passe à 33 secondes si on intègre les autres jeux.

Plus de 5000 clics sont collectés pour un total de 9063 indications. Il faut comprendre qu'un clic est associé à plus d'une indication au fil du jeu (les indices sont cumulatifs). Par exemple dans la figure 1, le clic victorieux est simultanément "on the right of the dog", "on the cat" et "on the cat's head". Ce clic apporte donc plus d'une indication.

Figure 8: *Les traces d'un jeu visualisées au travers de l'interface web : à gauche sont visibles l'image, la région cachée et les clics. A droite, les* tags *émis par l'enquêteur durant un jeu.*

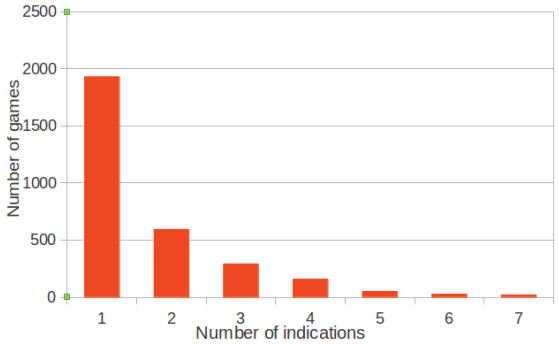

Figure 9: *Distribution du nombre d'indications par jeu.*

La figure 9 montre la distribution du nombre d'indications par jeu. Un grand nombre de jeux joués (pratiquement deux tiers) ont seulement une indication. Une stratégie coopérative efficace est évidemment de cacher la région au centre de l'objet principal ou dominant dans une image. Par exemple la figure 11 montre une situation où il est efficace de placer la région sur la tête d'un animal ou d'un humain présents dans l'image. Quand il y a plusieurs animaux (comme le chien et le chat) il faut statistiquement un peu plus d'une indication par jeu pour gagner.

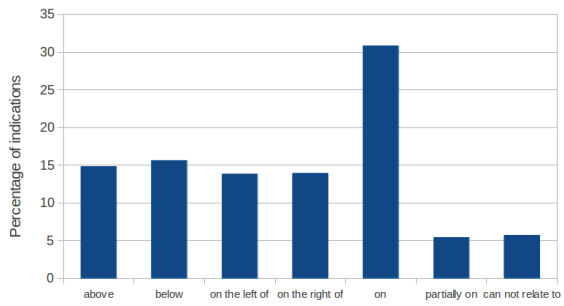

Figure 10: *Statistiques à propos des occurrences des relations spatiales.*

La figure 10 montre la distribution des relations spatiales indiquées par le maître à l'enquêteur. Il est intéressant de noter que les relations *on* sont de loin les plus utilisées. C'est une conséquence directe des faits expliqués ci-dessus : le maître a tendance à cacher la région sur une région saillante qui facilite le rôle de l'enquêteur et maximise le score. Les clics sur les objets les plus saillants sont donc majoritaires, ce qui est intéressant du point de vue de la compréhension d'images. Le second aspect intéressant est que les relations *above*, *below*, *on the left of*, et *on the right of* sont équitablement réparties. Ce qui signifie probablement que la position des objets les plus saillants vers lesquels le maître se penche sont plutôt uniformément répartis dans les images.

Figure 11: *Position des régions cachées dans l'image du chien et du chat.*

Exemples extraits des données. Pour bien montrer le potentiel des données collectées nous avons systématiquement fait joué les joueurs sur une image PASCAL (celle du chien et du chat). Cette image fait partie de celles qui sont assez adaptées au jeu, assez faciles "à gagner". Nous avons enregistré 99 jeux sur cette image particulière. La figure 11 montre les positions des 99 régions cachées. Il est notable que ces régions couvrent les deux deux objets saillants de l'image. Les densités sont plus fortes sur les têtes du chat et du chien. Nous observons logiquement que les deux *tags* les plus utilisés sont évidemment "cat" et "dog" avec respectivement 67 and 66 occurrences. Dans ce décompte nous n'ajoutons pas les mots "cat" et "dog" lorsqu'ils apparaissent dans des textes structurés comme "cat's head" ou "dog's leg".

La figure 12 montre tous les clics issus de tous les enquêteurs relativement au chien. Les clics en jaune sont "on the dog". Les clics en rouge sont "above", "below", "on the left of" ou "on the right of" du chien. On observe des erreurs. Par exemple, des clics théoriquement attendus sur le chien sont en dehors. Il y aussi aussi des points qui devraient être à l'arrière plan qui sont sur le chien. Un des sujets ouverts que nous soumettons à la communauté de recherche est l'élimination des ces erreurs soit via une approche robuste statistiquement soit avec des raisonnements plus sophistiqués intégrant, éventuellement, les propriétés de l'image. C'est une de nos pistes actuelles de recherche.

Figure 12: *Les clics des enquêteurs relativement au chien.*

7. Utilisations possibles du corpus

Le corpus peut être utilisé pour de nombreuses applications et pas uniquement dans le domaine de l'imagerie ou du multimédia. Voilà une liste non exhaustive des travaux qu'il serait pertinent d'envisager à partir des traces du jeu.

Analyses textuelles. Le traitement des informations textuelles collectées par notre jeu est un premier enjeu important. De très nombreuses inférences à propos des catégories d'images ou d'objets, à propos des instances multiples d'objets, à propos des parties d'objets seraient un plus évident dans une démarche de compréhension sémantique globale des images : S'agit-il d'une scène intérieure ou extérieure ? Y a-t-il plusieurs chats dans cette image ? La patte dont on parle serait-elle celle du chat ?

