Transformée de Hough sans a priori pour la segmentation

Adrien CHAN-HON-TONG[1], Catherine ACHARD[2], Laurent LUCAT[1]

[1]CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Centre d'études de Saclay, Route Nationale, 91400 Gif-sur-Yvette, France

[2]Institut des Systèmes Intelligents et Robotique
CC 173 - 4 Place Jussieu, 75005 Paris, France

adrien.chan-hon-tong@cea.fr, catherine.achard@upmc.fr
laurent.lucat@cea.fr

Résumé – Le paradigme de la transformée de Hough permet de réaliser une segmentation multi-classes dans des données multi-dimensionnelles i.e. des séries temporelles, des images, des vidéos... Cependant, la version standard de la transformée de Hough dite *Implicit Shape Model* (*ISM*) est peu discriminante. Des extensions comme *Max-Margin Hough Transform* (*MMHT*) ou *Implicit Shape Kernel* (*ISK*) consistant à ajouter des paramètres discriminatifs aux paramètres génératifs de *ISM* offrent des améliorations de performances. Dans cet article, nous optimisons directement toutes les variables de la transformée de Hough de façon discriminative. Les performances de cette nouvelle méthode de Hough sont significativement meilleures que celles des précédentes sur le jeu de données *honeybee*.

Abstract – Hough-Transform framework allows to compute multi-class segmentation in multi-dimensional data (e.g. time series - images - videos). However, standard Hough methods like *Implicit Shape Model* (*ISM*) provide poor accuracy. Several extensions like *Max-Margin Hough Transform* (*MMHT*) or *Implicit Shape Kernel* (*ISK*) which consists to add discriminative parameters to the generative ones introduced by *ISM* have reported experimental improvements. In this paper, we directly optimize all variables of the Hough transform in discriminative way. Our new Hough Transform significantly outperforms previous ones on *honeybee* dataset.

1 Introduction

1.1 La transformée de Hough

La transformée de Hough probabiliste, introduite dans [9], connaît une attention croissante suite aux importants gains de performance qu'elle a récemment apporté dans de nombreuses applications dont les plus emblématiques sont l'extraction de squelette depuis une carte de profondeur [5] et la détection d'actions [13]. Dans cet article, on se concentre sur la segmentation temporelle d'actions dans des vidéos : l'objectif est de décider de l'action présente en chaque image/instant d'une vidéo contenant plusieurs actions successives. Dans ce contexte, ce formalisme s'articule autour de trois étapes :

1 : des primitives sont extraites de la vidéo

2 : chaque primitive vote pour chaque activité en chaque instant (dans un large voisinage autour de son instant d'extraction) avec un poids spécifique appris durant une étape d'apprentissage

3 : tous les votes sont agglomérés pour former le score de Hough à partir duquel sont prises les décisions de segmentations

Ce formalisme est une alternative aux méthodes de type *Linear Dynamic Systems* (*LDS*) [11] ainsi qu'aux méthodes basées sur la programmation dynamique réalisant un pavage temporel [6] (*OP* pour *Optimal Padding*).

1.2 Abstraction du processus de vote

Formellement, cette méthode (pour les étapes 2-3) repose sur la définition d'une fonction $w\left(\right)$ (traditionnellement positive) qui lie primitives, activités, écarts temporels et poids du vote. Ainsi, chaque primitive p extraite au temps t vote avec un poids $w\left(a, \delta_t, p\right)$ pour l'hypothèse que l'activité au temps $t + \delta_t$ est a (ce vote ne dépend pas de l'instant t mais uniquement de p, a, δ_t). Typiquement, une valeur $w\left(a, \delta_t, p\right)$ grande doit correspondre à un lien fort entre l'extraction d'une primitive p à l'instant t et la présence d'une action a en $t + \delta_t$ et inversement $w\left(a, \delta_t, p\right) = 0$ signifiant que ce lien est nul comme par exemple pour δ_t supérieur en valeur absolue à la durée maximale d'une action (quelque soit p, a).

