

Titre: Reconnaissance automatisée des activités et des comportements humains à travers l'analyse de vidéos

Auteur: Soufiane Lamghari

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Lamghari, S. (2023). Reconnaissance automatisée des activités et des comportements humains à travers l'analyse de vidéos [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. <https://publications.polymtl.ca/55100/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/55100/>

Directeurs de recherche: Guillaume-Alexandre Bilodeau, & Nicolas Saunier

Programme: Génie informatique

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Reconnaissance automatisée des activités et des comportements humains à
travers l'analyse de vidéos**

SOUFIANE LAMGHARI

Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Août 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

**Reconnaissance automatisée des activités et des comportements humains à
travers l'analyse de vidéos**

présentée par **Soufiane LAMGHARI**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Farida CHERIET, présidente

Guillaume-Alexandre BILODEAU, membre et directeur de recherche

Nicolas SAUNIER, membre et codirecteur de recherche

Lama SÉOUD, membre

Christian DESROSIERS, membre externe

DÉDICACE

*À ma mère et mon père,
avec amour et gratitude éternels...*

REMERCIEMENTS

Je tiens d’abord à remercier sincèrement mon directeur de recherche, Guillaume-Alexandre Bilodeau, pour sa présence, son expertise, sa patience et son dévouement. Votre encadrement rigoureux et vos conseils précieux ont grandement contribué à orienter mes réflexions de manière claire et cohérente. Je suis extrêmement reconnaissant d’avoir eu la chance de travailler à vos côtés. Je tiens également à remercier profondément mon codirecteur de recherche, Nicolas Saunier, pour sa disponibilité, son soutien inestimable tout au long de mon doctorat, ainsi que pour les discussions inspirantes et constructives lors de nos rencontres. Ses précieux commentaires et conseils ont grandement enrichi mes recherches.

J’aimerais également remercier les membres de mon comité de thèse pour avoir accepté de consacrer leur temps et leur expertise à l’évaluation de ce travail de recherche.

Je souhaite aussi exprimer ma reconnaissance envers le conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) pour son précieux soutien.

Je tiens à remercier profondément ma famille, qui a toujours cru en moi et m’a soutenu tout au long de mon parcours académique. J’aimerais également remercier mes collègues du LITIV pour les bons moments que nous avons passés ensemble.

RÉSUMÉ

La reconnaissance d'activités humaines dans les vidéos revêt une importance majeure en offrant une compréhension automatisée du comportement humain dans une variété de contextes. Elle permet de détecter, analyser et interpréter les activités des individus, ouvrant ainsi la voie à un large éventail d'applications. Par exemple, dans le domaine de l'aménagement urbain, la reconnaissance d'activités permet de détecter et comprendre les comportements des usagers des espaces publics. Elle peut ainsi aider les experts en aménagement urbain à prendre des décisions éclairées qui favorisent la conception d'aménagements à la fois fonctionnels et adaptés aux besoins de la population. Toutefois, reconnaître les activités humaines dans des environnements réalistes peut être problématique à cause de plusieurs défis tels que l'occlusion, les conditions d'éclairage, et la variation d'apparence intra-classe.

Dans cette thèse, nous présentons différentes méthodes permettant une reconnaissance automatisée des activités à partir de séquences vidéos. Les approches proposées visent à relever les défis spécifiques rencontrés dans des environnements réalistes et complexes comme les espaces publics. Ces approches se basent sur l'idée que l'intégration de l'information de pose humaine fournit des caractéristiques précieuses de l'activité. Aussi, se concentrer sur les informations d'apparence spatio-temporelles les plus discriminantes permet de mieux comprendre et reconnaître les activités dans les séquences vidéos.

La première méthode que nous proposons est une méthode de reconnaissance d'activités humaines basée sur la pose, qui extrait de manière efficace les caractéristiques spatio-temporelles les plus discriminantes des activités et les encodent en une représentation compacte en grille. Notre approche se concentre sur l'analyse des informations les plus pertinentes liées à l'activité, en mettant en évidence les poses représentatives à l'aide d'un mécanisme d'attention explicite. Ce mécanisme permet de gérer efficacement les situations difficiles, telles que les occlusions et les variations intra-classes.

Ensuite, nous abordons les défis spécifiques liés à la reconnaissance des activités dans des séquences vidéos où seule l'activité principale est étiquetée. Nous avons utilisé des séquences vidéos dans le spectre infrarouge comme cas d'étude. En plus de requérir la localisation de l'activité principale, la reconnaissance d'activités humaines dans le spectre infrarouge présente des particularités et des défis uniques, tels que les occlusions partielles et les apparences similaires des individus présents dans la scène observée, qui peuvent rendre difficile la distinction des activités effectuées. Pour relever ces défis, nous proposons une architecture de reconnaissance d'activités adaptée à ces deux situations. Cette architecture permet d'apprendre

simultanément à détecter les activités humaines dans le spectre infrarouge et à identifier automatiquement les acteurs clés qui réalisent ces activités, sans recourir à des connaissances préalables ou à des annotations explicites. Afin de compenser l'absence d'informations de texture et de couleur dans le spectre infrarouge, nous exploitons l'information de la pose humaine pour extraire des informations riches. De plus, nous effectuons une sélection plus précise des poses clés afin de limiter l'analyse aux caractéristiques d'activité temporelles pertinentes, minimisant ainsi la complexité du modèle. Finalement, nous utilisons les poses clés résultantes pour construire une représentation qui encode à la fois l'information infrarouge et l'information de la pose.

Étant donné que la précision de la reconnaissance d'activités humaines dépend à la fois de la complexité des mouvements et de leurs durées, le dernier travail présenté consiste en une méthode de reconnaissance d'activités humaines qui permet de modéliser efficacement les dépendances temporelles à court terme et à long terme. Grâce à notre méthodologie de regroupement de pose non-paramétrique, notre approche est capable de gérer efficacement des séquences représentant des activités de longueurs variables, ce qui la rend particulièrement pratique pour les scénarios de reconnaissance d'activités réalistes. De plus, nous entraînons un réseau auto-attentif pour capturer les caractéristiques représentatives des activités.

Nos évaluations approfondies sur plusieurs jeux de données représentant des scènes dans des espaces publics confirment les bonnes performances de nos méthodes proposées dans cette thèse. Elles ont démontré leurs capacités à reconnaître efficacement les actions humaines dans des environnements réalistes non-contrôlés. De plus, nos approches se sont révélées robustes face aux multiples défis auxquels peuvent faire face les systèmes de reconnaissance d'activités. Même en présence d'occlusions partielles, de variations d'apparence au sein d'une même classe, ainsi que dans des scénarios utilisant le spectre infrarouge, nos méthodes ont démontré leur capacité à maintenir des performances de reconnaissance d'activité élevées.

ABSTRACT

Human activity recognition in videos is of great importance as it provides automated understanding of human behaviour in a variety of contexts. It enables the detection, analysis, and interpretation of the activities of individuals, thereby paving the way for a wide range of applications. For example, in the field of urban planning, activity recognition helps detect and understand the behaviours of users in public spaces. It can assist urban planners in making informed decisions that promote the design of functional and population-tailored environments. However, recognizing human activities in realistic environments can be challenging due to several obstacles such as occlusion, lighting conditions, and intra-class appearance variation.

In this thesis, we present various methods for automated activity recognition from video sequences. The proposed approaches aim to address specific challenges encountered in realistic and complex environments, such as public spaces. These approaches are based on the idea that integrating human pose information provides valuable activity features. Additionally, focusing on the most discriminative spatio-temporal appearance information helps to better understand and recognize activities in video sequences.

The first method we propose is a pose-based human activity recognition method that efficiently extracts the most discriminative spatio-temporal features of activities and encodes them into a compact grid representation. Our approach focuses on analyzing the most relevant activity-related information, highlighting representative poses using an explicit attention mechanism. This mechanism effectively handles challenging situations, such as occlusions and intra-class variations.

Next, we address the specific challenges related to video where only the main activity is labelled. We use videos in the infrared spectrum for this study. In addition of requiring to identify the main activity, human activity recognition in the infrared spectrum presents unique characteristics and challenges, such as partial occlusions and similar appearances of individuals in the observed scene, which can make it difficult to distinguish performed activities. To overcome these challenges, we propose an activity recognition architecture tailored to address them. This architecture enables simultaneous learning of detecting human activities in this spectrum and automatically identifying key actors performing these activities, without relying on prior knowledge or explicit annotations. To compensate for the lack of texture and colour information in the infrared spectrum, we leverage human pose information to extract rich features. Furthermore, we perform more precise selection of key poses to limit

the analysis to relevant temporal activity characteristics, thereby minimizing the complexity of the model. Finally, we use the resulting key poses to construct a representation that encodes both the infrared information and pose information.

Since the accuracy of human activity recognition depends on both the complexity of movements and their duration, the last proposed work consists of a method for human activity recognition that effectively models short-term and long-term temporal dependencies. With our non-parametric pose clustering methodology, our approach can efficiently handle sequences of variable lengths, making it particularly practical for realistic activity recognition scenarios. Additionally, we train a Transformer network to capture representative features of activities.

Our thorough evaluations on multiple datasets representing scenes in public spaces confirm the good performance of our proposed methods in this thesis. They have demonstrated their ability to efficiently recognize human actions in realistic and uncontrolled environments. Furthermore, our approaches have proven to be robust against multiple challenges that action recognition systems may encounter. Even in the presence of partial occlusions, appearance variations within the same class, as well as scenarios using the infrared spectrum, our methods have shown a great capacity in maintaining high activity recognition performance.

TABLE DES MATIÈRES

DÉDICACE	iii
REMERCIEMENTS	iv
RÉSUMÉ	v
ABSTRACT	vii
TABLE DES MATIÈRES	ix
LISTE DES TABLEAUX	xiii
LISTE DES FIGURES	xiv
LISTE DES SIGLES ET ABRÉVIATIONS	xvii
CHAPITRE 1 INTRODUCTION	1
1.1 Définitions et concepts de base	2
1.2 Éléments de la problématique	5
1.2.1 Reconnaissance des activités humaines dans le spectre visible	5
1.2.2 Reconnaissance des activités humaines principales dans le spectre infrarouge thermique	7
1.2.3 Modélisation temporelle des actions à court et à long terme	8
1.3 Objectifs de recherche	8
1.4 Contributions	9
1.5 Plan de la thèse	10
CHAPITRE 2 REVUE DE LITTÉRATURE	12
2.1 Méthodes classiques	12
2.1.1 La détection des humains	12
2.1.2 Approches globales	16
2.1.3 Approches locales	18
2.1.4 Discussion	20
2.2 Méthodes basées sur l'apprentissage profond	20
2.2.1 Introduction	21
2.2.2 Architectures spatiales	21

2.2.3	Architectures spatio-temporelles	24
2.2.4	Architectures à deux flux spatial et temporel	27
2.2.5	Architectures spatiales et temporelles à long terme	29
2.2.6	Architectures auto-attentives	31
2.2.7	Discussion	38
2.3	Méthodes de reconnaissance d'activité basées sur l'apprentissage profond ver- sus méthodes classiques	39
2.4	Ensembles de données	40
2.4.1	Environnements contrôlés	40
2.4.2	Environnements réalistes	41
CHAPITRE 3 SURVOL DES MÉTHODES PROPOSÉES		45
CHAPITRE 4 ARTICLE 1 : A GRID-BASED REPRESENTATION FOR HUMAN ACTION RECOGNITION		48
4.1	Introduction	48
4.2	Related work	50
4.2.1	Deep Learning-based Action Recognition	50
4.2.2	Pose-based Video Action Recognition	51
4.2.3	Collective Activity Recognition	51
4.3	Grid-based Representation For Human Action Recognition	52
4.3.1	Human Pose Estimation	53
4.3.2	Relevant Features Selection	54
4.3.3	Grid Representation Learning	55
4.4	Experiments	57
4.4.1	Datasets	58
4.4.2	Implementation Details	58
4.4.3	Ablation Studies	59
4.4.4	Comparison with the State-of-the-Art	61
4.5	Conclusion	62
CHAPITRE 5 ARTICLE 2 : ACTOR-DRIVEN POSE EMBEDDINGS FOR VIDEO ACTION RECOGNITION		64
5.1	Introduction	64
5.2	Related work	67
5.2.1	Human Action Recognition in Videos	67
5.2.2	Pose-based Action Recognition	69

5.3	Proposed method	69
5.3.1	Overview	69
5.3.2	Key-Actor Identification	70
5.3.3	Human Pose Estimation	72
5.3.4	Scale-invariant Polar Encoding and Embedded Pose Filtering	73
5.3.5	Human Pose Features Selection	73
5.3.6	Human Action Representation Learning	75
5.4	Experiments	76
5.4.1	Experimental Settings	76
5.4.2	Ablation Study	76
5.4.3	Comparison with the State-of-the-Art Methods	78
5.5	Conclusion	78
CHAPITRE 6 ARTICLE 3 : TRANSFORMER-BASED HUMAN ACTION RECOGNITION WITH DYNAMIC FEATURE SELECTION		81
6.1	Introduction	81
6.2	Related Work	83
6.3	Proposed Method	84
6.3.1	Overview	84
6.3.2	Human Pose Estimation	85
6.3.3	Key Pose Identification	86
6.3.4	Object Embedding	87
6.3.5	Transformer-based Action Representation Learning	88
6.4	Experiments	89
6.4.1	Datasets	89
6.4.2	Implementation Details	90
6.4.3	Ablation studies	90
6.4.4	Comparison with the State-of-the-Art	92
6.5	Conclusion	94
CHAPITRE 7 DISCUSSION GÉNÉRALE		95
7.1	Reconnaissance des activités humaines dans le spectre visible	95
7.2	Reconnaissance des activités avec identification d'acteur principal	96
7.3	Modélisation temporelle des activités à court et à long terme	97
7.4	Utilité des différents éléments	98
CHAPITRE 8 CONCLUSION		100

8.1 Synthèse des travaux	100
8.2 Recommandations pour travaux futurs	101
RÉFÉRENCES	103

LISTE DES TABLEAUX

Tableau 4.1	Impact of the clustering algorithms and number of key poses on the performance of our model on the CAE dataset. Boldface : Best result.	59
Tableau 4.2	Impact of different modules on the accuracy of GRAR based on CAE dataset. Boldface : Best result.	61
Tableau 4.3	Comparison of the activity recognition performance of state-of-the-art methods versus our model evaluated on CA dataset. Boldface : Best result.	63
Tableau 4.4	Comparison of the activity recognition performance of state-of-the-art methods versus our model evaluated on CAE dataset. Boldface : Best result.	63
Tableau 4.5	Evaluation of action recognition performance of state-of-the-art methods versus our proposed model on the Volleyball dataset. Boldface : Best result.	63
Tableau 5.1	Performance contribution of each component in ActAR on InfAR dataset.	76
Tableau 5.2	Comparison of the average recognition precision of ActAR with state-of-the-art methods on InfAR dataset.	79
Tableau 6.1	Impact of the integration of different modules on the recognition performance of our model on the CAE dataset.	92
Tableau 6.2	Comparison of the performance of our model with state of the art models on the CA Dataset. Boldface : Best results.	93
Tableau 6.3	Comparison of the performance of our model with state of the art models on the CAE Dataset. Boldface : Best results.	93
Tableau 6.4	Comparison of the performance of our model with state of the art models on the UCLA Courtyard Dataset. Boldface : Best results.	94

LISTE DES FIGURES

Figure 1.1	Catégories des activités humaines.	2
Figure 1.2	Exemple d'une trame du jeu de données Collective Activity représentant plusieurs individus dans un espace public [1].	3
Figure 1.3	Un résultat d'estimation de pose en 2D sur une trame du jeu de données Collective Activity Extended [2].	5
Figure 2.1	Exemple de soustraction d'arrière-plan avec la méthode ViBe [3]. . .	14
Figure 2.2	Détection d'humain avec flot optique [4].	14
Figure 2.3	Images MEI et MHI pour représenter trois différents types de mouvements [5].	17
Figure 2.4	Représentations spatio-temporelles pour trois différents types d'activités [6].	17
Figure 2.5	Processus de reconnaissance des activités humaines par l'approche de sac de mots visuels [7].	19
Figure 2.6	Descripteurs spatio-temporels par trajectoires denses [8].	20
Figure 2.7	L'architecture CNN originale du réseau LeNet-5 [9].	22
Figure 2.8	Les différentes approches de fusion pour donner une représentation au niveau de la vidéo [10].	24
Figure 2.9	L'architecture du réseau CNN à deux flux basés uniquement sur les informations spatiales [10].	25
Figure 2.10	Opération de convolution en 3D comparée à la convolution 2D [11]. .	26
Figure 2.11	L'architecture du réseau CNN 3D C3D [12].	26
Figure 2.12	L'architecture CNN à deux flux spatial et temporel [13].	27
Figure 2.13	Schéma global du réseau de segments temporels à deux flux (TSN) [14].	29
Figure 2.14	L'architecture d'un réseau LSTM [15].	30
Figure 2.15	L'architecture du réseau de convolution récurrent (LRCN) [16]. . . .	31
Figure 2.16	Modèle d'attention appliqué à la reconnaissance d'activités humaines [17].	32
Figure 2.17	L'architecture du modèle auto-attentif [18].	33
Figure 2.18	Illustration du mécanisme d'attention de produit scalaire pondéré (Scaled Dot-Product Attention) (à gauche), et du mécanisme d'attention multi-têtes (Multi-Head Attention) (à droite) [18].	34
Figure 2.19	Architecture du modèle ViT [19].	35
Figure 2.20	Architecture du modèle Action Transformer Network [20].	36
Figure 2.21	Architecture du modèle RViT [21].	37