Détection d'objets. La contribution majeure des traces de notre jeu est à l'évidence l'association entre données textuelles et informations spatiales dans l'image. La manière d'opérer ces associations est originale car, selon notre connaissance de l'état de l'art, elle intègre des relations spatiales pour la première fois dans un jeu GWAP. Chaque clic contribue à contraindre la recherche d'un objet dans l'image. Il est naturellement possible d'augmenter la probabilité de détection d'un objet grâce à chaque triplet (*tag*, *spatial relation*, *click*) [CMC12] et aux effets cumulatifs des indications.

Une utilisation "asymptotique" du jeu (c'est-à-dire un cas où beaucoup de jeux sont joués sur chaque image) est possible en vue de l'établissement d'une vérité terrain pour la détection ou la segmentation d'objets.

Cependant et de manière plus réaliste, c'est dans un scénario semi-supervisé que le potentiel du corpus réside plutôt, selon nous. Il nous semble que l'intérêt des traces collectées est de semi-superviser des algorithmes de vision par ordinateur : éliminer des fausses détections d'un algorithme d'OpenCV, renforcer ou combiner des conclusions de détecteurs différents ? Tout cela grâce à une contribution peu coûteuse en nombre de jeux par image.

Elimination d'erreurs. Nous avons écrit plus haut que certaines erreurs (humaines) contaminaient le corpus et nous venons de souligner que les algorithmes de vision commettent aussi des erreurs. Ces imprécisions ou aberrations

qu'elles soient humaines ou algorithmiques s'avèrent de nature si différentes qu'il nous semble possible d'utiliser la vision pour corriger l'humain et réciproquement ! C'est un sujet intéressant vers lequel nous nous dirigeons. Une forme d'apprentissage actif semble aussi envisageable : un algorithme pourrait-il demander (via un jeu ou quelques jeux) une aide pour conforter sa conclusion ? des jeux pourraient-ils faire appel aux sorties d'algorithmes de vision pour détecter des traces aberrantes ?

Segmentation ou **Image Parsing**. La lecture (*parsing*) ou annotation sémantique complète d'une image est probablement l'application la plus intéressante de nos données. Ce corpus est un pas vers l'établissement d'une méthodologie de résolution semi-supervisée de ce problème, dès lors qu'on admet un humain dans la boucle d'interprétation. Les approches semi-supervisées sont ou bien vues comme des méthodologies supervisées avec peu d'exemples d'apprentissage ou bien vues comme des techniques non-supervisées avec des contraintes. Ces deux approches d'utilisation du corpus sont possibles.

8. Conclusion

En conclusion, nous livrons à la communauté un corpus de données visuelles et sémantiques qui possèdent un potentiel intéressant. Le corpus est accessible à l'URL suivante (http ://TBC). Ces données sont issues d'un jeu GWAP original que nous avons proposé il y a deux ans. Le jeu est en ligne à l'URL suivante (http ://TBC). Ce délai de deux ans a été nécessaire pour collecter un nombre significatif de traces qui sont aujourd'hui exploitables et qui pourraient alimenter des recherches au-delà de notre laboratoire. Nos propres travaux sur ces données sont partagés avec des collègues étrangers [CMC12, SCGiN*13]. Les auteurs remercient en particulier Ogé Marques (FAU, USA) et Xavi Giro-i-Nieto (UPC Barcelona) qui seront naturellement associés, dans le futur, à une version ou extension anglaise de cet article.

Références

[CMC12] CARLIER A., MARQUES O., CHARVILLAT V. : Ask'nseek : A new game for object detection and labeling. In *ECCV'12 Workshops*. 2012, pp. 249–258.

[EVGW*10] EVERINGHAM M., VAN GOOL L., WILLIAMS C. K., WINN J., ZISSERMAN A. : The pascal visual object classes (voc) challenge. *IJCV. Vol. 88*, Num. 2 (2010), 303–338.

[GRC*10] GULSHAN V., ROTHER C., CRIMINISI A., BLAKE A., ZISSERMAN A. : Geodesic star convexity for interactive image segmentation. In *CVPR'10* (2010), IEEE, pp. 3129–3136.

[LYS*11] LIU T., YUAN Z., SUN J., WANG J., ZHENG N., TANG X., SHUM H.-Y. : Learning to detect a salient object. *PAMI. Vol. 33*, Num. 2 (2011), 353–367.

[MFTM01] MARTIN D., FOWLKES C., TAL D., MALIK J. : A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV'01* (2001), vol. 2, IEEE, pp. 416–423.

[MO10] MCGUINNESS K., O'CONNOR N. E. : A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition. Vol. 43*, Num. 2 (2010), 434–444.

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A. : Grabcut : Interactive foreground extraction using iterated graph cuts. In *TOG* (2004), vol. 23, ACM, pp. 309–314.

[RTMF08] RUSSELL B. C., TORRALBA A., MURPHY K. P., FREEMAN W. T. : Labelme : a database and web-based tool for image annotation. *IJCV. Vol. 77*, Num. 1-3 (2008), 157–173.

[SCGiN*13] SALVADOR A., CARLIER A., GIRO-I NIETO X., MARQUES O., CHARVILLAT V. : Crowdsourced object segmentation with a game. CrowdMM '13, ACM, pp. 15–20.

[VAD04] VON AHN L., DABBISH L. : Labeling images with a computer game. In *CHI'04* (2004), ACM, pp. 319–326.