Ainsi, étant donné l'ensemble des primitives localisées de la vidéo $\mathcal{V} = \{p, t\}$ (p les primitives, t leurs instants d'extraction), le score de Hough $\mathcal{H}$ pour l'activité a à l'instant $\overline{t}$ est :

$$\mathcal{H}\left(\overline{t}, a\right) = \sum_{(p,t)\in\mathcal{V}} w\left(a, \overline{t} - t, p\right) \quad (1)$$

Ce score $\mathcal{H}\left(\overline{t}, a\right)$ correspond simplement à l'agglomération des votes de l'ensemble des primitives de la vidéo pour l'action a à l'instant $\overline{t}$. Il dépend de l'action a car les primitives votent

différemment selon les actions et de $\bar{t}$ car les primitives votent différemment dans le temps (relativement à leur position d'extraction). L'activité prédite au temps $\bar{t}$ est alors :

$$\widehat{a}\left(\bar{t}\right) = \arg\max_{a} \left(\mathcal{H}\left(\bar{t}, a\right)\right) \quad (2)$$

Cette méthode prédit une action à chaque instant de la vidéo.

Il convient donc de déterminer durant l'apprentissage des valeurs de $w\left(a, \delta_t, p\right)$ qui permettront de déterminer le plus possible d'activités correctes lors de l'application des équations (1) et (2) à de nouvelles vidéos. Les travaux qui s'intéressent à ce sujet sont présentés en section 2. Ces travaux sont tous basés sur les votes génératifs de *ISM*. À l'opposé, nous proposons dans la section 3 d'optimiser directement et simultanément toutes les valeurs $w\left(a, \delta_t, p\right)$. Une série d'expériences menées sur le jeu de données *honeybee*, présentées en section 4, révèlent que les performances de notre méthode y sont significativement meilleures que celles des autres méthodes de Hough, ce qui valide la pertinence de notre approche. De plus, notre méthode est adaptée à des jeux de données de taille importante via une approximation présentée en section 5 permettant de ramener le calcul de $w\left(\right)$ à un problème de type *support à vaste marge* (*SVM*). Enfin la section 6 présente la conclusion et les perspectives.

2 Etat de l'art

Les différentes méthodes de la littérature pour apprendre $w\left(\right)$ sont *ISM* [9] et ses extensions [10, 12, 14].

ISM : Dans [9], chaque valeur $w\left(a, \delta_t, p\right)$ est associée à $\mathcal{P}\left(a, \delta_t | p\right)$ la probabilité que l'action en $t + \delta_t$ soit a sachant que la primitive p a été extraite en t (cette probabilité ne dépendant pas de t mais uniquement de a, δ_t, p). Cette probabilité est mesurée sur l'ensemble d'apprentissage.

MMHT [10] : Les votes de la méthode *ISM* sont influencés par un poids supplémentaire associé à chaque primitive p de manière à privilégier les plus discriminantes.

ISK [14] : Les votes sont associés à la même probabilité que dans *ISM* mais mesurée indépendamment sur chaque exemple d'apprentissage. Le vote final est alors une somme pondérée des différentes probabilités.

ISM + SVM : Dans [12], il est établi que l'application d'un *SVM* sur une fenêtre de taille fixe sur la carte de Hough $\mathcal{H}$ est équivalente au fait de rajouter un poids supplémentaire associé à chaque écart temporel δ_t.

Ici, nous ne présentons pas les méthodes *forêt de Hough* [13] car ces méthodes se comportent comme *ISM* durant les étapes 2-3 de la méthode de Hough. L'apport de ces méthodes vis à vis de *ISM* tient dans l'extraction des primitives (étape 1). Mais chacune des méthodes de Hough présentées dans cet article (résumées dans le tableau 1) pourrait s'appliquer aux primitives extraites par la *forêt de Hough*.