Figure 2.22	Architecture du modèle VidTr [22].	38
Figure 2.23	Exemples de l'ensemble de données "KTH" [23].	40
Figure 2.24	Exemples de l'ensemble de données "Weizmann" [6].	41
Figure 2.25	Exemples de l'ensemble de données "Collective Activity" [1].	41
Figure 2.26	Un exemple d'image du jeu de données "Volleyball" [24].	42
Figure 2.27	Un exemple d'images du jeu de données "InfAR" [25].	43
Figure 2.28	Un exemple d'image du jeu de données "UCLA Courtyard" [26].	44
Figure 4.1	The pipeline of our proposed GRAR model. From each frame, we extract human poses using a pose estimation method. We select then the most representative key poses for each human based on an unsupervised clustering method. These estimated key poses are combined with their corresponding RGB sources and then concatenated. This results in a compact grid representation G_h , that encodes the relevant RGB and pose information related to the performed action. Finally, we train a deep convolutional neural network on the obtained grid representations to predict the corresponding action category A_h	52
Figure 4.2	Examples from the Volleyball and CAE datasets of (a) a pose estimation failure case considered as outlier during key poses selection. (b) and (c) show occluded humans handled with our explicit attention on pose.	56
Figure 4.3	Example of a grid image from the CAE dataset.	57
Figure 5.1	Example of two actions involving single and multiple persons from the InfAR dataset labeled as "skip". (a) Considering all people in a scene to recognize an action can be uninformative. (b) Identifying the key-actor performing the main action provides accurate guidance for the action recognition model.	66
Figure 5.2	An overview of the proposed ActAR model. Given an infrared input video, we identify the key-actors based on optical flow. To compensate for the missing texture and color information in the infrared domain, we extract the human body keypoints. We project the estimated keypoints into the polar system, and then we filter out incorrect poses that may mislead the action representation. Next, we select the key-poses based on a deep clustering method to form a compact representation encoding the most relevant infrared and pose information. In the last stage, we train a 2D deep convolutional neural network to predict the human action.	71

Figure 5.3	Impact of different pose clustering algorithms on the performance of ActAR on InfAR dataset.	79
Figure 6.1	The pipeline of our proposed ActiViT model. The model comprises four stages. In the first stage, we estimate the pose keypoints of all individuals in each frame of a video sequence. The second stage uses a non-parametric approach to cluster the extracted body poses, enabling dynamic key pose selection. In the third stage, we localize and recognize objects appearing in the scene that individuals interact with. Finally, we feed the resulting key image patches, along with their corresponding objects embeddings, to the transformer model to learn discriminative features for improved action recognition.	85
Figure 6.2	Overview of the action Vision Transformer model. E_{pos} is a positional embedding vector and A_{cls} is a learnable embedding token.	89
Figure 6.3	Impact of the clustering strategy on the performance of our model on the CAE dataset. "NP" refers to our model using the non-parametric approach, and "P_5", "P_7", "P_10", "P_12", refer to the parametric approach with fixed numbers of poses $\in \{5, 7, 10, 12\}$	91

LISTE DES SIGLES ET ABRÉVIATIONS

CNN	<i>Convolutional Neural Network</i>
EM	<i>Expectation–Maximization</i>
GRU	<i>Gated Recurrent Units</i>
HAR	<i>Human Action Recognition</i>
HOF	<i>Histogram of Optical Flow</i>
HOG	<i>Histogram of Oriented Gradients</i>
HRNet	<i>High-Resolution Network</i>
GMM	<i>Gaussian Mixture Model</i>
I3D	<i>Inflated 3D Convnet</i>
KLT	<i>Kanade-Lucas-Tomasi</i>
LN	<i>Layer Normalization</i>
LSTM	<i>Long Short-Term Memory</i>
LRCN	<i>Long-term Recurrent Convolutional Network</i>
MBH	<i>Motion Boundary Histogram</i>
MEI	<i>Motion Energy Image</i>
MHI	<i>Motion History Image</i>
MHV	<i>Motion History Volume</i>
MLP	<i>Multi-Layer Perceptron</i>
MSA	<i>Multi-head Self-Attentions</i>
OF-MHI	<i>Optical flow-motion history image</i>
PAM	<i>Partitioning Around Medoids</i>
Relu	<i>Rectified Linear Unit</i>
RGB	<i>Red Green Blue</i>
RNN	<i>Recurrent Neural Network</i>
SGD	<i>Stochastic Gradient Descent</i>
SVM	<i>Support Vector Machine</i>
TSN	<i>Temporal Segment Network</i>
ViT	<i>Vision Transformer</i>

CHAPITRE 1 INTRODUCTION

L’urbanisation connaît actuellement une croissance rapide à l’échelle mondiale, avec environ 55% de la population mondiale résidant dans les villes, selon les données de l’Organisation des Nations Unies [27]. Pour accompagner cette évolution, il est essentiel de mettre en place des plans d’aménagement efficaces, capables de guider et de modeler l’évolution des villes de manière durable et équilibrée. Des aménagements urbains de qualité, tels que des espaces verts bien entretenus, des infrastructures de transport efficaces et des logements adéquats, contribuent à améliorer notre quotidien. L’objectif derrière ces planifications est de développer des villes dynamiques, sûres, durables et saines. Cependant pour y arriver, il faut pouvoir mesurer la qualité des aménagements existants et futurs. Comme les espaces publics et les rues constituent une partie majeure de ces aménagements, il s’avère donc nécessaire de mesurer comment ces espaces sont utilisés, notamment en ce qui concerne leur capacité à répondre aux besoins de leurs usagers et à encourager des pratiques socialement et physiquement bénéfiques, par exemple.

La disponibilité croissante de capteurs vidéos et de caméras dans les environnements urbains offre une opportunité précieuse pour collecter des données riches sur l’activité des personnes dans les espaces publics, fournissant ainsi des informations essentielles pour comprendre l’utilisation de ces aménagements urbains. Cependant, le volume important de ces données visuelles rend leur traitement manuel difficile. Il est donc impératif de recourir à des méthodes informatiques avancées pour un traitement automatique.

Grâce aux avancées technologiques dans le domaine de la vision par ordinateur, en particulier celui du traitement vidéo, et grâce à la puissance de calcul des nouveaux ordinateurs, la tâche d’extraction automatique des connaissances à partir d’une séquence vidéo est devenue possible. Cela ouvre ainsi la voie à de nouvelles possibilités d’analyse et d’utilisation des données visuelles à grande échelle. Dans ce contexte, nous nous intéressons dans cette thèse à reconnaître les comportements des usagers dans les espaces publics et les rues à travers des données vidéos, afin de faciliter l’aménagement urbain.

En se basant sur la sortie fournie par notre système, les experts en transport et en aménagement urbain pourront évaluer la manière par laquelle les aménagements urbains sont utilisés. Ils pourront également diagnostiquer les problèmes potentiels présents dans ces espaces (sécurité routière, cheminement des piétons, etc.) afin qu’ils puissent proposer des stratégies qui assureront des aménagements plus agréables et adéquats aux besoins des citoyens.

1.1 Définitions et concepts de base

Le sujet de la reconnaissance d'activités humaines auquel nous nous intéressons est un sujet très actif en vision par ordinateur. Il consiste à identifier et classifier les activités exercées par des êtres humains à partir d'une séquence d'observations visuelles. Dans un système de vision par ordinateur, la reconnaissance d'activités humaines constitue l'étape finale d'un processus comprenant plusieurs tâches préliminaires qui viennent après l'acquisition des données. Ces tâches incluent la segmentation, la détection et le suivi des humains.

Les activités humaines ont une structure hiérarchique inhérente à différents niveaux dépendamment de leur durée et complexité. Ces activités sont considérées dans une catégorisation à quatre niveaux comme illustrée dans la figure 1.1.

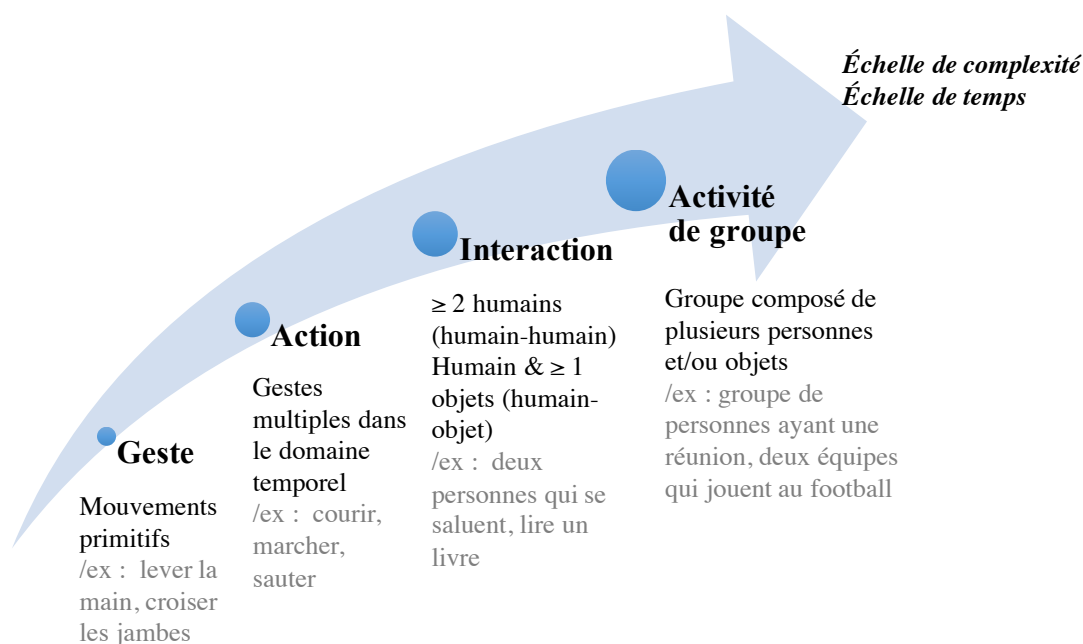


FIGURE 1.1 Catégories des activités humaines.

Ces types d'activités vont des gestes jusqu'aux activités de groupes définies plus précisément de la façon suivante :

Les gestes : ce sont des mouvements primitifs des parties du corps d'une personne et correspondent à une action particulière telle que "lever la main" ou "croiser les jambes". Leur complexité est assez basse et ils sont exercés sur de courtes périodes de temps.

Les actions : elles sont plus complexes et représentent la suite de plusieurs gestes pour

une période de temps donnée. Par exemple "marcher", "courir" ou "agiter la main".

Les interactions : les interactions humain-humain et humain-objet sont des activités humaines où au moins deux personnes et/ou des objets interagissent. Par exemple "deux personnes qui se saluent" est une interaction de type humain-humain et "une personne lisant un livre" est une interaction de type humain-objet.

Les activités de groupes : sont les plus complexes dans le temps, elles représentent les activités exercées par deux humains ou plus et peuvent inclure également les interactions avec les objets. Par exemple "un groupe de personnes ayant une réunion" ou "deux équipes qui jouent au football".

Dans cette thèse, nous considérons plusieurs types d'activités dans des environnements réalistes non contrôlés, où plusieurs personnes sont impliquées dans des activités distinctes en parallèle. Un exemple représentatif d'un tel environnement est illustré dans la figure 1.2.



FIGURE 1.2 Exemple d'une trame du jeu de données Collective Activity représentant plusieurs individus dans un espace public [1].

La reconnaissance d'activités humaines implique généralement l'extraction de caractéristiques distinctives, la représentation et la classification des activités. La première étape consiste à extraire des caractéristiques discriminantes à partir de chaque séquence. Ces caractéristiques peuvent être liées aux apparences visuelles, telles que les couleurs, les textures ou les formes, ainsi qu'aux mouvements. Elles doivent être soigneusement choisies ou apprises pour capturer

les informations pertinentes qui permettront de distinguer efficacement les différentes activités effectuées. Ensuite, il est nécessaire de les transformer en un vecteur de caractéristiques adéquat qui servira d'entrée au classificateur. Cette représentation peut prendre différentes formes, telles que la réduction de dimensionnalité, la normalisation ou la fusion des caractéristiques, et doit être compréhensible et exploitable par les algorithmes de classification. Enfin, la dernière étape consiste à utiliser la représentation des caractéristiques obtenues afin de catégoriser les activités observées. Dans ce contexte, les classificateurs permettent d'apprendre les caractéristiques relatives à chaque activité à partir de séquences préalablement étiquetées et doivent être capables de généraliser les relations et les caractéristiques apprises pour classer les activités des individus dans de nouvelles séquences vidéos. Chacune de ces étapes contribue de manière significative à la qualité de la reconnaissance d'activités, mais il est essentiel de souligner que les performances globales d'un modèle de reconnaissance d'activités reposent largement sur la qualité des représentations des activités. En effet, la façon dont les caractéristiques des activités sont extraites et représentées peut avoir un impact significatif sur la capacité du modèle à identifier et à classer correctement celles-ci.

L'estimation de la pose humaine permet de caractériser et de localiser les positions des articulations du corps humain dans des images ou des vidéos. Les recherches dans ce domaine démontrent les avantages considérables de cette technique pour de nombreuses tâches et applications en vision par ordinateur [28, 29]. Dans le cas de la reconnaissance d'activités par exemple, l'estimation de pose peut fournir des informations précieuses sur les activités réalisées. En effet, en identifiant les positions spatiales des mains, des jambes et des autres parties du corps, il devient possible de différencier entre plusieurs activités qui peuvent être décrites par les mouvements de ces parties, telles que sauter, courir, danser ou d'autres activités spécifiques.

L'extraction de la pose est généralement réalisée en 3D, ce qui permet de représenter la position des individus dans un espace tridimensionnel [30, 31]. Cette approche nécessite toutefois l'accès à des capteurs spéciaux capables de fournir des mesures de profondeur. En conséquence, l'acquisition de pose en 3D peut s'avérer non pratique dans les scénarios réalistes, notamment, en raison de son coût élevé en termes de matériel et de temps de calcul. En alternative, l'estimation de pose en 2D peut être bénéfique dans ce sens, car elle est basée uniquement sur la position relative des individus dans les séquences d'images en entrée et ne nécessite aucun matériel spécifique [32, 33]. La figure 1.3 illustre un exemple de pose estimée en 2D d'un individu qui court. Motivés par l'utilité de la pose sur la reconnaissance d'activités humaines, nous proposons de nouvelles méthodes basées sur cette technique pour extraire les caractéristiques spatiales et les mouvements du corps de chaque individu afin de mieux reconnaître leurs activités.

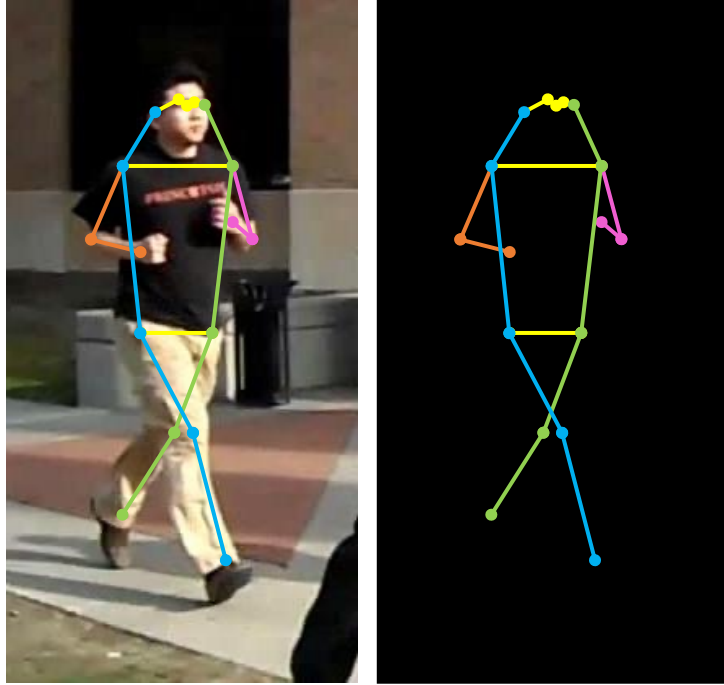


FIGURE 1.3 Un résultat d'estimation de pose en 2D sur une trame du jeu de données Collective Activity Extended [2].

1.2 Éléments de la problématique

La reconnaissance d'activités humaines à partir de vidéos dans des environnements réalistes tels que les espaces publics présente de nombreux défis complexes. Malgré les avancées scientifiques considérables réalisées dans le domaine de la vision par ordinateur, plusieurs problématiques inhérentes à cette tâche demeurent encore irrésolues. Dans ce qui suit, nous détaillerons les principales problématiques que nous avons tenté de résoudre dans le cadre de cette thèse.

1.2.1 Reconnaissance des activités humaines dans le spectre visible

La reconnaissance d'activités est une tâche intrinsèquement complexe, et assurer sa précision dans des environnements réalistes représente un véritable défi. L'occlusion constitue l'un des problèmes majeurs qui compliquent cette tâche. En effet, l'occlusion se produit lorsqu'un individu d'intérêt est partiellement ou totalement caché à un moment donné par un autre objet ou un autre individu présent dans la même scène. Lorsqu'un individu est partiellement occulté, il devient difficile d'extraire les caractéristiques visuelles nécessaires à la compréhension de son activité, alors qu'en cas d'occlusion totale, l'individu peut être temporairement

perdu de vue, ce qui peut entraîner des erreurs dans la reconnaissance de son activité. En plus, le problème d’occlusion est d’autant plus prononcé dans les scènes où plusieurs personnes ou objets sont en mouvement, car ils peuvent se superposer aux autres individus en train d’exercer des activités, ce qui entraînera ainsi une confusion dans la reconnaissance de l’activité. Par exemple, la reconnaissance de l’activité d’un individu qui traverse la route ou qui s’est simplement arrêté pour attendre peut être difficile si celui-ci est partiellement occulté par des véhicules en mouvement. L’occlusion causée par la présence de ces véhicules peut rendre la distinction entre ces deux activités problématique.

