Détection spatio-temporelle d'actions dans les vidéos de football : l'apport du langage du jeu

Jérémie Ochin, Raphael Chekroun, Bogdan Stanciulescu, Sotiris Manitsaris

DÉTECTION SPATIO-TEMPORELLE D'ACTIONS DANS LES VIDÉOS DE FOOTBALL : L'APPORT DU LANGAGE DU JEU

Jérémie OCHIN, Raphaël CHEKROUN, Bogdan STANCIULESCU, Sotiris MANITSARIS

*CAOR - **Mines Paris - PSL** | Computer Vision team - **Footovision***

Motivations

- La **détection spatio-temporelle des actions (STAD)** dans les sports collectifs est une tâche particulièrement complexe. Les vidéos de sports collectifs sont sujettes à des problèmes tels que le flou de mouvement, l'occlusion, l'ambiguïté visuelle, ou encore la petite taille des joueurs dans l'image…

- L'annotation des matchs de football reste une tâche coûteuse et fortement manuelle, pourtant essentielle pour la production d'analyses

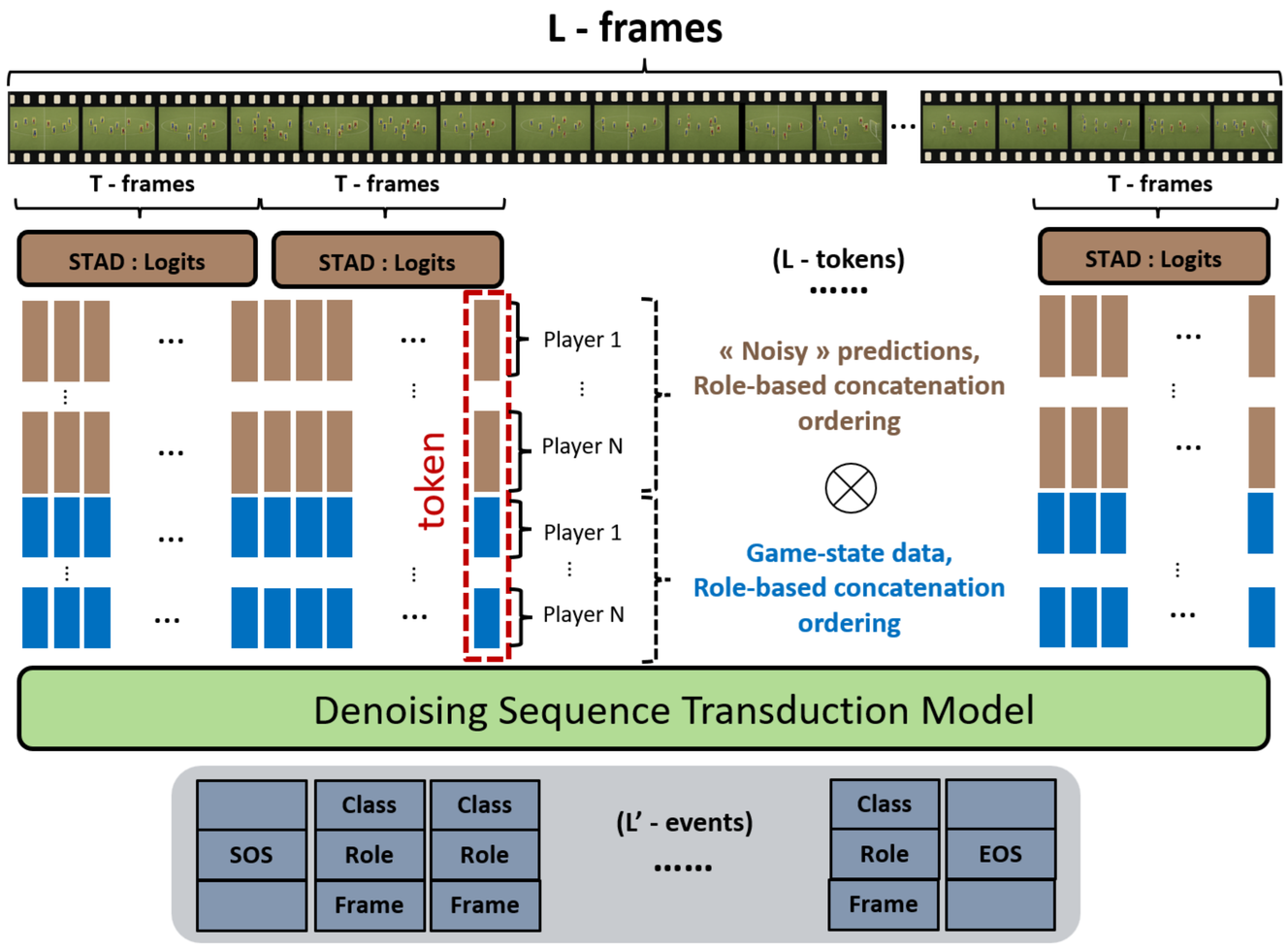
Hypothèses scientifiques

- Limite des méthodes STAD basées sur la vidéo: ces approches s'appuient principalement sur des indices visuels locaux (joueur par joueur, faible horizon temporel). Elles ne peuvent pas comprendre la dynamique du jeu et conduisent à des séquences de détection qui n'ont pas de sens.