Au vu de l'état de l'art, il apparaît que l'ajout de paramètres discriminatifs améliore les performances de *ISM*. Aussi nous proposons d'optimiser directement l'ensemble des valeurs de $w\left(\right)$ en se démarquant complètement des poids *ISM*. Notre approche n'utilise pas une forme prédéfinie pour $w\left(\right)$ dans laquelle certains paramètres sont à optimiser mais bien une optimisation complète indicée à la fois par les primitives (p) comme dans *MMHT*, les actions (a) et les écarts temporels (δ_t) comme dans *ISM+SVM*.

Notre approche est appelée la transformée de Hough Sans A Priori (*HSAP*).

3 *HSAP*

L'objectif de l'apprentissage est de déterminer une fonction $w\left(\right)$ telle que pour tous les instants $\bar{t}$ dans les exemples $\mathcal{V}$ de la base d'apprentissage, l'annotation décidée $\widehat{a}$ soit l'annotation réelle a^* (connue à l'apprentissage) : $\forall \bar{t}, \mathcal{V}, \widehat{a}\left(\bar{t}\right) = a^*\left(\bar{t}\right)$.

Vu la définition de l'annotation décidée $\widehat{a}$ (éq. (2)), cet objectif est équivalent à ce que pour toutes les étiquettes $a \neq a^*\left(\bar{t}\right)$, on ait $\mathcal{H}\left(\bar{t}, a\right) < \mathcal{H}\left(\bar{t}, a^*\left(\bar{t}\right)\right)$. Comme il est possible de diviser $w\left(\right)$ par le plus petit des écarts des inégalités précédentes (en n'imposant pas à $w\left(\right)$ d'être bornée), ce problème est équivalent à $\mathcal{H}\left(\bar{t}, a\right) + 1 \leq \mathcal{H}\left(\bar{t}, a^*\left(\bar{t}\right)\right)$ lui même équivalent à : $\forall \mathcal{V}, \bar{t}, a \neq a^*\left(\bar{t}\right)$

$$\sum_{(p,t)\in\mathcal{V}} w\left(a, \bar{t} - t, p\right) + 1 \leq \sum_{(p,t)\in\mathcal{V}} w\left(a^*\left(\bar{t}\right), \bar{t} - t, p\right)$$

en remplaçant $\mathcal{H}$ par sa définition (éq. (1)).

Cependant, si les votes de la primitive p pour deux actions différentes correspondent naturellement à deux quantités indépendantes, les votes de la primitive p pour une action a et pour deux décalages temporels proches doivent être proches puisque les actions varient doucement. De plus, une primitive extraite à l'instant t ne peut pas donner plus d'information sur l'instant $t \pm 1$ que sur l'instant t. Pour tenir compte de cette propriété, $w\left(\right)$ est contrainte à être décroissante vis à vis de la valeur absolue de δ_t.

Sur des jeux de données réels, l'ensemble des contraintes associées à $w\left(\right)$ peut ne pas avoir de solution, aussi comme dans [4] une approche de type marges souples est adoptée : des termes ξ sont introduits pour autoriser la non satisfaction de certaines des contraintes. On cherche alors à minimiser une fonction de perte L de ces termes pour tenter de minimiser le nombre de contraintes non satisfaites. Cela conduit ainsi au problème d'optimisation suivant :

$$\min_{w\geq 0, \xi \geq 0} \left(\sum_{\bar{t}} L\left(\xi\left(\bar{t}\right)\right) \right)$$
$$\text{sous contraintes : } \forall \mathcal{V}, \bar{t}, a \neq a^*\left(\bar{t}\right),$$
$$\sum_{(p,t)\in\mathcal{V}} \left(w\left(a^*\left(\bar{t}\right), \bar{t} - t, p\right) - w\left(a, \bar{t} - t, p\right)\right) + \xi\left(\bar{t}\right) \geq 1$$
$$\text{et } \forall p, a, \delta_t,\ w\left(a, \delta_t + \text{sign}\left(\delta_t\right), p\right) \leq w\left(a, \delta_t, p\right)$$

où sign est la fonction signe.