Les variations d’apparence intra-classe constituent d’autres défis que nous devons résoudre afin d’assurer une compréhension précise des comportements des individus dans les espaces publics. En effet, plusieurs personnes peuvent exécuter la même activité tel que "Courir", mais leurs apparences peuvent différer en termes de couleurs et style de vêtement, de styles de course, etc. Ces variations d’apparence pour la même catégorie d’activité rendent la tâche de reconnaissance d’activités plus complexe. Dans ce contexte, un bon modèle de reconnaissance d’activités humaines doit être capable de généraliser et de prendre en compte ces variations pour identifier correctement l’activité réalisée, même lorsque les apparences des individus varient considérablement.

En plus des problèmes d’occlusion et d’apparence intra-classe, les conditions lumineuses et météorologiques posent également un défi pour la reconnaissance d’activités humaines. En particulier dans les espaces publics, nous sommes confrontés à des variations de luminosité, causées par des facteurs tels que l’ombre des objets, passage de nuages ou la transition entre le jour et la nuit. Ces variations peuvent avoir un impact significatif sur la qualité des vidéos capturées, ce qui complique l’extraction de caractéristiques discriminantes pour l’activité en cours. Les conditions météorologiques telles que la pluie, le brouillard ou la neige rendent difficile la reconnaissance d’activités humaines, car elles entraînent une perte d’informations sur l’activité et son contexte. Ainsi, il devient plus complexe de distinguer les actions et les comportements des individus dans de telles conditions. Dans ce sens, il est essentiel que le modèle de reconnaissance soit robuste face à ces variations et capable d’inférer les activités humaines malgré le manque d’informations visuelles claires.

En outre, les méthodes conventionnelles de reconnaissance d’activités humaines utilisent des trames individuelles en RGB pour représenter l’action de chaque individu et ne modélisent pas correctement l’information temporelle [34]. Cependant, la reconnaissance d’activités nécessite souvent une compréhension globale de la dynamique des mouvements, car le facteur le plus important pour la compréhension des activités réside dans les indices de mouvement qui existent implicitement dans la dimension temporelle. En se basant uniquement sur

des images individuelles sélectionnées aléatoirement, ces méthodes ne considèrent pas adéquatement l’aspect temporel de l’activité, ce qui peut entraîner une perte d’informations contextuelles essentielles pour une reconnaissance précise des activités.

1.2.2 Reconnaissance des activités humaines principales dans le spectre infrarouge thermique

Le spectre infrarouge thermique (dans la suite de la thèse, par infrarouge, on entend infrarouge thermique) peut être une alternative intéressante pour résoudre quelques défis liés à la reconnaissance d’activités dans le spectre visible. En effet, les caméras infrarouges peuvent capturer des informations qui ne sont pas visibles à l’œil nu, ce qui les rend particulièrement adaptées aux conditions de faible luminosité ou d’obscurité totale. Ainsi, elles sont résistantes aux éléments perturbateurs de l’arrière-plan, ce qui les rend moins sensibles aux variations présentes dans la scène. De plus, les caméras infrarouges sont moins intrusives sur la vie privée par rapport aux autres types de caméras, car elles ne captent pas d’images détaillées des individus, mais plutôt des émissions thermiques, ce qui réduit les risques de reconnaissance ou d’identification personnelle. Cependant, la reconnaissance d’activités dans le spectre infrarouge présente également ses propres défis. Premièrement, les méthodes d’extractions de caractéristiques ne peuvent pas être directement appliquées aux vidéos infrarouges, notamment lorsqu’il s’agit d’estimer la pose des individus. Comme la plupart des modèles d’estimation de pose sont entraînés sur des données du spectre visible, ils peuvent échouer dans certains cas lorsqu’ils sont directement appliqués au domaine infrarouge en raison des variations spécifiques des caractéristiques des images infrarouges, qui diffèrent significativement des images RGB.

Dans le spectre infrarouge, l’occlusion demeure un défi majeur à relever. Comparée aux caméras du spectre visible, la résolution spatiale des caméras infrarouges peut être relativement basse. Cela signifie que les détails fins du contexte peuvent être moins visibles, ce qui rend la distinction entre les actions des individus occultés plus difficile. Il est aussi courant que dans le spectre infrarouge, différents individus aient des apparences thermiques similaires. Cela peut rendre la tâche de reconnaissance plus complexe. Par conséquent, il devient plus difficile de se fier uniquement aux caractéristiques de texture qui sont indiscernables et redondants. Cela souligne la nécessité de prendre en compte d’autres informations dans le processus de reconnaissance d’activités en infrarouge, afin d’améliorer sa précision.

1.2.3 Modélisation temporelle des actions à court et à long terme

La complexité de la tâche de reconnaissance des activités humaines peut dépendre de la nature des mouvements ainsi que de leurs durées. Par exemple, l'activité "applaudir" peut être détectée avec un nombre limité de mouvements temporels, principalement basés sur les gestes des mains et la détection des moments de contact et de séparation entre les paumes. Cette activité est relativement plus facile à identifier, car elle peut être simplifiée en un ensemble de mouvements spécifiques. En revanche, l'activité "danser" est généralement bien plus complexe, car elle implique une séquence de mouvements plus longue et variée, englobant des gestes coordonnés du torse, des bras et des jambes. Ainsi, la reconnaissance de cette activité demande une analyse approfondie des mouvements dans leur contexte temporel. Un bon modèle de reconnaissance d'activités doit être en mesure de modéliser et de capturer les formes et les relations temporelles à court et à long terme associés aux mouvements de l'individu, afin de différencier de manière précise et fiable les différentes activités.

En plus des défis mentionnés ci-haut, il existe également un problème lié à la nature des annotations. Ces dernières peuvent être au niveau de la trame plutôt que au niveau des individus. En effet, dans de nombreux ensembles de données, les annotations fournies ne représentent qu'une seule activité parmi plusieurs présentes dans la scène, ce qui rend difficile la distinction entre les individus impliqués dans l'activité étiquetée. Ce problème inhérent aux jeux de données nécessite donc le développement d'une méthode capable d'identifier et de représenter l'action d'intérêt en utilisant des caractéristiques discriminantes appropriées relatives aux individus qui exercent l'activité.

1.3 Objectifs de recherche

Les objectifs de recherche de cette thèse sont centrés sur la résolution des défis liés à la reconnaissance d'activités humaines dans les vidéos. Plus spécifiquement, nos objectifs de recherche sont les suivants :

1. Développer une méthode de reconnaissance d'activités dans des environnements réalistes. Cette méthode nécessite en entrée une séquence vidéo composée d'une série de trames temporellement consécutives représentant un ou plusieurs individus dans un espace public. Elle doit ensuite fournir en sortie une prédiction des activités effectuées par chaque individu présent dans la séquence vidéo. Cette méthode doit pouvoir reconnaître les activités dans la scène observée, garantissant une reconnaissance fiable dans des conditions réelles, tout en s'adaptant aux changements d'éclairage, aux occlusions partielles et à d'autres perturbations visuelles. Ainsi, cette méthode doit permettre

une modélisation spatio-temporelle efficace pour mieux comprendre la dynamique des activités observées.

2. Développer une méthode de reconnaissance d'activités humaines, permettant d'identifier l'acteur principal dans une séquence, impliqué dans l'activité d'intérêt. Cette approche se révèle particulièrement bénéfique lorsque les données disponibles ont des annotations incomplètes, et lorsque l'objectif est d'identifier l'activité la plus intense. Les difficultés abordées dans ce travail sont l'ambiguïté liée à l'identification de la personne principale, les cas d'occlusions partielles, de densité de personnes, et de similarité d'apparence des individus.
3. Développer une méthode de reconnaissance d'activités visant à modéliser la dynamique des activités à court terme et à long terme. Cette méthode doit être conçue de manière à modéliser des séquences de longueurs variables et d'apprendre les informations clés relatives à la dynamique et à l'apparence des activités. Cette approche doit permettre d'analyser davantage d'informations séquentielles de manière efficace et précise. Ce travail traite aussi des défis liés aux occlusions partielles, à la visibilité et à la densité des personnes présentes dans la scène surveillée.

1.4 Contributions

La présente thèse propose plusieurs contributions significatives dans le domaine de la reconnaissance d'activités humaines dans les vidéos. Nous présentons ci-dessous les principales contributions de cette thèse :

Reconnaissance d'activités humaines dans des environnements réalistes : pour cet objectif, nous avons proposé et publié "A Grid-based Representation for Human Action Recognition" [35] (article 1). Dans ce travail, nous présentons une méthode de reconnaissance d'activités qui utilise une représentation en grille pour encoder les postures clés de chaque individu dans une scène. Cette représentation compacte permet de capturer efficacement toutes les caractéristiques importantes de l'activité exécutée, en se concentrant uniquement sur les zones et les informations les plus pertinentes. En utilisant cette méthode, nous avons obtenu des résultats prometteurs en termes d'efficacité et de précision dans la reconnaissance d'activités humaines dans des environnements réalistes. Cette contribution est particulièrement importante car elle permet de traiter des scènes complexes, où les activités des individus peuvent être affectées par des facteurs externes tels que l'éclairage ou les mouvements de caméra.

Reconnaissance d'activités humaines avec identification de l'acteur principal :

pour cet objectif, nous avons proposé et publié "Actor-driven Pose Embeddings for Video Action Recognition" (article 2). Dans ce travail, nous présentons une nouvelle méthode de reconnaissance d'activités humaines, qui permet de déterminer l'acteur principal en utilisant le flot optique. Cela nous permet d'identifier l'individu impliqué dans l'activité la plus intense de la scène observée. De plus, notre méthode est également adaptée pour le spectre infrarouge, ce qui la rend particulièrement pertinente dans des situations réelles où les conditions de luminosité sont faibles, comme en présence de brouillard, de fumée ou dans des environnements sombres. Cette approche complète améliore la précision et la fiabilité de la reconnaissance d'activités humaines dans des environnements complexes en spectre infrarouge.

Reconnaissance et modélisation temporelle des activités humaines à court et à long terme : pour cet objectif, nous avons proposé et publié "Transformer-based Human Action Recognition with Dynamic Feature Selection" (article 3). Dans ce travail, nous proposons une nouvelle méthode de reconnaissance d'activités humaines basée sur les réseaux auto-attentionnels. Notre approche intègre une sélection dynamique des caractéristiques basées sur la pose humaine, permettant ainsi d'identifier les informations les plus discriminantes pour chaque activité spécifique. En utilisant un réseau auto-attentionnel, nous exploitons efficacement les relations temporelles dans les séquences vidéo, ce qui améliore la précision et la robustesse de la reconnaissance d'activités. De plus, notre méthode permet une adaptation automatique aux variations de longueur des séquences vidéo, offrant ainsi une flexibilité accrue dans des scénarios où les séquences et les activités peuvent avoir une durée variable. Notre approche présente plusieurs avantages significatifs par rapport aux méthodes conventionnelles, ouvrant ainsi de nouvelles perspectives dans le domaine de la reconnaissance d'activités humaines dans les vidéos.

1.5 Plan de la thèse

Le plan de la présente thèse est structuré de la manière suivante. Le chapitre 2 sera consacré à une revue de littérature des travaux clés dans le domaine de la reconnaissance d'activités humaines dans les vidéos. Dans le chapitre 3, nous présenterons un aperçu des méthodes proposées constituant le corps de la thèse. Le chapitre 4 présentera une nouvelle méthode de reconnaissance d'activités dans les vidéos (article 1). Ensuite, dans le chapitre 5, nous présenterons une méthode de reconnaissance d'activités et d'identification de l'acteur principal adaptée au spectre infrarouge (article 2). Dans le chapitre 6, nous présenterons une méthode de reconnaissance d'activités à court et à long terme dans les vidéos, basée sur des réseaux auto-attentionnels (article 3). Le chapitre 7 sera dédié à une discussion approfondie des travaux réalisés et des améliorations potentielles. Enfin, nous concluons la thèse dans le chapitre 8, en

résumant les principaux résultats obtenus et en proposant des pistes de recherches futures.

CHAPITRE 2 REVUE DE LITTÉRATURE

Dans ce chapitre, nous allons faire un aperçu sur les méthodes les plus pertinentes pour la reconnaissance d'activités humaines dans les vidéos. Pour avoir une vision assez large sur l'état de l'art lié à cette thèse, nous avons couvert à la fois les méthodes classiques et les méthodes basées sur l'apprentissage profond. Les méthodes classiques sont basées sur le couplage entre les modèles traditionnels d'apprentissage machine et les représentations classiques des activités dont les paramètres sont choisis manuellement sous la supervision de l'humain, alors que les méthodes d'apprentissage profond sont basées sur les réseaux de neurones artificiels dont les paramètres sont appris automatiquement. Dans ce qui suit, nous allons discuter les approches classiques dans la section 2.1 et les approches d'apprentissage profond dans la section 2.2. Nous discuterons également des différences entre les méthodes de reconnaissance d'activité basées sur l'apprentissage profond face aux méthodes classiques dans la section 2.3. Quant à la section 2.4, elle sera consacrée à la description des jeux de données intéressants pour la reconnaissance d'activités humaines dans les vidéos.

2.1 Méthodes classiques

Dans la littérature, plusieurs travaux ont proposé des solutions d'analyse vidéo qui étudient les comportements humains. Selon la représentation donnée aux activités, on distingue les approches globales et les approches locales. Comme la plupart de ces approches sont basées sur la détection de l'humain, nous commencerons par discuter les méthodes de détection dans la partie 2.1.1. Ensuite nous discuterons les approches globales dans la partie 2.1.2. Quant à la partie 2.1.3, elle sera consacrée aux approches locales.

2.1.1 La détection des humains

La détection de l'humain est un processus essentiel pour analyser leurs comportements. L'objectif derrière cette étape est d'extraire les régions correspondantes à des personnes du reste d'une image. La détection est généralement faite par segmentation ou par classification.

Détection par segmentation

La segmentation du mouvement consiste à détecter les régions mobiles qui correspondent aux éléments en mouvement. Pour y arriver, on trouve plusieurs familles de méthodes telles que

les méthodes de soustraction d'arrière-plan [36, 37] ou les méthodes de flot optique [38, 39].

La soustraction d'arrière-plan : pour distinguer les éléments d'intérêt (avant-plan) des parties statiques de la scène (arrière-plan), la technique de soustraction d'arrière-plan construit un modèle d'arrière-plan et compare ce modèle à la trame actuelle afin de détecter les zones où une différence significative se produit. La soustraction d'arrière-plan implique deux processus distincts, à savoir la modélisation d'arrière-plan et la détection d'avant-plan. Dans la modélisation d'arrière-plan, on construit le modèle d'arrière-plan qu'on met à jour périodiquement par exemple pour prendre en compte les modifications de l'éclairage. Alors que dans la détection d'avant-plan, on décide si une nouvelle intensité correspond au modèle d'arrière-plan.

Les méthodes de soustraction d'arrière-plan peuvent être séparées en deux familles : méthodes paramétriques et méthodes non paramétriques. Dans les méthodes paramétriques comme [40], pour construire le modèle d'arrière-plan, on modélise chaque pixel d'arrière-plan par un mélange de distributions Gaussiennes. Ensuite en regardant si les observations font partie ou non au mélange de distributions Gaussiennes de l'arrière-plan, on considère un pixel comme appartenant à l'arrière-plan ou à l'avant-plan. Ces méthodes ont tendance à être robustes face au bruit des capteurs.

Pour les méthodes non paramétriques on modélise les probabilités d'appartenance des pixels à l'arrière-plan par des histogrammes. La méthode la plus populaire dans cette famille est la méthode ViBe [3] : pour chaque pixel, cette méthode stocke un ensemble de valeurs prises antérieurement au même endroit ou dans le voisinage. Ensuite on compare cet ensemble de valeurs à la valeur du pixel actuel afin de déterminer si ce pixel appartient à l'arrière-plan. Pour adapter le modèle, contrairement aux autres méthodes de cette famille, dans ViBe on ne remplace pas les valeurs les plus anciennes mais on choisit plutôt de manière aléatoire les valeurs à enlever du modèle d'arrière-plan. Finalement, lorsqu'un pixel fait partie de l'arrière-plan, sa valeur est propagée dans le modèle d'arrière-plan d'un pixel voisin. Ici, la figure 2.1 montre un exemple de résultat obtenu par la méthode ViBe.

En résumé, les méthodes de soustraction d'arrière-plan s'avèrent utiles pour décrire les contours des régions d'intérêt. Cependant, elles sont sensibles aux scènes dynamiques (objets d'arrière-plan en mouvement et variations d'illuminations) et nécessitent que la caméra observant la scène soit stationnaire. De plus, ces méthodes nécessitent souvent des opérations morphologiques pour éliminer les bruits de segmentation et améliorer la forme des régions segmentées.

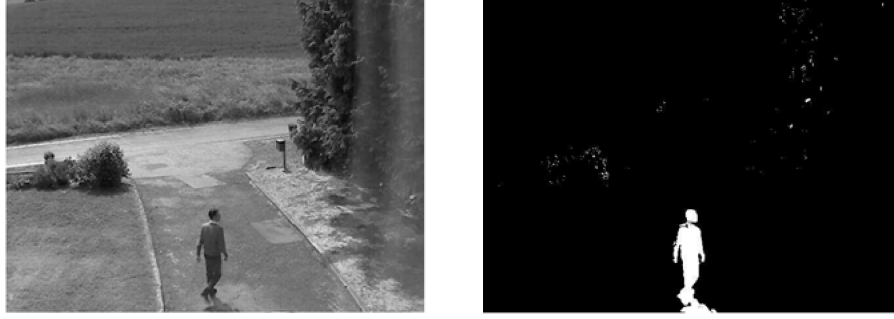


FIGURE 2.1 Exemple de soustraction d'arrière-plan avec la méthode ViBe [3].