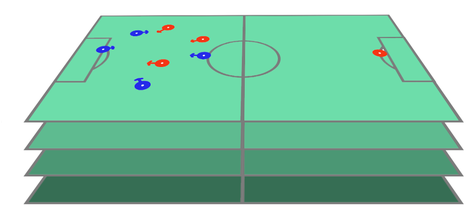
- Existence d'une ``langue tactique'' du football: le football présente une structure séquentielle forte, où les actions des joueurs sont corrélées à des schémas tactiques prédictibles. Cette régularité peut être exploitée comme un ``langage'' pour améliorer la détection des événements.

- Les séquences de prédictions d'actions peuvent être débruitées en utilisant des méthodes de transduction de séquences inspirées du traitement automatique du langage naturel (NLP), par exemple BART.

Méthode



- Un modèle STAD produit d'abord des prédictions d'actions par joueur et par image sur une longue séquence, par morceaux.
- Ces prédictions sont ensuite concaténées selon un ordre cohérent basé sur les rôles des joueurs, à l'aide de métadonnées.
- Chaque token est construit en combinant ces prédictions structurées avec des informations supplémentaires sur l'état du jeu, telles que les positions et vitesses des joueurs.
- La séquence résultante de tokens est ensuite injectée dans un Transformer de type encodeur-décodeur (**Denoising Sequence Transduction – DST**), entraîné à générer de manière auto-régressive des séquences d'actions débruitées — en prédisant, pour chaque action, sa classe, le joueur associé et le numéro d'image.



Sequence of game state

Résultats

Class	Samples	TAAD		TAAD + DST without Game State		TAAD + DST with Game State	
		PR	REC	PR	REC	PR	REC
Pass	26,428	62.6	71.1	79.1	74.5	**83.0**	**79.8**
Ball-drive	20,772	37.5	68.0	74.0	71.6	**79.6**	**77.5**
Header	2,232	26.9	49.4	44.9	44.1	**46.3**	**45.9**
Cross	1,291	60.2	55.8	65.2	65.5	**67.4**	**67.2**
Throw-in	1,149	5.8	24.2	51.7	47.8	**68.5**	**65.3**
Ball-block	764	10.5	44.5	35.8	22.3	**38.2**	**29.6**
Shot	669	47.8	59.3	66.5	59.9	**69.7**	**63.5**
Tackle	134	3.0	6.0	**9.3**	**3.0**	7.0	2.2
Overall	53,439	43.5	66.9	74.1	70.2	**78.7**	**75.8**

	δ = 25 frames						δ = 12 frames					
	L = 100 Frames		L = 250 Frames		L = 750 Frames		L = 100 Frames		L = 250 Frames		L = 750 Frames	
	PR	REC	PR	REC	PR	REC	PR	REC	PR	REC	PR	REC
Overall	77,1	73,3	80,4	75,4	**80,9**	**77,9**	74,9	71,2	78,3	73,4	**78,7**	**75,8**

- Comparaison de la précision (PR) et du rappel (REC) avec un seuil de confiance de 15 % et δ = 12 images, entre la méthode de STAD retenue (TAAD [1]) et deux variantes de notre méthode, avec et sans information sur l'état du jeu. Les résultats sont cohérents pour δ = 25. Les métriques sont multipliées par 100.

- Influence du contexte temporel sur la précision (PR) et du rappel (REC), toutes choses égales par ailleurs. Les métriques sont multipliées par 100.

Références

1 Singh, G., Choutas, V., Saha, S., Yu, F. and Van Gool, L.: Spatio-temporal action detection under large motion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 5998-6007. IEEE, Waikoloa, HI, USA, (2023)

2 Simpson, I., Beal, R.J., Locke, D., and Norman, T.J.: Seq2Event: Learning the Language of Soccer Using Transformer-based Match Event Prediction. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3898–3908. Association for Computing Machinery, Washington DC, USA, (2022)

3 Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871--7880. Association for Computational Linguistics, Online (2020)