Enfin, afin de limiter le sur-apprentissage, un terme de régularité est ajouté comme dans [2]. Un coefficient C pondère alors le terme d'attache aux données et le terme de régularité,

Table 1: les méthodes d'apprentissage des votes de la transformée de Hough de l'état de l'art.

nom de la méthode	forme de w	variables à optimiser
ISM [9]	$w_{ISM}(a,\delta_t,p) = \mathcal{P}(a,\delta_t\|p)$	-
MMHT [10]	$w_{MMHT}(a,\delta_t,p) = \lambda_p \times \mathcal{P}(a,\delta_t\|p)$	λ_p
ISK [14]	$w_{ISK}(a,\delta_t,p) = \sum_i (\lambda_i \times \mathcal{P}_i(a,\delta_t\|p))$	λ_i
ISM+SVM [12]	$w_{ISM+SVM}(a,\delta_t,p) = \lambda_{\delta_t}\mathcal{P}(a,\delta_t\|p)$	λ_{δ_t}
HSAP	$w_{HSAP}(a,\delta_t,p) = \lambda_{a,\delta_t,p}$	$\lambda_{a,\delta_t,p}$

$\mathcal{P}(a,\delta_t|p)$ est la probabilité que l'action en $t+\delta_t$ soit a sachant que la primitive p a été extraite en t (cette probabilité ne dépendant pas de t mais uniquement de a,δ_t,p). Cette probabilité est mesurée sur l'ensemble d'apprentissage. $\mathcal{P}_i(a,\delta_t|p)$ est cette même probabilité mesurée uniquement sur l'exemple d'apprentissage i.

ce qui aboutit au problème :

$$\min_{w\geq 0,\xi\geq 0}\left(\sum_{a,p,\delta_t}\widehat{L}\left(w\left(a,\delta_t,p\right)\right)+C\sum_{\bar{t}}L\left(\xi\left(\bar{t}\right)\right)\right)$$
$$\text{sous contraintes : } \forall\mathcal{V},\bar{t},a\neq a^*\left(\bar{t}\right),$$
$$\sum_{(p,t)\in\mathcal{V}}\left(w\left(a^*\left(\bar{t}\right),\bar{t}-t,p\right)-w\left(a,\bar{t}-t,p\right)\right)+\xi\left(\bar{t}\right)\geq 1$$
$$\text{et } \forall p,a,\delta_t,\ w\left(a,\delta_t+\text{sign}\left(\delta_t\right),p\right)\leq w\left(a,\delta_t,p\right)$$

Afin que l'apprentissage soit résoluble en pratique, nous proposons d'utiliser la norme 1 pour L et $\widehat{L}$. L'apprentissage est ainsi codé par un programme linéaire qui est un type de problème bien étudié dans la littérature [8].

4 Résultats

Nous avons évalué les différentes méthodes de Hough (tab. 1) sur le jeu de données *honeybee* [11] (sauf *ISK* étant spécifique au problème de détection). Ce jeu de données fournit les résultats de suivi d'abeilles ayant 3 types d'actions corrélés avec leur trajectoire. Il est bien adapté au problème de segmentation d'actions multi-classes car chaque image est associée à une action et qu'il y a plus de 2 types d'actions. Afin de fournir des résultats comparables avec [11], chaque algorithme décide une action à chaque image et la proportion des actions correctement prédites mesure la performance à travers une *leave-one out cross validation*.