Le flot optique : appelé aussi vecteurs de flux optique [38], il permet d'estimer le mouvement à partir de trames consécutives. Il repose sur le calcul d'un champ vectoriel qui exprime le déplacement des pixels d'une image. Le but consiste à trouver les valeurs de $V_x = \frac{\partial x}{\partial t}$ (flot optique en X) et $V_y = \frac{\partial y}{\partial t}$ (flot optique en Y) de l'équation suivante :

$$I_x V_x + I_y V_y = -I_t \quad (2.1)$$

où I_x , I_y et I_t sont les dérivés de l'intensité en X , en Y , et dans le temps qui sont calculées à partir des trames en entrée. Pour résoudre cette équation à deux variables, plusieurs méthodes ont été proposées. Une des plus populaires est basée sur l'algorithme de Gunnar Farneback [41] qui permet de calculer le flot optique dense pour tous les pixels de la trame.

Le flot optique permet de donner des informations importantes sur la disposition spatiale des objets en mouvement. À travers les discontinuités dans le flot optique, on peut segmenter les régions d'intérêt dans la scène. Ainsi en fonction des vitesses individuelles des éléments présents dans la scène on peut détecter les êtres humains ou tout autre élément d'intérêt [42]. La figure 2.2 montre un exemple d'utilisation du flot optique pour la détection d'humain.

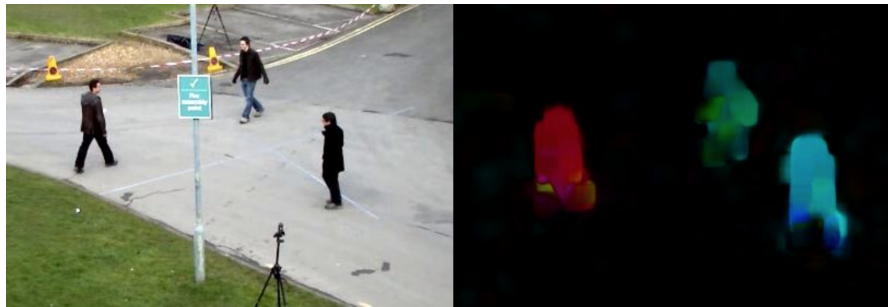


FIGURE 2.2 Détection d'humain avec flot optique [4].

Bien que le calcul du flot optique soit coûteux (notamment lorsqu'on calcule du flot optique dense), il reste à ce jour nécessaire, voire inévitable, pour la plupart des applications en analyse vidéo.

Détection par classification

La détection par classification consiste à distinguer les éléments d'intérêt dans la scène. En revanche, ces éléments doivent appartenir à des catégories spécifiques et connues à l'avance. Comparées aux autres méthodes de détection d'éléments d'intérêt, les méthodes basées sur la classification permettent de donner également la nature de l'objet et non pas seulement sa position dans la scène. Dans la littérature, plusieurs méthodes de classification ont été proposées pour la détection des êtres humains, la méthode la plus populaire est basée sur les descripteurs HOG (Histogram of Oriented Gradients) [43]. Pour identifier l'humain, le détecteur HOG commence par diviser la région d'intérêt en blocs. Ensuite pour obtenir les descripteurs HOG, on calcule l'histogramme d'orientations du gradients pour chaque bloc. Une fois que nous avons les descripteurs qui représentent les régions de l'humain, on utilise un classificateur SVM pour apprendre les représentations possibles de l'humain. Finalement afin de détecter les humains présents dans la scène, on fait glisser des rectangles du haut à gauche vers le bas à droite dans l'image entière pour différentes échelles spatiales.

Malgré l'efficacité de ce genre de méthodes, elles restent très coûteuses du moment qu'elles parcourent toute l'image à différentes échelles. Pour pallier ce problème, d'autres méthodes basées sur la proposition d'objets ont été proposées. L'une des méthodes les plus populaires dans ce sens est la méthode Edge Boxes [44]. Cette méthode restreint la recherche dans l'image en se basant sur les objets dont les contours sont fermés. Aujourd'hui, les méthodes de détection d'objets les plus efficaces comme Faster R-CNN [45] ou YOLO [46] sont basées principalement sur les réseaux de neurones artificiels profonds que nous discuterons dans la section 2.2.

En résumé, la détection des humains à partir des vidéos est une tâche difficile en raison de leur apparence qui varie selon la pose, la distance à la caméra, l'éclairage ou les mouvements de cette dernière. De plus, comme les méthodes de détection des humains ne sont souvent qu'une première étape pour la reconnaissance d'activités, elles doivent être rapides et précises pour qu'on puisse opérer convenablement les étapes suivantes.

2.1.2 Approches globales

Les approches globales consistent à décrire la trame dans son ensemble, elles représentent les régions d'intérêt par leurs formes. Les méthodes basées sur la dynamique de la silhouette de l'humain sont les toutes premières méthodes qui ont été utilisées pour représenter les activités humaines dans les vidéos. Pour capturer les informations de mouvement de l'humain dans le temps, ces méthodes consistent à empiler les informations spatiales de la pose humaine à des moments consécutifs. L'une des méthodes les plus classiques dans cette catégorie est la méthode par énergie et historique de mouvement [5]. Pour représenter une activité, cette méthode vise à encoder une activité humaine (vidéo) sous forme de deux images seulement. Elle commence par extraire la silhouette via la soustraction d'arrière-plan et ensuite accumule les différences des silhouettes entre les trames consécutives. Ces différences sont ensuite utilisées pour construire deux images. Une image d'énergie de mouvement binaire (MEI) qui représente la position du mouvement et une image d'historique de mouvement (MHI) qui représente la manière par laquelle le mouvement a été exercé. L'intensité de chaque pixel dans les images MHI est en fonction de l'historique temporel du mouvement à l'emplacement spatial correspondant dans la séquence des trames (les pixels les plus récents sont les plus lumineux, voir la figure 2.3). Une fois que les deux images MEI et MHI sont extraites, elles sont utilisées pour donner une représentation discriminative des activités. Pour chaque catégorie d'action, on extrait toutes les images MEI et MHI de l'ensemble d'entraînement. Ensuite, on calcule les descriptions statistiques de ces images en utilisant les moments de Hu afin de leur donner des représentations plus discriminatives. Finalement pour reconnaître l'activité humaine à travers une séquence vidéo de test, on calcule la distance de Mahalanobis entre les descriptions statistiques de la vidéo de test et chacune des vidéos d'entraînement. Selon cette distance, les descriptions les plus proches sont considérées pour déterminer la classe de l'action.

Bien que cette méthode permettait d'obtenir de bons résultats à l'époque, son applicabilité était limitée uniquement aux activités très simples (comme les exercices d'aérobic). En plus, les représentations MEI et MHI sont sensibles aux changements de point de vue. Pour pallier au problème de dépendance au point de vue, les auteurs de [47] ont utilisé des représentations 3D, où chaque action est représentée par des volumes d'historique de mouvement (MHV) dans des repères cylindriques. Mais en conséquence, ces dernières nécessitent l'utilisation de plusieurs caméras. Sans utiliser plusieurs caméras, grâce aux formes spatio-temporelles, les auteurs de [6] ont proposé un modèle basé sur les formes induites par les silhouettes dans le volume spatio-temporel. Comme pour les autres méthodes, on extrait les silhouettes par soustraction d'arrière-plan, ensuite on empile ces silhouettes sur une séquence donnée pour

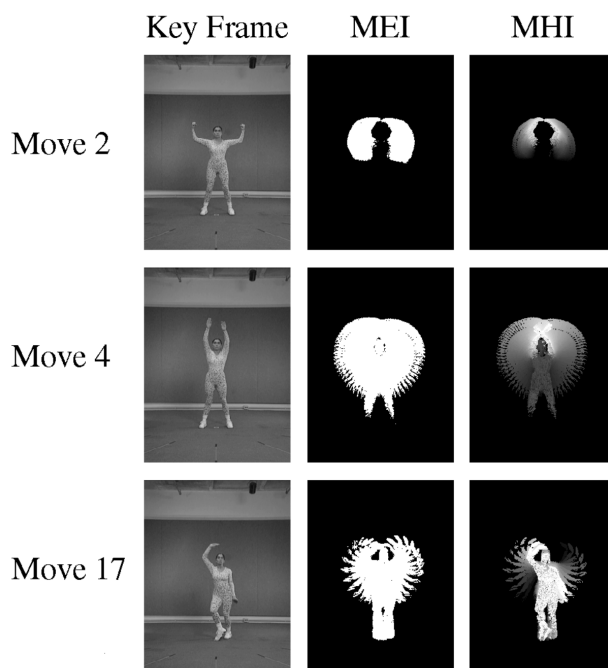


FIGURE 2.3 Images MEI et MHI pour représenter trois différents types de mouvements [5].

former un volume spatio-temporel (voir figure 2.4). Après, on utilise les propriétés de la solution de l'équation de Poisson pour extraire des caractéristiques spatio-temporelles qui représentent l'action exercée. Une fois que nous avons les représentations pour chaque vidéo, on reconnaît une activité humaine en utilisant l'algorithme du plus proche voisin avec la distance euclidienne.



FIGURE 2.4 Représentations spatio-temporelles pour trois différents types d'activités [6].

Pour conclure, malgré l'utilité des méthodes globales pour la reconnaissance d'activités humaines à l'époque, leurs précision est étroitement liée à la qualité de la segmentation puis-

qu'elles reposent principalement sur la silhouette de l'humain. En plus, leur applicabilité est limitée aux environnements contrôlés.

2.1.3 Approches locales

Face aux limites des représentations globales pour la reconnaissance d'activités humaines, des chercheurs ont proposé des méthodes basées sur les représentations locales. Ces dernières ont montré beaucoup de succès lors des deux dernières décennies. L'avantage principal de ces méthodes est qu'elles identifient uniquement les régions locales contenant des informations de mouvement saillantes (où les informations de mouvement sont pertinentes). Ainsi, elles ne nécessitent aucune information sur le modèle du corps humain ou sur sa localisation. En fait, au lieu d'extraire explicitement des caractéristiques sur la silhouette humaine, on utilise un détecteur de caractéristiques locales qui permet d'identifier les régions les plus saillantes dans l'image. Ensuite on décrit les voisinages spatio-temporels autour des caractéristiques détectées à l'aide d'un descripteur de caractéristiques locales. Parfois les détecteurs de caractéristiques locales peuvent également servir comme descripteurs, comme c'est le cas des points SIFT [48].

Dans la littérature, les approches locales sont souvent basées sur les points spatio-temporels et les trajectoires de mouvement. La méthode proposée par Laptev et al. [23] est l'une des approches les plus populaires basées sur les points spatio-temporels. Cette méthode consiste à donner une représentation sous forme d'un vecteur caractéristique pour chaque vidéo d'action. Pour ce faire, on commence par détecter les points d'intérêt spatio-temporels à l'aide de détecteurs Harris3D [49]. Ensuite on extrait des caractéristiques autour de chaque point d'intérêt détecté. Une fois que nous avons obtenu toutes les caractéristiques pour chaque vidéo, on utilise l'approche de sac de mots visuels. Dans cette approche, on commence par quantifier les caractéristiques obtenues de chaque vidéo par un regroupement avec k-moyennes. Ensuite, on construit par regroupement un vecteur d'occurrences des centres de groupes (mots visuels). Enfin, chaque vidéo est représentée par un vecteur de la fréquence de mots visuels, permettant ainsi la reconnaissance d'activités à l'aide d'un classificateur SVM. La figure 2.5 illustre le processus typique pour reconnaître les activités humaines par l'approche de sac de mots visuels.

Dans le même contexte, au lieu d'utiliser des détecteurs de points d'intérêt, d'autres méthodes ont proposé de faire de l'échantillonnage dense [50]. Le but de ces méthodes est de considérer un grand nombre de points pour obtenir une description plus riche des activités. Cet échantillonnage consiste à extraire les points d'intérêt à des positions et à des échelles régulières dans l'espace et dans le temps.

Comparées aux approches globales, les approches basées uniquement sur les points d'inté-

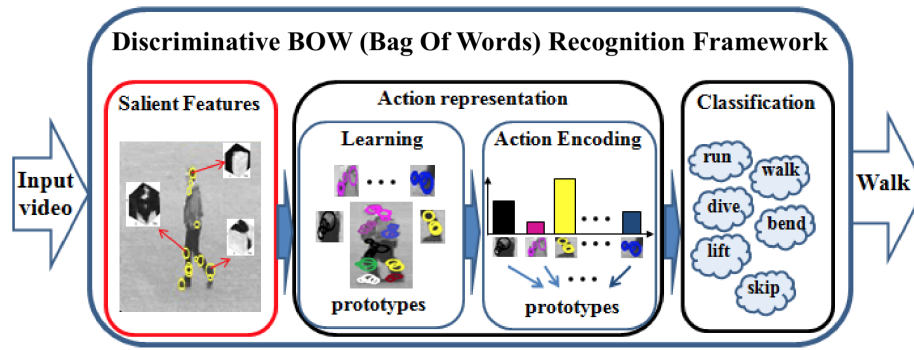


FIGURE 2.5 Processus de reconnaissance des activités humaines par l’approche de sac de mots visuels [7].

rêt spatio-temporels ont montré des succès remarquables, notamment face au problème de variation de la pose. Cependant ces approches permettent de capturer uniquement les informations sur de courtes durées temporelles, typiquement pour des séquences de 10 trames. Ainsi, elles ne permettent pas de considérer les informations sur de longs intervalles temporels puisqu’elles ne prennent pas en compte les caractéristiques temporelles des activités humaines.

Pour capturer des informations de longues durées on a proposé des méthodes basées sur les trajectoires de caractéristiques [51–53]. Ces trajectoires sont généralement extraites en détectant des points d’intérêt et en les suivant d’une trame à l’autre. L’une des méthodes de suivi des points d’intérêt les plus utilisées pour extraire les trajectoires des activités humaines est la méthode KLT (Kanade-Lucas-Tomasi) [54]. KLT commence par localiser les bonnes caractéristiques de suivi, ensuite suit ces caractéristiques à l’aide d’une recherche locale basée sur les gradients de l’image.

Toujours en se basant sur les trajectoires de caractéristiques et spécialement pour surpasser les performances de KLT, l’une des méthodes les plus robustes à ce jour a été proposée par les auteurs de [8]. Cette méthode consiste à générer un grand nombre de trajectoires (trajectoires denses) afin de donner une représentation plus riche aux activités. Pour ce faire, on commence par un échantillonnage dense à multiples échelles spatiales pour extraire les points d’intérêt. Ensuite, pour construire les trajectoires denses, on suit ces points à l’aide du flot optique dense afin de considérer uniquement les points appartenant à des régions en mouvement. Une fois que nous avons extrait les trajectoires, on les décrit par des descripteurs de forme de trajectoire, HOG, HOF et également MBH pour limiter les effets des mouvements de la caméra (voir la figure 2.6). Finalement, on utilise l’approche de sac de mots visuels et on reconnaît les activités avec SVM comme classificateur. Comparée à toutes les autres approches

locales, la combinaison des descripteurs HOG, HOF et MBH pour les trajectoires denses [8,55] a permis d'améliorer les performances des méthodes de reconnaissance des activités dans les vidéos. Aujourd'hui la méthode des trajectoires denses améliorées [56] reste la méthode la plus performante de la littérature. Pour améliorer les méthodes proposées auparavant, cette méthode utilise des détecteurs d'humains pour estimer l'homographie afin de réduire encore plus les effets de mouvement de la caméra et ainsi éliminer le maximum de trajectoires qui appartiennent à l'arrière-plan.

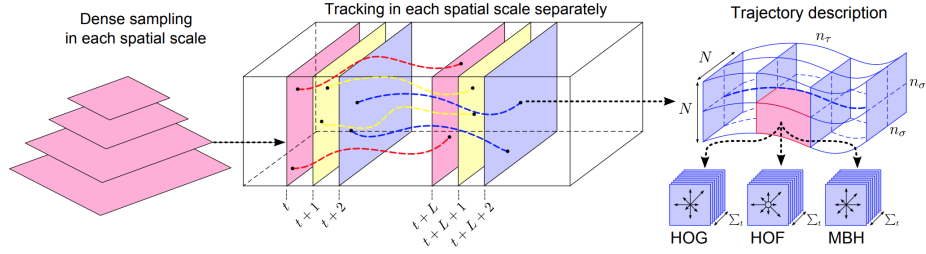


FIGURE 2.6 Descripteurs spatio-temporels par trajectoires denses [8].

2.1.4 Discussion

Contrairement aux approches globales qui permettent de reconnaître uniquement les activités très basiques dans des environnement contrôlés, les approches locales couvrent des activités plus réalistes. Ces approches sont fiables face au bruit d'arrière-plan, aux changements d'illumination, et permettent de représenter plusieurs catégories d'activités.

Aujourd'hui, bien que les méthodes d'apprentissage profond soient dominantes dans la littérature, les méthodes classiques de reconnaissance des activités humaines restent encore utiles. Elles sont faciles à entraîner et généralisent assez bien sur les petits ensembles de données. Quelques méthodes classiques donnent même de bons résultats sur les ensembles de données les plus larges, comme c'est le cas des méthodes basées sur les trajectoires denses améliorées [56] dont les performances restent comparables aux méthodes les plus récentes.

2.2 Méthodes basées sur l'apprentissage profond

Comme la section précédente était consacrée aux approches classiques proposées pour faire la reconnaissance d'activités humaines dans une vidéo, nous allons discuter dans cette section les méthodes basées sur l'apprentissage profond. Nous allons commencer par introduire l'apprentissage profond qui représente la base des approches récentes dans la partie 2.2.1. Ensuite, dépendamment de l'architecture utilisée pour reconnaître les activités humaines dans

les vidéos, nous allons séparer les approches d'apprentissage profond selon cinq types d'architectures : architectures spatiales dans la partie 2.2.2, architectures spatio-temporelles dans la partie 2.2.3, architectures à deux flux spatiale et temporel dans la partie 2.2.4, architectures spatiales et temporelles à long terme dans la partie 2.2.5 et architectures auto-attentives dans la partie 2.2.6.

2.2.1 Introduction

Récemment, grâce à la disponibilité croissante de données numériques et aux nouvelles capacités de calcul disponibles (principalement les architectures GPU), les recherches et les applications sont de plus en plus dirigées vers l'exploitation des réseaux de neurones artificiels pour des tâches d'intelligence artificielle. Comme plusieurs autres branches de l'intelligence artificielle, aujourd'hui la vision par ordinateur profite pleinement des avancées des recherches liées aux réseaux de neurones artificiels profonds [57] [58]. Actuellement ces avancées sont d'un grand intérêt pour différentes tâches telles que la reconnaissance d'activité humaine dans les vidéos.

Alors que les approches d'apprentissage automatique les moins récentes (évoquées dans la section 2.1) adoptent des caractéristiques classiques et choisies manuellement (*hand crafted*) couplées avec des méthodes d'apprentissage machine classiques [59, 60], les approches basées sur l'apprentissage profond utilisent actuellement des caractéristiques apprises et permettent souvent l'apprentissage de bout en bout. Contrairement aux approches classiques, l'apprentissage profond donne souvent une représentation de haut niveau directement à partir des données (vidéos). L'idée sous-jacente est que plus l'extraction des caractéristiques donne de bonnes représentations, plus les résultats de classification augmentent.

2.2.2 Architectures spatiales

Les réseaux de neurones convolutifs sont inspirés principalement par le mécanisme de perception visuelle des êtres humains. Les premiers réseaux de neurones convolutifs (CNN) [9, 61] ont obtenu les meilleures performances et constituent l'état de l'art sur plusieurs ensembles de données de reconnaissance de patrons pour des tâches de classification des images et de détection d'objets [62]. Comme pour la plupart des réseaux de neurones artificiels, les CNN contiennent plusieurs couches et sont entraînés par l'algorithme de rétropropagation. La figure 2.7 montre l'architecture originale du réseau LeNet-5 [9]. Ce réseau est composé de 2 couches convolutives, 2 couches de sous-échantillonnage, 2 couches pleinement connectées et une couche de sortie Softmax. Le but des couches convolutives est de repérer les caractéristiques présentes dans l'image d'entrée. Pour ce faire, elles effectuent des opérations de convolution entre

l'image et les filtres de convolutions dont les paramètres constituent les poids du modèle. Une fois ces opérations effectuées, chaque couche de convolution donne en sortie des cartes de caractéristiques. Ensuite vient le rôle des couches de sous-échantillonnage qui permettent d'alléger les calculs en réduisant la taille des caractéristiques extraites. Finalement les cartes de caractéristiques résultantes sont prises comme entrée de la couche suivante. D'une couche à l'autre, on répète les mêmes processus de convolution et de sous-échantillonnage jusqu'à ce qu'on obtienne les caractéristiques de la dernière couche (caractéristiques profondes). Comme dernière étape, les couches pleinement connectées prennent les caractéristiques de la dernière couche pour produire des prédictions avec la couche Softmax.

Les CNN sont conçus à la base pour extraire des caractéristiques spatiales 2D à partir des images. En 2012 lors de la compétition ImageNet, l'architecture profonde du réseau AlexNet [57] était la première à inclure l'optimisation Dropout pour atténuer le sur-apprentissage et à remplacer les unités Tanh (tangente hyperbolique) par les unités Relus (Rectified Linear Units) pour accélérer l'apprentissage. Ainsi AlexNet contenant 5 couches convolutives suivies par 3 couches pleinement connectées était l'une des architectures les plus performantes à l'époque.

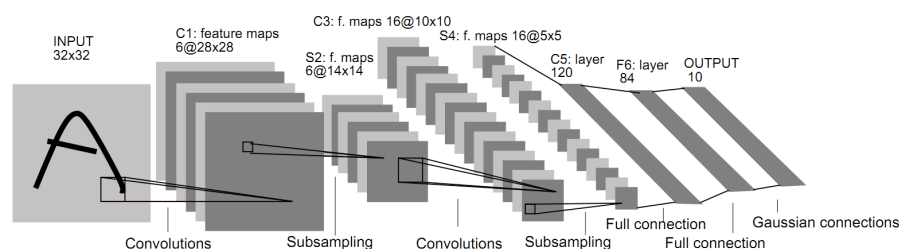


FIGURE 2.7 L'architecture CNN originale du réseau LeNet-5 [9].

Suite au succès de ce dernier réseau, plusieurs autres travaux s'en sont inspirés pour mettre en œuvre des architectures encore plus profondes. VGGNet [63] proposait deux différentes versions de 16 et de 19 couches tout en utilisant des filtres de convolution très petits de 3×3 . GoogleNet [64] proposait une architecture de 22 couches tout en augmentant également la largeur du réseau. Plus récemment, Wang et al. [65] ont proposé High-Resolution Net (HRNet), une architecture qui se distingue par sa capacité à maintenir une résolution élevée à toutes les étapes du processus de traitement, ce qui permet de capturer des détails fins et de mieux représenter les informations spatiales.

Après les résultats prometteurs que permettaient d'avoir ces architectures dans différentes applications, il a été démontré que plus le réseau est profond, plus on génère des représentations de caractéristiques discriminantes et ainsi on obtient une meilleure approximation de

la fonction de perte du réseau. En plus de la profondeur, plusieurs stratégies d'optimisation des réseaux de neurones ont montré beaucoup d'efficacité. La stratégie la plus populaire à ce jour reste la normalisation par lots [66]. Pour garder la même distribution des entrées avant et après l'activation d'une couche lors de l'entraînement des réseaux de neurones artificiels, la normalisation par lots consiste à normaliser les activations des neurones entre toutes les données d'un lot. En résultat, cela permet au réseaux de neurones d'accélérer l'entraînement et d'améliorer les performances tout en assurant un entraînement plus stable. Le succès des caractéristiques CNN pour l'analyse d'images a motivé l'utilisation des caractéristiques profondes pour analyser les vidéos. L'idée est de traiter une vidéo comme une séquence de trames, où pour chaque trame on extrait des caractéristiques discriminantes afin de donner des représentations des activités dans notre cas. À travers ce processus d'extraction des caractéristiques, on cherche à avoir des représentations discriminantes de manière à ce que les exemples de la même classe aient des représentations similaires et en même temps différentes des exemples des autres classes. Pour ce faire, on extrait souvent les caractéristiques via des réseaux CNN profonds pré-entraînés sur des bases de données très grandes comme ImageNet [58].

Après l'extraction des représentations des séquences, la reconnaissance d'activités est faite par des classificateurs tels que les SVMs. La reconnaissance peut être aussi faite par l'apprentissage de bout en bout via l'intégration de couches pleinement connectées et d'une couche de classification (souvent Softmax) dans l'architecture du réseau CNN.

L'un des travaux les plus populaires qui utilise uniquement les informations spatiales des CNN 2D pour reconnaître les activités humaines est [10]. Dans ce travail, plusieurs techniques qui étendent les CNN 2D pour analyser les vidéos des activités sont exploitées. Pour donner une représentation au niveau de la vidéo et faire la classification par la suite, on extrait avec l'architecture AlexNet les caractéristiques des trames qu'on fusionne dans le domaine temporel. Pour ce faire, on a essayé plusieurs techniques de fusion telles que la fusion au début, la fusion à la fin (fusion classique) et la fusion lente. La figure 2.8 illustre ces différentes techniques comparées à l'utilisation d'une seule trame.

La fusion à la fin est faite en utilisant deux réseaux différents qui prennent chacun une seule trame en entrée, ensuite les caractéristiques pour chaque trame sont fusionnées dans la sortie des deux réseaux. La fusion au début est faite en modifiant la première couche convolutive de manière à ce que ses filtres de convolution tiennent en considération la dimension temporelle (considération de 10 trames). Quant à la fusion lente, son principe est similaire à peu près au principe de la fusion au début sauf pour le fait qu'on modifie toutes les couches convolutives et non seulement la première. Considérant 10 trames en entrée, les filtres de la première couche

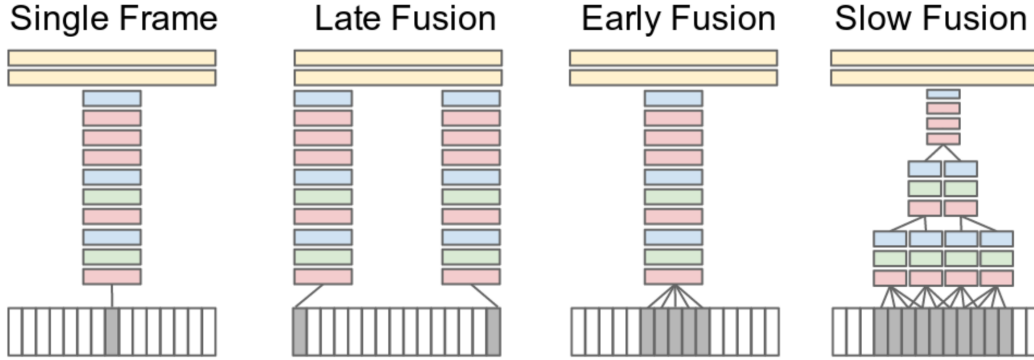


FIGURE 2.8 Les différentes approches de fusion pour donner une représentation au niveau de la vidéo [10].

sont étendus pour fusionner 4 trames tandis que les filtres de la seconde et la troisième couche sont étendus pour fusionner 2 trames. Suite aux différents tests effectués sur les techniques de fusion, les auteurs ont qualifié la fusion lente comme étant la meilleure technique.

Dans ce travail, on a proposé aussi une architecture qui permet d'accélérer l'entraînement des réseaux CNN pour la tâche d'analyse des vidéos, toutefois en gardant les mêmes performances et sans nécessité de réduire la taille du réseau. L'architecture proposée consiste à prendre deux entrées de très faible résolution (89×89) au lieu de prendre une entrée de résolution originale (178×178). On a utilisé une architecture à deux flux où le premier flux (flux de contexte) prend la trame entière à moitié de la résolution originale et le deuxième flux (flux fovea) prend la région centrale de la trame à la résolution originale. Une fois les entrées passées par les deux flux (AlexNet), les sorties de ces deux derniers sont concaténées pour faire la prédiction. La figure 2.9 illustre l'architecture proposée.

Bien que les architectures basées uniquement sur les CNN 2D présentent des avantages, la manière par laquelle on intègre les informations dans le temps reste limitée face à la nature dynamique des activités humaines. Dans ce qui suit nous discuterons des approches qui considèrent l'aspect temporel d'une manière plus adéquate.

2.2.3 Architectures spatio-temporelles

Même si les CNN 2D tirent avantage des modèles pré-entraînés sur des ensemble de données larges, ils ne permettent pas de modéliser efficacement les informations temporelles. En conséquence, ils ne permettent pas d'apprendre des caractéristiques discriminatives dans le temps. Comme on s'intéresse à l'analyse des vidéos qui sont des signaux spatio-temporels 3D, l'idée principale derrière les travaux de littérature que nous discuterons dans cette section est

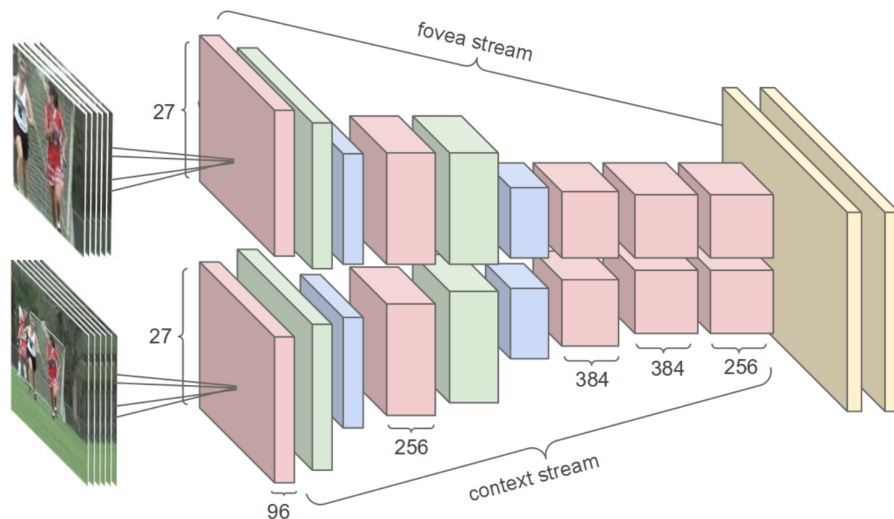


FIGURE 2.9 L'architecture du réseau CNN à deux flux basés uniquement sur les informations spatiales [10].

d'étendre les CNN conçus à la base pour extraire les caractéristiques spatiales 2D à partir des images, afin qu'ils explorent les informations temporelles contenues dans les vidéos.

Pour intégrer l'évolution temporelle, les auteurs de [67] ont proposé un modèle CNN 3D qui considère des convolutions en 3D sur des trames empilées afin d'opérer dans le domaine spatio-temporel. Les CNN 3D utilisent des filtres de convolution en 3D étendus au domaine temporel pour apprendre les informations de mouvement entre des trames consécutives (voir figure 2.10). Les filtres utilisés dans les CNN 3D sont similaires aux filtres utilisés pour faire la fusion dans les architectures spatiales (la fusion au début et la fusion lente). Cependant au lieu d'utiliser ces filtres uniquement dans quelques couches et juste entre certaines trames quelconques, dans les réseaux CNN 3D les filtres sont utilisés de manière à ce qu'il considèrent à chaque couche un ensemble continu de trames consécutives. Dans l'approche originale [67], on a utilisé des filtres qui considèrent 7 trames de chaque séquence vidéo.

L'une des architectures de CNN 3D la plus populaire est l'architecture C3D [12]. Cette architecture consiste à utiliser uniquement des filtres de convolution de $3 \times 3 \times 3$ qui considèrent 3 trames de la séquence vidéo. Comme illustré dans la figure 2.11, cette architecture est composée de 8 couches convolutives et 2 couches pleinement connectées.

Afin d'intégrer les poids appris par les modèles de CNN 2D spatiaux pré-entraînés sur les ensembles de données larges comme ImageNet, les auteurs de [68] ont proposé l'initialisation des poids de convolution 3D par des poids de convolution 2D. Grâce à cette technique, on accélère l'entraînement et on minimise le sur-apprentissage des CNN 3D sur les petits

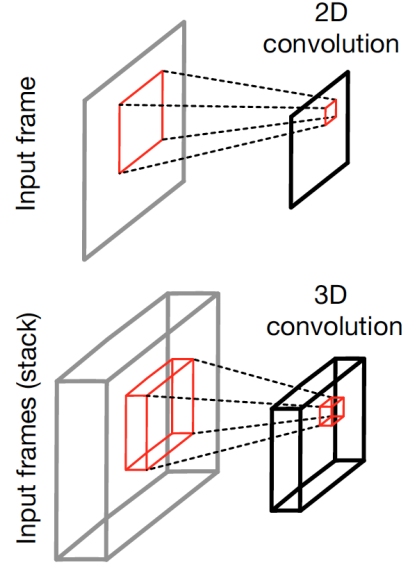


FIGURE 2.10 Opération de convolution en 3D comparée à la convolution 2D [11].

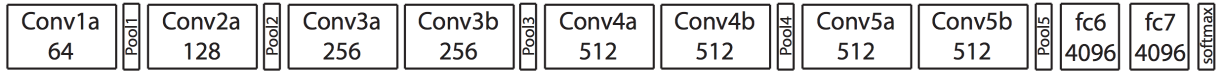


FIGURE 2.11 L'architecture du réseau CNN 3D C3D [12].

ensembles de données.

Quoique les CNN 3D représentent une alternative qui considère l'évolution temporelle dans les vidéos, leur entraînement est très coûteux car ils contiennent beaucoup de paramètres surtout lorsqu'ils sont comparés aux CNN 2D. L'une des techniques pour réduire le nombre des paramètres des C3D est F_{STCN} [69] qui consiste à utiliser des réseaux de convolution spatio-temporels factorisés. Ces réseaux factorisent les filtres de convolution 3D à une séquence de filtres spatiaux 2D dans les couches inférieures suivis par des filtres temporels 1D (dimension temporelle) dans les couches supérieures. D'autres techniques comme MiCT [70] consistent tout simplement à utiliser des architectures composées à la fois de convolutions en 3D et de convolutions en 2D.

En résumé, les caractéristiques spatio-temporelles sont très complexes à capturer uniquement par des modèles CNN sur des images RGB. Dans la littérature, on a aussi montré que les performances des modèles CNN 3D sont comparables à celles des CNN 2D [12]. En plus les CNN 3D nécessitent de très larges ensembles de données d'entraînement et ne permettent pas de modéliser efficacement les longs intervalles temporels. Dans la partie suivante nous discuterons d'une alternative qui considère l'aspect temporel d'une manière implicite sans

alourdir l'entraînement.

2.2.4 Architectures à deux flux spatial et temporel

Motivés par le fait que les vidéos sont naturellement décomposables en informations spatiales et temporelles, les auteurs de [13] ont proposé l'architecture à deux flux d'entrées différents qui intègre l'aspect temporel des activités humaines présentes dans les vidéos. Cette architecture permet d'apprendre séparément les caractéristiques spatiales et les caractéristiques temporelles entre les trames. L'approche typique consiste à utiliser deux CNN : entraîner un CNN spatial avec les informations d'apparences spatiales des trames (RGB) et entraîner un autre CNN temporel avec les informations de mouvements entre les trames adjacentes (voir la figure 2.12). Pour entraîner le CNN temporel, on utilise le flot optique dense ou les vecteurs de mouvement (pour accélérer les calculs) [71], ces vecteurs sont obtenus généralement à l'aide de la compression de vidéos. Cette dernière est utilisée pour réduire la grande taille que peuvent occuper les vidéos, notamment lorsqu'elles ont une haute résolution spatiale et temporelle. Pour ce faire, la compression de vidéos exploite la similarité des trames consécutives afin de retenir uniquement quelques trames complètes et ainsi reconstruire les autres trames en se basant sur les différences (vecteurs de mouvements) entre les trames.