Notre objectif étant d'évaluer l'influence de l'optimisation des votes et non l'influence des primitives, la partie extraction de primitives (étape 1) est identique pour les méthodes de Hough. Le signal entrant est constitué de la séquence des positions 2D et des orientations de l'abeille (x_t,y_t,α_t). On note $R(\beta)$ la rotation d'angle $-\beta$ et $p_t=(x_t,y_t)$, le vecteur $(R(\alpha_t)(p_{t-\tau}-p_t),...,R(\alpha_t)(p_{t+\tau}-p_t))$ constitue alors la primitive extraite à l'instant t pour la taille τ. Ces primitives sont quantifiées indépendamment pour chaque valeur de τ par l'algorithme des K-moyennes.

L'ensemble des paramètres de cette série d'expériences est ainsi l'ensemble des tailles τ considérées, les K pour les quantifications et le coefficient d'attache aux données C.

Les performances maximales des précédentes méthodes de Hough obtenues en faisant varier empiriquement l'ensemble des paramètres restent inférieures aux performances moyennes de *HSAP*. $\tau\in\{1,3,6\}$, $K=10$ et $C=1$ donne des résultats représentatifs, les pourcentages de prédictions correctes correspondant sont présentés dans le tableau 2.

Table 2: Les scores sur le jeu de données *honeybee* [11]

algorithme	scores
ISM [9]	71,9
MMHT [10]	78,8
ISM+SVM [12]	77,5
HSAP	**86,5**
PS-SLDS [11]	87,7
OP [6]	**89,3**

PS-SLDS et *OP* obtiennent des performances supérieures à *HSAP*. Néanmoins, *PS-SLDS* est une méthode à latence infinie : elle a besoin de l'ensemble des valeurs de la séquence pour prendre une décision concernant l'action à la première image de la vidéo alors que la latence d'une méthode de Hough est bornée par la durée maximale d'une action. De même, *OP* est une méthode basée sur les scores d'un *SVM* appliqué sur l'ensemble des intervalles temporels de la vidéo de durée compatible avec une action. Ainsi, la complexité de cette méthode est quadratique vis à vis de la durée maximale des actions alors qu'elle est linéaire pour une méthode de Hough.

Ainsi, les performances de la méthode de Hough proposée sont, sur *honeybee*, significativement meilleures que celles de *ISM*, *MMHT* et *ISM+SVM*, et équivalentes à celles de *PS-SLDS* et *OP* qui sont deux algorithmes inutilisables pour une segmentation temps-réel ou sur un jeu de données de taille importante. Cela soutient la pertinence de notre nouvelle approche.

Nos apprentissages ont été réalisés par CPLEX[1].

5 Discussion sur l'entraînement

L'apprentissage proposé peut être difficilement réalisable sur des jeux de données de grandes tailles. Aussi, nous proposons des approximations pour dépasser cette limitation.

[1] www-01.ibm.com/software/commerce/optimization/cplex-optimizer/

Introduisons χ la notation des fonctions caractéristiques i.e. si I est un intervalle temporel alors $\chi_I(t)$ vaut 1 si $t \in I$, 0 sinon. Soit $\mathcal{J}$ un ensemble d'intervalles temporels, alors : $w(a, \delta_t, p) \approx \sum_{I \in \mathcal{J}} w_{a,I,p} \chi_I(\delta_t)$. Un intérêt de l'approximation est que w introduit une variable pour chaque a, δ_t, p contre une pour chaque a, I, p pour son approximation or il y a autant de δ_t que deux fois la durée maximale d'une action contre un nombre constant d'intervalles I. Mais surtout, remarquons que si $0 \in I$, alors $\forall t$, $\chi_I(t + \text{sign}(t)) \leq \chi_I(t)$. Ainsi, si les $w_{a,I,p}$ sont positifs, l'approximation vérifie automatiquement les contraintes de décroissance. L'optimisation ne contient plus alors que des contraintes de type *SVM* multiclasses permettant une résolution spécifique :