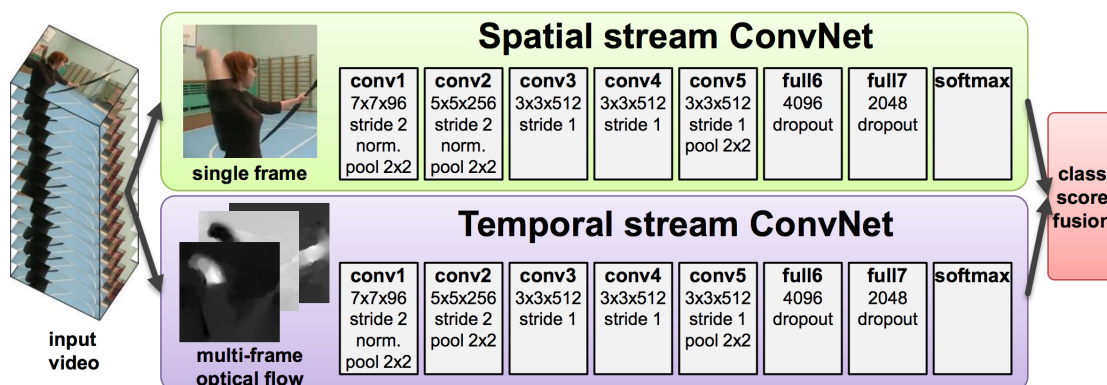


FIGURE 2.12 L'architecture CNN à deux flux spatial et temporel [13].

Dans l'approche originale [13], le flux spatial sert précisément à capturer les informations d'apparence relatives à la scène et aux objets présents. Il permet de tirer ces informations (statiques) à partir d'une seule trame qu'on choisit aléatoirement dans la vidéo. L'architecture du flux spatial (et temporel) est la même que les architectures appliquées pour classer les images. Dans l'approche que nous discutons on a utilisé l'architecture CNN-M [72]. Cette architecture est composée de 5 couches convolutives et de 2 couches pleinement connectées.

Quant au flux temporel, il sert à capturer les informations de mouvement liées aux éléments dynamiques de la scène. Pour considérer ces informations, le flux temporel prend en entrée un volume des trames empilées de flot optique dense entre les trames autour (avant et après) de la trame choisie de la vidéo. Afin de fusionner les deux flux, on fait une fusion à la fin dans laquelle on fusionne les scores de prédiction de chaque flux à l'aide d'un classificateur SVM ou en calculant simplement la moyenne des scores.

Encore une fois l'avantage de ce type d'approche est qu'il permet d'utiliser des modèles pré-entraînés sur ImageNet du moment que le CNN spatial est relatif à la description des images. Malgré le succès de l'approche originale, elle présentait tout de même des limites, principalement liées à la considération d'une seule trame temporelle et à la manière par laquelle on fusionne les deux flux.

Afin de surpasser ces limites, plusieurs travaux ont proposé des architectures à deux flux améliorées. Pour agréger les caractéristiques de convolution dans différentes positions spatiales ou temporelles, les auteurs de [73] ont introduit des couches de sous-échantillonnage par le maximum. L'avantage de ces couches est qu'elles permettent de considérer plusieurs trames. Pour ce faire, on extrait les caractéristiques profondes de chaque trame et ensuite grâce à la couche de sous-échantillonnage, on agrège toutes les caractéristiques de la vidéo avant de faire la prédiction. De façon similaire, la méthode TDD (Trajectory-pooled Deep-convolutional Descriptor) [74] a intégré le sous-échantillonnage par la somme. Ce sous-échantillonnage consiste à calculer la somme des descripteurs profonds de convolutions au long des trajectoires denses améliorées [56] obtenues tout au long de la séquence. L'avantage de cette approche est qu'elle permet de tirer avantage à la fois des caractéristiques classiques (trajectoires) et des caractéristiques profondes (cartes de caractéristiques). Encore une fois, afin de considérer le maximum de trames possibles, la méthode TSN (Temporal Segment Network) [14] a été introduite. Elle consiste à échantillonner aléatoirement les trames par des segments temporels de même longueur et d'obtenir ensuite les scores de prédiction de chaque segment afin de calculer le score de consensus pour chaque type de flux. Finalement, on calcule la prédiction finale avec une moyenne pondérée du score des deux flux. La figure 2.13 montre le schéma du réseau TSN. Comme les flux des architectures à deux flux sont indépendants l'un de l'autre et qu'il n'y a aucune interaction entre les deux, les auteurs de ST-ResNet [75] ont observé que les architectures à deux flux ne sont pas capables de modéliser les relations entre l'apparence et le mouvement (qu'est-ce qui bouge dans quel sens). Pour répondre à ce problème, ils ont proposé des fusions au milieu qui consistent à injecter des connexions résiduelles entre les deux flux, à partir du flux de temporel vers le flux spatial.

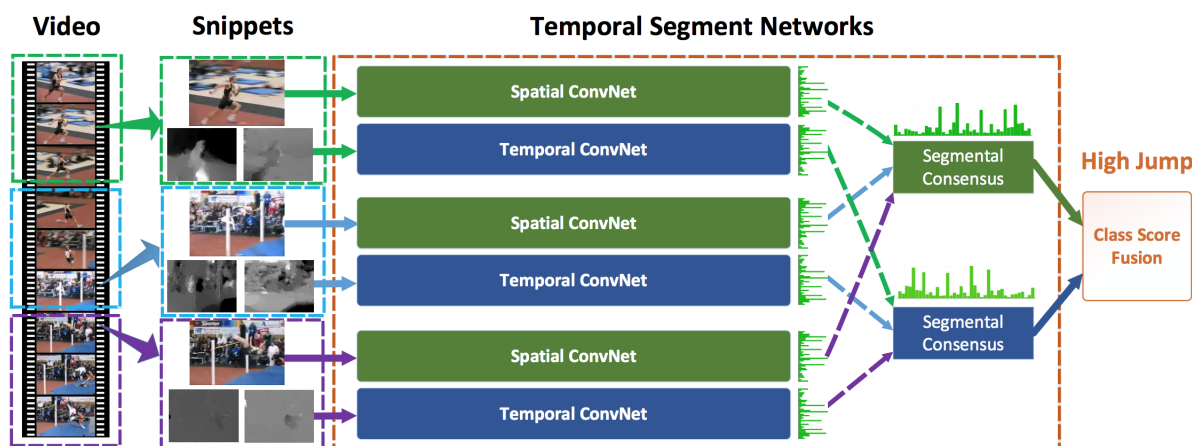


FIGURE 2.13 Schéma global du réseau de segments temporels à deux flux (TSN) [14].

2.2.5 Architectures spatiales et temporelles à long terme

La plupart des approches CNN à deux flux capturent uniquement les informations de mouvement entre les trames adjacentes. Cependant, elles ne permettent de représenter que les mouvements dans des intervalles temporels très courts, généralement entre 5 et 10 trames. De plus, lorsqu'on utilise plusieurs trames, l'agrégation des caractéristiques des CNN ne considère pas l'ordre de ces trames. Afin de mieux analyser les événements complexes qui se déroulent sur une longue période temporelle, on utilise souvent les réseaux récurrents [76]. Ces derniers permettent de modéliser les informations temporelles des données séquentielles en liant l'entrée courante aux entrées précédentes [77]. Puisque le modèle de base des réseaux récurrents souffrait des problèmes de disparition et d'explosion de gradients, il ne permettait pas de prédire convenablement les sorties lorsque les exemples sont longs. Pour pallier ce problème, on a proposé le modèle LSTM.

LSTM (Long short-term memory) est une variante de réseaux de neurones récurrents qui inclut une cellule mémoire et des portes non linéaires. Il est conçu pour stocker et accéder efficacement aux informations les plus importantes dans une longue séquence temporelle. Il contient une porte d'entrée, une porte d'oubli et une porte de sortie (voir figure 2.14). Dans les LSTM, l'information de la cellule mémoire est contrôlée par deux éléments, l'activation de la porte d'entrée (instant courant) et l'unité de mémoire de la cellule précédente (instant précédent) qu'on multiplie par l'activation de la porte d'oubli (instant courant) afin d'apprendre à oublier l'information des états précédents. De cette façon, les portes d'oubli permettent aux LSTM d'apprendre les étapes temporelles (des états cachés précédents) à partir desquelles on doit considérer les mises à jour du réseau.

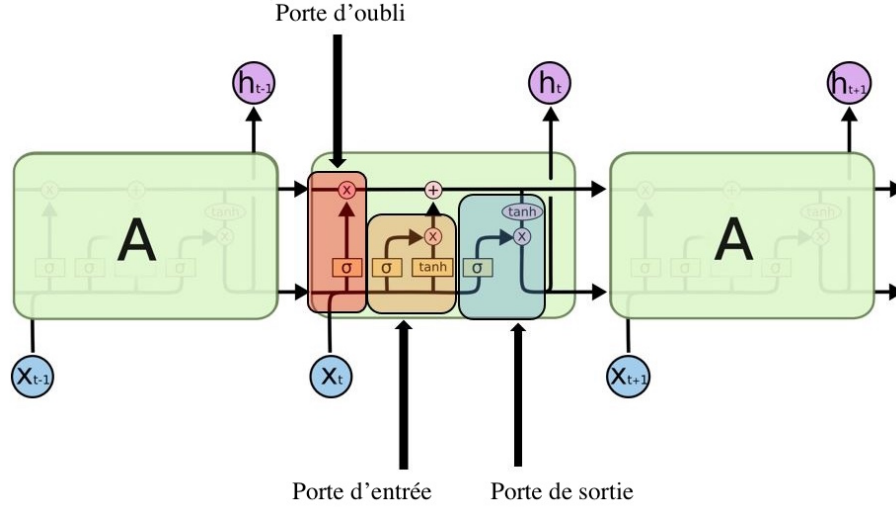


FIGURE 2.14 L'architecture d'un réseau LSTM [15].

Puisque les LSTM permettent de stocker les entrées et les états précédents pour calculer la sortie courante, ils permettent ainsi de capturer convenablement les dépendances temporelles. Ils sont perçus être meilleurs que les différentes techniques de sous-échantillonnage temporel car ils capturent la structure temporelle des trames à long terme. Dans le contexte de la reconnaissance d'activités humaines via les vidéos, les LSTM sont entraînés sur l'ordre des séquences des caractéristiques extraites des trames afin de prédire le résultat final. Pour ce faire, on ajoute des couches récurrentes de LSTM à la sortie des CNN [16, 73]. De cette façon, on exploite les avantages des CNN et des LSTM en capturant les caractéristiques spatiales ainsi que les dépendances à long terme. On peut également entraîner des LSTM sur les caractéristiques obtenues par les CNN 3D [78] afin d'apprendre les informations spatio-temporelles à long terme. Il est possible aussi d'utiliser des réseaux de convolution récurrents (LRCN) [16], qui consistent à entraîner de bout en bout les CNN et les LSTM conjointement [16] (voir figure 2.15). D'autres travaux utilisent les LSTM bi-directionnels au lieu des LSTM unidirectionnels [79]. Cependant, une limitation de ces approches est que les séquences doivent généralement avoir une taille fixe en termes de trames. Pour surmonter cette limitation, d'autres recherches ont exploré l'utilisation de réseaux auto-attentionnels plutôt que des LSTM pour modéliser les relations entre les éléments d'une séquence.

Pour optimiser les architectures spatiales et temporelles à long terme, les auteurs de [17] ont proposé un modèle d'attention pour apprendre à travers les trames les régions qui sont les plus pertinentes pour le processus de la reconnaissance d'activités. Par exemple, si une personne joue au golf, le modèle d'attention donne plus d'importance à la balle et à l'humain plutôt qu'aux autres éléments. Pour ce faire, le mécanisme d'attention prend en entrée les

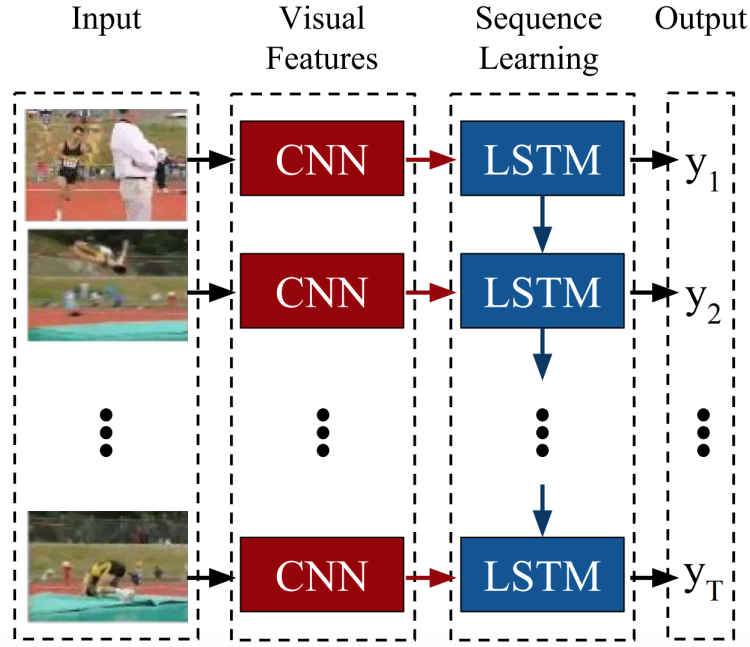


FIGURE 2.15 L'architecture du réseau de convolution récurrent (LRCN) [16].

caractéristiques profondes extraites de la trame par le CNN et produit en sortie des caractéristiques pondérées. Le poids de chaque caractéristique est affecté selon l'importance de la région correspondante à l'instant précédent. Pour générer la distribution des régions les plus importantes à l'instant courant, on apprend à prédire l'état caché de l'unité LSTM à l'instant précédent qui va correspondre à la sortie désirée. La figure 2.16 illustre le modèle d'attention appliqué dans [17].

2.2.6 Architectures auto-attentives

Un modèle auto-attentif est un modèle d'apprentissage profond proposé en 2017 par Vaswani et al. [18] qui se distingue par l'utilisation de mécanismes d'attention plutôt que de récurrences ou de convolutions. Cette approche permet d'établir des dépendances globales entre l'entrée et la sortie, ce qui permet de donner plus de poids à certaines parties de l'entrée en fonction de leur pertinence pour la tâche à accomplir. Bien que les modèles auto-attentifs aient été développés à l'origine pour le traitement du langage naturel, ils ont depuis trouvé des applications dans d'autres domaines, tels que la vision par ordinateur [80].

Les modèles auto-attentifs ont souvent une structure encodeur-décodeur. L'encodeur est composé d'une série de couches d'encodage qui traitent la séquence d'entrée couche par couche de manière itérative et la transforme en représentations riches, tandis que le décodeur prend ces

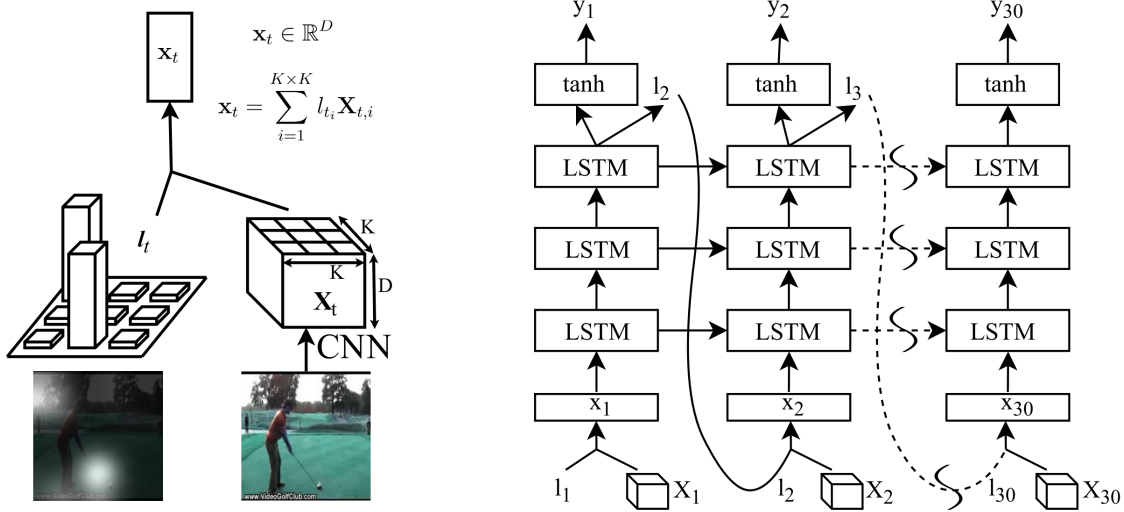


FIGURE 2.16 Modèle d'attention appliqué à la reconnaissance d'activités humaines [17].

représentations et les utilise pour générer la sortie désirée. Spécifiquement, l'encodeur consiste en plusieurs couches où chacune est composée d'une sous-couche d'auto-attention multi-têtes (*multi-head self-attention mechanism*) suivie par un réseau de neurones entièrement connecté (*fully connected feed-forward network*). Chaque couche est dotée d'une connexion résiduelle et est suivie d'une normalisation de couche (*layer normalization*), permettant ainsi de capturer les différentes relations entre les éléments de l'entrée. Le décodeur, est aussi composé de plusieurs couches identiques à l'encodeur, mais avec une sous-couche additionnelle, en plus de l'auto-attention multi-têtes et le réseau de neurones entièrement connecté. Cette sous-couche effectue une attention multi-têtes sur la sortie de l'encodeur. La figure 2.17 illustre l'architecture du modèle auto-attentif, où la moitié gauche correspond à l'encodeur et la moitié droite représente le décodeur.

L'architecture originale du modèle auto-attentif proposé par Vaswani et al. [18] utilise une fonction d'attention de produit scalaire pondéré (*Scaled Dot-Product Attention*), qui associe un vecteur de requête et un ensemble de paires de vecteurs clé-valeur à un vecteur de sortie. La figure 2.18 à gauche illustre le mécanisme d'attention de produit scalaire pondéré. Formellement, cette fonction d'attention est définie par :

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (2.2)$$

où Q , K et V représentent respectivement l'ensemble des requêtes, des clés et des valeurs regroupées en matrices, et d_k représente la dimension des requêtes et des clés. Il est à noter

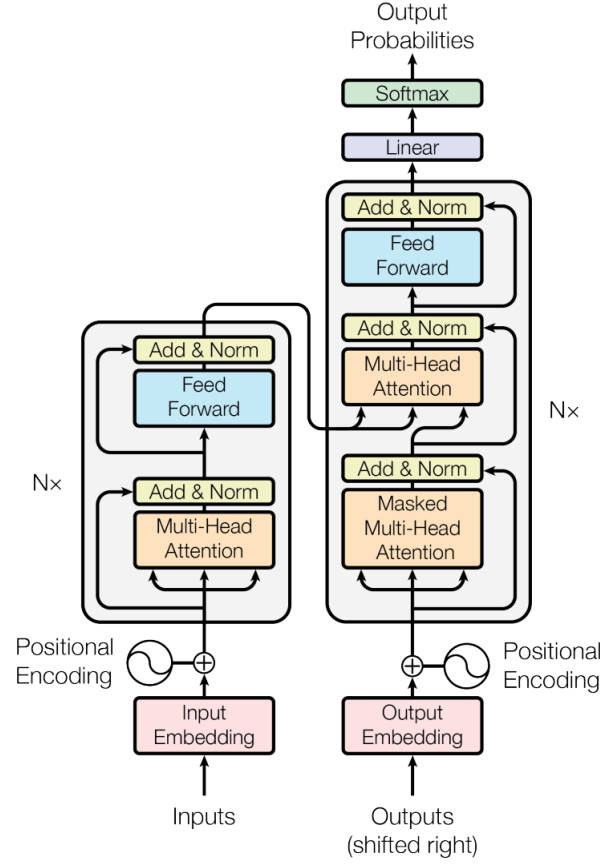


FIGURE 2.17 L'architecture du modèle auto-attentif [18].

que les valeurs des produits scalaires sont ajustées en les divisant par la racine carrée de d_k . Cette étape est nécessaire pour éviter que les produits scalaires ne deviennent trop grands et que la fonction Softmax ne soit entraînée dans des zones où les gradients sont extrêmement petits.

Le module d'auto-attention multi-têtes est un composant essentiel du modèle auto-attentif. Plutôt que de calculer l'attention sur la totalité de l'entrée en une seule fois, l'attention multi-têtes divise les entrées en sous-ensembles plus petits. Plus précisément, le modèle effectue plusieurs projections linéaires des vecteurs d'entrée, puis calcule l'attention de produits scalaires pondérés pour chacun de ces espaces de projection de manière parallèle. Les sorties d'attention indépendantes sont ensuite combinées et transformées pour obtenir les dimensions souhaitées. La figure 2.18 à droite, illustre le mécanisme d'attention multi-têtes. Formellement, ce mécanisme peut être exprimé comme suit :

$$\text{Multi-Head}(Q, K, V) = \text{Concat}(\text{head}_1; \text{head}_2; \dots; \text{head}_h)W^O \quad (2.3)$$

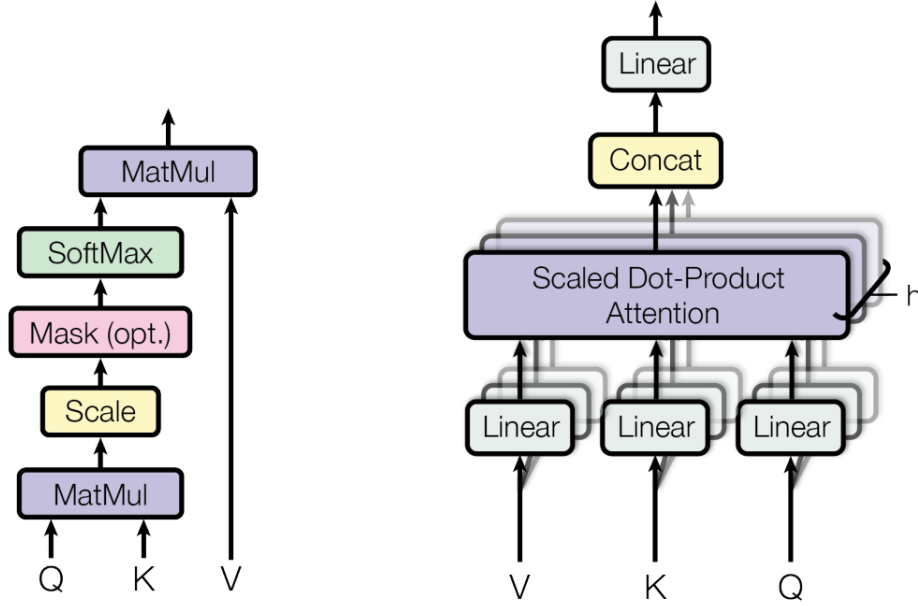


FIGURE 2.18 Illustration du mécanisme d'attention de produit scalaire pondéré (Scaled Dot-Product Attention) (à gauche), et du mécanisme d'attention multi-têtes (Multi-Head Attention) (à droite) [18].

avec :

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2.4)$$

où le symbole ; représente l'opération de concaténation, $W_i^Q, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, et $W^O \in \mathbb{R}^{h d_v \times d_{\text{model}}}$ représentent les projections des matrices de paramètres. d_k représente la dimension des valeurs, et d_{model} est un hyperparamètre qui définit la dimension de l'espace vectoriel contextuel.

L'attention multi-tête permet au modèle de se focaliser sur plusieurs aspects différents des données d'entrée simultanément, ce qui peut améliorer les performances du modèle. Elle facilite également l'apprentissage en permettant au modèle d'ajuster les poids d'attention de manière plus précise et plus fine pour chaque sous-ensemble d'entrée.

Dans les modèles auto-attentionnels, l'encodage de position permet d'ajouter l'information relative à la position de chaque jeton (*token*) dans la séquence. Cette étape est nécessaire, car ces modèles n'utilisent pas de convolutions ou de récurrences pour capturer l'ordre de la séquence. L'encodage de position est ajouté aux vecteurs contextuels denses d'entrée avant d'être passés à l'encodeur et au décodeur, étant donné que sa dimension est similaire à celle des vecteurs contextuels denses d'entrée. Le modèle auto-attentionnel de base [18] considère deux fonctions

d'encodages : sinus et cosinus de différentes fréquences :

$$\begin{aligned} PE(pos, 2i) &= \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \\ PE(pos, 2i + 1) &= \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \end{aligned} \quad (2.5)$$

où pos représente la position du jeton dans la séquence et i représente la dimension de l'encodage de position.

Suite au succès des modèles auto-attentionnels dans le domaine du traitement du langage naturel, ceux-ci ont récemment été adaptés à la vision par ordinateur avec les "Vision Transformers" (ViT) [19].

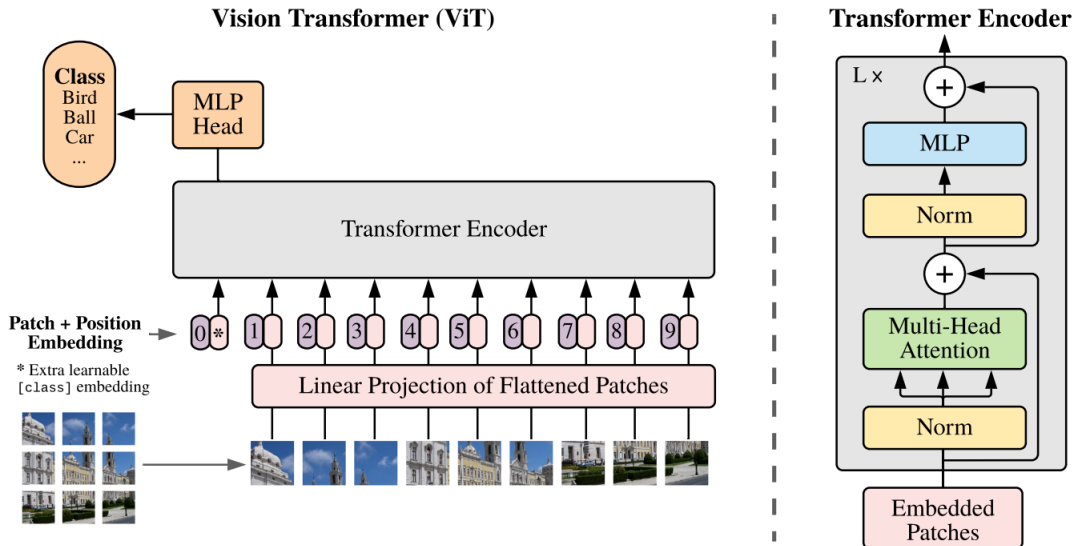


FIGURE 2.19 Architecture du modèle ViT [19].

L'architecture de ViT diffère de celle des modèles auto-attentionnels, car elle ne nécessite pas une structure encodeur-décodeur pour la classification d'images. Ainsi, le modèle ViT se compose uniquement d'une architecture d'encodeur, qui utilise une série de blocs d'auto-attention multi-têtes et de MLP, permettant ainsi de traiter les images de manière efficace. La figure 2.19 illustre l'architecture de ViT.

Dans le modèle ViT, les images sont divisées en plusieurs patches carrés de taille fixe, où chaque patch est équivalent à un "jeton" dans une séquence. Les vecteurs contextuels denses de patch sont ensuite passés à travers des couches d'encodage de position et d'attention multi-têtes

pour capturer les relations spatiales entre les différents patches et construire une représentation de l'image. Les mêmes techniques d'attention multi-têtes et d'encodage de position utilisées dans les modèles auto-attentionnels en traitement de langage naturel sont appliquées aux d'image en utilisant le modèle ViT. Ensuite, la représentation de l'image est alimentée à travers une tête de classification (*MLP Head*) pour classer l'image en fonction des catégories prédéfinies. Cette tête de classification est une série de couches entièrement connectées MLP qui prend en entrée la représentation de l'image générée par les blocs d'encodage et produit les probabilités de sortie pour chaque catégorie.

Le modèle ViT a été utilisé dans plusieurs applications en vision par ordinateur, y compris la classification d'images, la détection d'objets et la segmentation d'images, ainsi que la reconnaissance d'activités. Par exemple, Girdhar et al. [20] ont introduit une architecture appelée Action Transformer Network, basée sur un modèle auto-attentionnel qui incorpore un CNN pour apprendre des caractéristiques spécifiques au contexte spatio-temporel de l'activité humaine à classifier. Cette approche est considérée parmi les premières à avoir utilisé un réseau auto-attentionnel pour la reconnaissance d'activités. Ce modèle utilise des requêtes spécifiques à chaque personne mais agnostiques à la classe, permettant ainsi au modèle d'apprendre à suivre individuellement les personnes dans la séquence vidéo. Le mécanisme d'attention de l'Action Transformer Network apprend à mettre l'accent sur des régions clés, qui sont importantes pour discriminer les différentes activités. La figure 2.20 illustre l'architecture du modèle proposé par Girdhar et al. [20]. Bien que cette approche ait montré des performances remarquables, son intégration des réseaux de neurones convolutifs reste limitative. En effet, ces réseaux ne traitent pas efficacement les informations liées à la position et à la temporalité des activités dans les vidéos. Cette limitation a des impacts directs sur l'efficacité du réseau auto-attentionnel ensuite pour la reconnaissance d'activités.

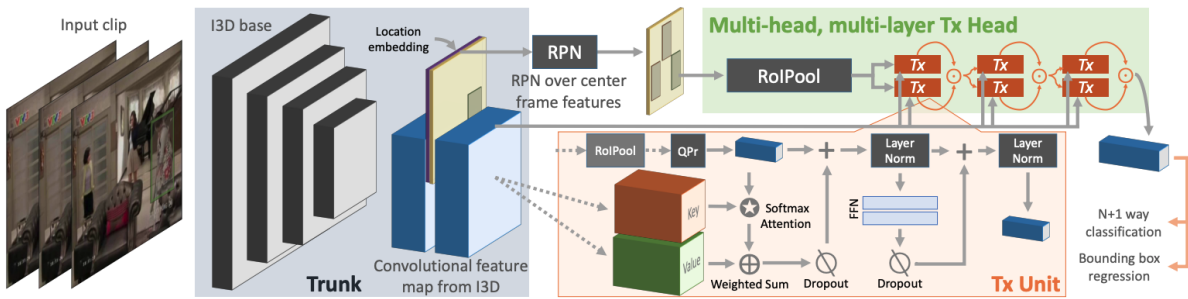


FIGURE 2.20 Architecture du modèle Action Transformer Network [20].

Dans le même contexte, Yang et al [21] ont aussi proposé Recurrent Vision Transformer (RViT) pour la reconnaissance d'activités dans les vidéos. L'architecture de RViT utilise

un processus récurrent pour capturer les caractéristiques spatiales et temporelles grâce à une porte d'attention. La figure 2.21 illustre l'architecture du modèle RViT à deux couches, où les informations spatiales et temporelles sont agrégées verticalement et horizontalement. Étant donné la trame actuelle $x^{(t)}$ et l'état caché $h^{(t-1)}$ de la trame précédente, une sortie $O^{(t)}$ et un état caché $h^{(t)}$ sont générés à partir de l'unité RViT actuelle. Ensuite, une porte d'attention est intégrée pour transférer les attributs temporels à travers l'état caché au lieu d'extraire les attributs temporels à partir de chaque trame en lots. Les attributs temporels agrégés via l'état caché sont utilisés pour mettre en correspondance les attributs spatiaux dans le flux de traitement ultérieur [21].

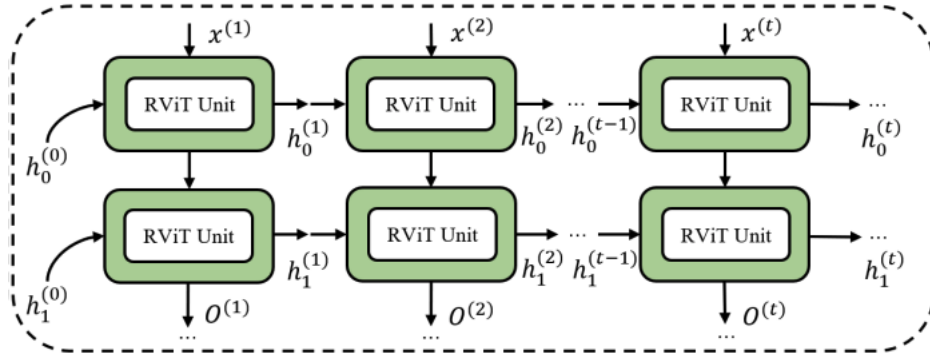


FIGURE 2.21 Architecture du modèle RViT [21].

Bertasius et al. [81] ont développé Timesformer, une méthode de classification vidéo entièrement basée sur l'auto-attention dans l'espace et le temps, qui ne nécessite aucune convolution. Cette approche adapte l'architecture standard des réseaux auto-attentifs aux vidéos en permettant l'apprentissage des caractéristiques spatio-temporelles directement à partir d'une séquence de patches au niveau des images. Timesformer prend en entrée un extrait vidéo et décompose chaque image en N patches non chevauchants, garantissant que l'ensemble des N patches représente l'intégralité de l'image. Par la suite, il projette linéairement chaque patch en un vecteur en utilisant une matrice entraînable. Le réseau auto-attentif de Timesformer est composé de L blocs d'encodage. À chaque bloc, un vecteur de requête, une clé et une valeur sont calculés pour chaque patch, en utilisant la représentation encodée par le bloc précédent. Les poids d'auto-attention sont calculés en utilisant le produit scalaire entre les vecteurs de requête et de clé. Ensuite, la concaténation de ces vecteurs provenant de toutes les têtes d'attention est projetée et passe à travers un MLP, avec des connexions résiduelles après chaque opération. Finalement, la projection de l'extrait vidéo est obtenue à partir du dernier bloc, en utilisant le jeton de classification [81].

Un autre modèle qui incorpore les réseaux auto-attentifs est Vidéo Transformer (VidTr) [22].

Ce dernier adopte une architecture basée sur des empilements d'attention spécialement conçus pour la reconnaissance d'activités dans les vidéos. Comparé aux réseaux convolutifs en 3D, VidTr agrège les informations spatio-temporelles via des attentions empilées. Une attention séparable multi-têtes (MSA) est également utilisé dans ce modèle qui consiste à dissocier l'auto-attention 3D en attention spatiale et en attention temporelle. De plus, il utilise un échantillonnage temporel pour éliminer la redondance dans la représentation des données vidéo [22]. L'architecture du modèle est illustré dans la figure 2.22, elle prend en entrée des patches de pixels et apprend les caractéristiques spatiales et temporelles via l'attention séparable.