$$\min_{w \geq 0, \xi \geq 0} \left(\sum_{a,I,p} \widehat{L}(w_{a,I,p}) + C \sum_{\bar{t}} L(\xi(\bar{t})) \right)$$
$$\text{sous contraintes : } \forall \mathcal{V}, \bar{t}, a \neq a^*(\bar{t}),$$
$$\left\langle \left(w_{a^*(\bar{t}),I,p} - w_{a,I,p} \right), N_{p,I \oplus \bar{t}} \right\rangle \geq 1 - \xi(\bar{t})$$

où $I \oplus \bar{t}$ est l'intervalle I translaté de $\bar{t}$ et où chaque composante $N_{p,I \oplus \bar{t}}$ du vecteur $N_{\bar{t}}$ est le nombre de primitives p extraites de la vidéo $\mathcal{V}$ dans l'intervalle $I \oplus \bar{t}$ et où le produit scalaire $\langle\rangle$ se fait sur p et I. La contrainte de positivité des w s'obtient en rajoutant, à l'ensemble des points $N_{\bar{t}}$, un point sur chaque demi-axe négatif de l'espace des primitives.

Cette approximation peut se compléter par l'utilisation d'un sous-échantillonnage des images $\bar{t}$ au niveau des contraintes.

Avec la norme 1 pour L et $\widehat{L}$, cette formulation devient un *lp-SVM* [1]. Alternativement, en utilisant la norme 1 pour L et le carré pour $\widehat{L}$, cette formulation devient un *SVM* standard [3, 7]. Au vu de nos expérimentations, nous recommandons cette dernière formulation (*SVM* standard) qui obtient des performances sensiblement égales à la version sans approximation ($85,1\%$ contre $86,5\%$) pour un apprentissage 50 fois plus rapide sur *honeybee*.

6 Conclusion

Dans cet article, nous étudions la transformée de Hough dans le but de prédire une action au niveau de chaque image dans une vidéo. C'est une méthode de votes dans laquelle chaque vote est classiquement associé à une probabilité observée sur l'ensemble d'apprentissage. Nous proposons d'améliorer cette méthode en optimisant la valeur de tous les votes de façon discriminative. Les performances de notre méthode sont significativement meilleures que celles des autres méthodes de Hough sur le jeu de données *honeybee*. Dans de futurs travaux, nous évaluerons notre méthode dans d'autres contextes tels que le traitement d'image, le traitement de flux sonore...

References

[1] Kristin P Bennett and Olvi L Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1992.

[2] S. Chakrabartty and G. Cauwenberghs. Gini support vector machine: Quadratic entropy based robust multi-class probability regression. *Journal of Machine Learning*, 2007.

[3] Olivier Chapelle and S Sathiya Keerthi. Efficient algorithms for ranking with svms. *Information Retrieval*, 2010.

[4] C. Cortes and V. Vapnik. Support-vector networks. *Journal of Machine learning*, 1995.

[5] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Internationnal Conference on Computer Vision and Pattern Recognition*, 2011.

[6] M. Hoai, Z.Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *Internationnal Conference on Computer Vision and Pattern Recognition*, 2011.

[7] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural svms. *Journal of Machine Learning*, 2009.

[8] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Theory of computing*, 1984.

[9] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision*, 2004.

[10] S. Maji and J. Malik. Object detection using a max-margin hough transform. In *Internationnal Conference on Computer Vision and Pattern Recognition*, 2009.

[11] S.M. Oh, J.M. Rehg, T. Balch, and F. Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *Journal of Computer Vision*, 2008.

[12] Paul Wohlhart, Samuel Schulter, Martin Kostinger, Peter Roth, and Horst Bischof. Discriminative hough forests for object detection. In *Conference of British Machine Vision Conference*, 2012.

[13] Angela Yao, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Does human action recognition benefit from pose estimation? In *Conference of British Machine Vision Conference*, 2011.

[14] Yimeng Zhang and Tsuhan Chen. Implicit shape kernel for discriminative learning of the hough transform detector. In *Conference of British Machine Vision Conference*, 2010.