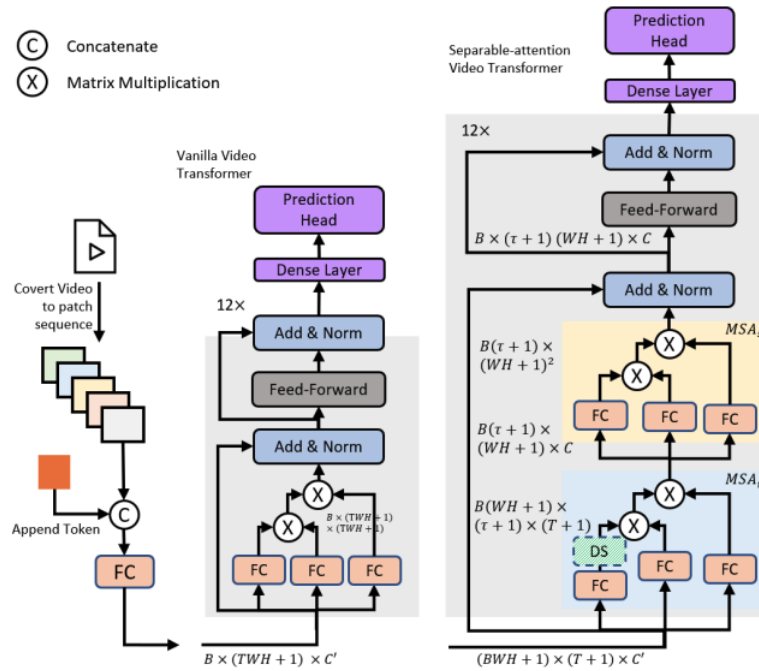


FIGURE 2.22 Architecture du modèle VidTr [22].

Inspiré par le succès de ces approches, nous proposons une méthode de reconnaissance des activités qui exploite les capacité des réseaux auto-attentifs à capturer les relations complexes dans les données. En outre, nous incorporons une sélection non-paramétriques des poses clés visant à relever les défis persistants inhérents à la modélisation des dynamiques des activités, qu'elles soient de courte ou de longue durée.

2.2.7 Discussion

Dans cette section, nous avons discuté les méthodes d'apprentissage profond pour la reconnaissance d'activités. Nous avons vu que les CNN 2D permettent de modéliser efficacement

les informations spatiales contenues dans les images. Nous avons vu également que les architectures à deux flux différents et les architectures CNN 3D sont importantes pour apprendre les informations spatio-temporelles. Cependant les CNN 3D requièrent beaucoup de données d'entraînement étant donnée la complexité d'entraîner des filtres 3D. Nous avons évoqué que les informations temporelles à long terme sont modélisées en appliquant un réseau LSTM sur les caractéristiques profondes extraites [16, 73]. En complément de ces méthodes, les modèles auto-attentionnels se différencient des CNN et des LSTM en ce qu'ils n'ont pas besoin de convolutions ou de réseaux récurrents pour modéliser les relations spatiales et temporelles dans les vidéos. Au lieu de cela, ils utilisent une architecture d'auto-attention pour modéliser les relations entre tous les éléments d'une séquence, y compris les séquences de tailles variables, ce qui leur permet de capturer des informations à la fois spatiales et temporelles à long terme.

2.3 Méthodes de reconnaissance d'activité basées sur l'apprentissage profond versus méthodes classiques

Les méthodes classiques et les méthodes par apprentissage profond sont largement exploitées et souvent complémentaires pour la reconnaissance d'activités humaines. Le choix entre ces deux approches dépend du contexte, de la taille des données et des ressources de calcul disponibles. Cependant, il est clair que les méthodes par apprentissage profond sont devenues plus populaires ces dernières années en raison de leur capacité à capturer des informations complexes contenues dans des séquences vidéos brutes. Lorsqu'on compare les caractéristiques des méthodes classiques aux caractéristiques profondes, ces dernières semblent être supérieures pour la classification. Elles ont une grande capacité de généralisation sur les ensembles de données récents. Toutefois, les méthodes par apprentissage profond nécessitent généralement de grandes quantités de données vidéos annotées pour l'entraînement, ce qui peut être difficile et coûteux à obtenir dans de nombreux cas. De plus, elles requièrent des ressources de calcul substantielles pour l'apprentissage et l'inférence, ce qui peut limiter leur applicabilité dans des environnements avec des contraintes de puissance de calcul ou dans les systèmes embarqués.

D'un autre côté, les approches basées uniquement sur les caractéristiques classiques comme les trajectoires denses améliorées [56] offrent à ce jour des performances encore comparables à celles des approches basées sur l'apprentissage profond. Cependant, pour améliorer les performances des méthodes d'apprentissage profond, on combine souvent les représentations profondes avec les caractéristiques classiques. Des travaux de recherche ont montré que cette combinaison peut améliorer significativement les performances de la reconnaissance d'activités dans les vidéos. Par exemple, le travail proposé par Wang et al. [74] a démontré l'efficacité

de cette approche en combinant les caractéristiques classiques des trajectoires avec des représentations profondes apprises par un réseau de neurones convolutif. Les résultats ont montré une amélioration significative des performances par rapport à l'utilisation isolée de l'une ou l'autre de ces approches.

2.4 Ensembles de données

Il existe plusieurs jeux de données disponibles pour la reconnaissance d'activités humaines dans les vidéos. Parmi ces ensembles on peut citer KTH [23], Weizmann [6], Collective Activity [1], Collective Activity Extended [2], Volleyball [24], InfAR [25], et UCLA Courtyard [26]. Selon l'environnement, les ensembles de données sont généralement répartis en deux catégories : environnement contrôlé ou environnement réaliste. Dans ce qui suit, nous présentons les jeux de données existants qui sont pertinents.

2.4.1 Environnements contrôlés

KTH [23] est l'un des ensembles de données les plus anciens en reconnaissance d'activités. Il contient 600 courtes vidéos de 6 activités humaines (marche, jogging, course, boxe, mouvement de main, applaudissement) exercées par 25 personnes dans quatre scénarios différents (intérieur et extérieur avec différents vêtements). La figure 2.23 montre des exemples de l'ensemble de données KTH.



FIGURE 2.23 Exemples de l'ensemble de données "KTH" [23].

Weizmann [6] est un ensemble de données qui contient 90 vidéos de 10 classes d'activités (marcher, courir, sauter, galoper latéralement, se plier, bouger une seule main, bouger deux

mains, sauter sur place, saut jack et skipping) exercés par 9 personnes différentes, toutes les vidéos sont acquises par des caméras fixes. La figure 2.24 montre des exemples de l'ensemble de données Weizmann.

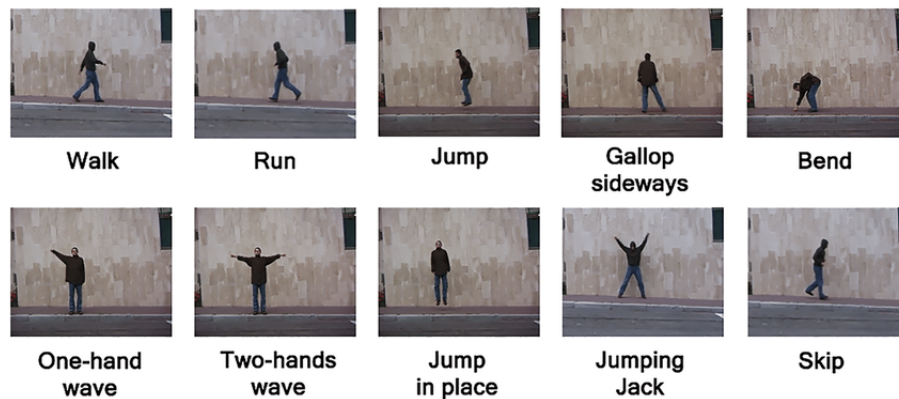


FIGURE 2.24 Exemples de l'ensemble de données "Weizmann" [6].

2.4.2 Environnements réalistes

Collective Activity (CA) [1] est un jeu de données qui contient 5 différentes activités collectives (traverser, marcher, attendre, parler et faire la queue). En plus de l'activité de chaque personne, on fournit également l'activité la plus représentative de la scène. Les vidéos sont acquises par des caméras tenues par des humains. Cet ensemble est constitué de 44 vidéos. La figure 2.25 montre des exemples de l'ensemble de données Collective Activity.

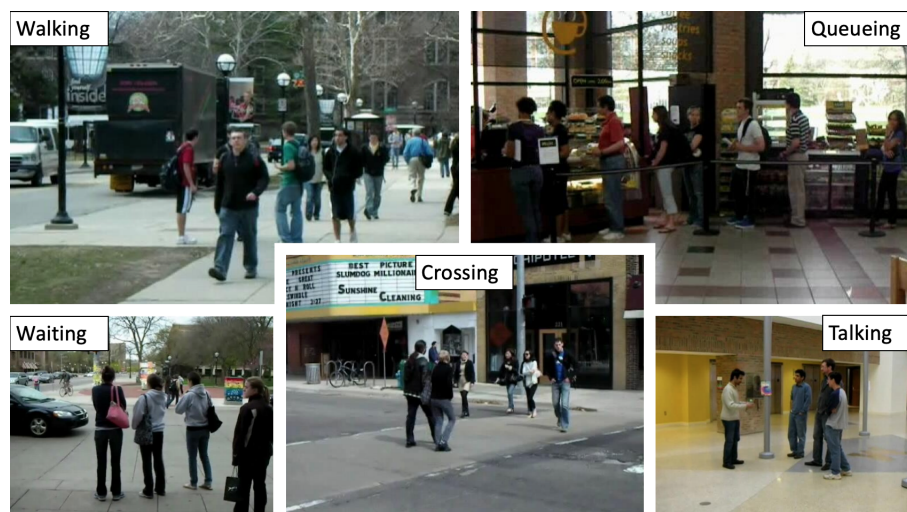


FIGURE 2.25 Exemples de l'ensemble de données "Collective Activity" [1].

Collective Activity Extended (CAE) [2] est une version augmentée du jeu de données Collective Activity [1] dans laquelle deux nouvelles activités "danser" et "courir" ont été ajoutées. Le nombre total des vidéos dans cette nouvelle version est de 75.

Volleyball [24] est un jeu de données spécifiquement conçu pour la reconnaissance d'activités dans les vidéos du sport de volleyball. Il est constitué de 55 séquences vidéo, à partir desquelles 4 830 images ont été annotées manuellement. Ce jeu de données fournit des annotations détaillées pour 9 étiquettes d'activités individuelles des joueurs ainsi que pour 8 étiquettes d'activités d'équipe. Les activités individuelles comprennent : Attente, Passe, Défense, Chute, Attaque, Blocage, Saut, Déplacement, Position debout. En ce qui concerne les activités de groupe, elles incluent : Passe droite, Attaque droite, Réception droite, Winpoint droit, Winpoint gauche, Réception gauche, Attaque gauche, Passe gauche. La résolution des vidéos varie entre 1920×1080 et 1280×720 . La figure 2.26 montre un exemple de l'ensemble de données Volleyball.

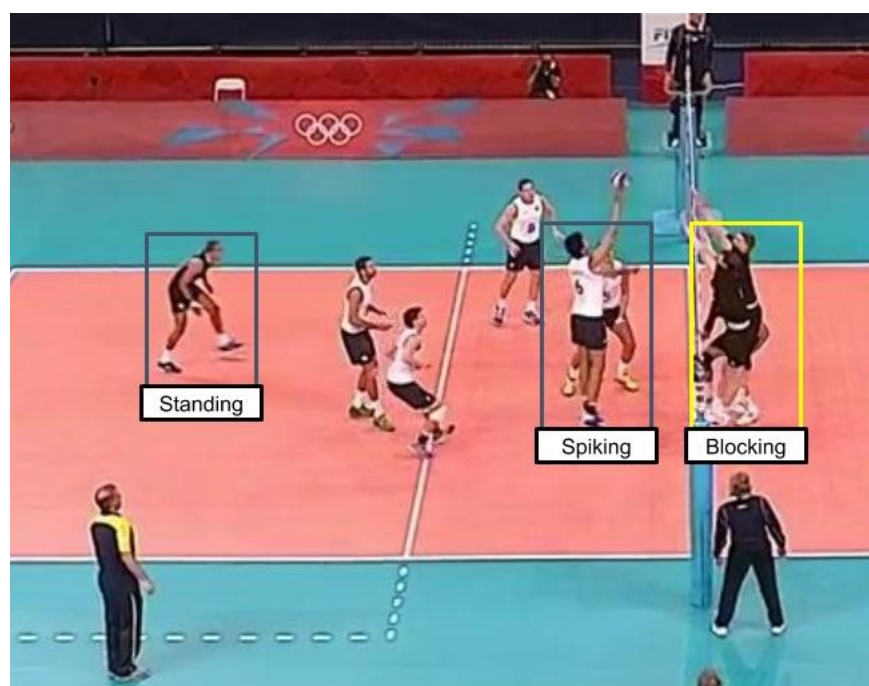


FIGURE 2.26 Un exemple d'image du jeu de données "Volleyball" [24].

InfAR [25] est un ensemble de données spécialement conçu pour la reconnaissance d'activités dans le spectre infrarouge. Il contient 12 annotations d'activités humaines, où chacune des activités est représentée par 50 vidéos. Les annotations comprennent les activités suivantes : faire signe d'une main, faire signe de deux mains, applaudir, marcher, courir, sauter, sauter à la corde, serrer la main, étreindre, pousser, donner un coup de poing et se battre. Les vidéos

ont une résolution de 293×256 pixels avec une fréquence d'images de 25 images par seconde. La figure 2.27 montre des exemples du jeu de données InfAR.

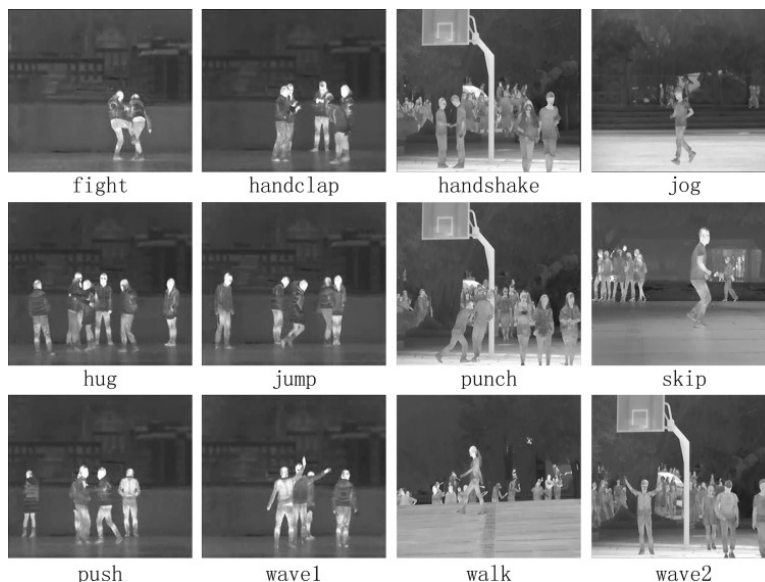


FIGURE 2.27 Un exemple d'images du jeu de données "InfAR" [25].

UCLA Courtyard [26] est un jeu de données qui contient un ensemble de vidéos de 106 minutes prises depuis une vue en hauteur, montrant une scène d'une cour du campus de l'UCLA. Les vidéos ont une résolution de 2560×1920 pixels avec une fréquence de 30 images par seconde. Ce jeu de données est étiqueté avec six activités de groupe (marcher ensemble, faire la queue, discuter en groupe, s'asseoir ensemble, attendre en groupe et visite guidée) et dix activités individuelles (faire du skateboard, faire du vélo, faire du scooter, conduire une voiture, marcher, parler, attendre, lire, manger et s'asseoir). La figure 2.28 montre un exemple de l'ensemble de données UCLA Courtyard.

L'objectif de cette thèse est de développer des techniques et des modèles de reconnaissance d'activité plus robustes et adaptés aux contextes complexes et dynamiques. En effet, nous nous concentrerons sur les jeux de données de reconnaissance d'activité dans des environnements réalistes, qui reflètent des conditions du monde réel, plutôt que de se limiter aux environnements contrôlés, qui présentent des conditions expérimentales spécifiques et prédéfinies.



FIGURE 2.28 Un exemple d'image du jeu de données "UCLA Courtyard" [26].

CHAPITRE 3 SURVOL DES MÉTHODES PROPOSÉES

Le format de cette thèse est structuré en articles, présentant les travaux de recherche effectués pendant mon doctorat. Les trois chapitres suivants sont constitués d'articles publiés à des congrès, chacun répondant aux objectifs fixés et proposant des solutions adaptées à plusieurs problématiques liées à la reconnaissance d'activités humaines dans les vidéos. Il convient de souligner que l'un des défis majeurs auxquels nous avons été confrontés au cours de cette thèse était de trouver des jeux de données adaptés au contexte de la reconnaissance d'activité humaine dans les espaces publics. En effet, les données disponibles étaient souvent limitées en termes de diversité des scénarios, de qualité ou de disponibilité d'annotations. Ceci nous a conduit à explorer et proposer des méthodes de reconnaissance d'activité robustes qui font face aux limitations des données disponibles.

Comme discuté dans les éléments de la problématiques au chapitre 1, les occlusions partielles ainsi que les variations d'apparence au sein d'une même classe constituent des défis majeurs de la reconnaissance d'activités, particulièrement dans des environnements réalistes. Les travaux de recherche existants ont tenté d'aborder ces problèmes en exploitant diverses approches [34, 82], telles que l'utilisation de modèles de réseau de neurones profonds et l'intégration de mécanismes d'attention pour se concentrer sur les parties visibles malgré les occlusions. Cependant, trouver une solution universelle reste difficile en raison de la complexité et de la variété des situations réelles.

C'est dans cette optique que nous avons exploré l'utilisation de la pose humaine pour la reconnaissance d'activité. Plus spécifiquement, l'estimation de la pose consiste à extraire des informations de posture à partir des séquences vidéos pour chaque individu dans la scène observée. Nous avons choisi cette approche, car l'estimation de la pose peut fournir des informations riches sur les mouvements et les positions des différentes parties du corps, permettant ainsi de capturer les caractéristiques clés de l'activité humaine et ce même en cas d'occlusions partielles ou de variations d'apparence. Contrairement aux méthodes traditionnelles qui se basent principalement sur l'analyse globale de la scène [83], notre approche basée sur la pose humaine offre une représentation abstraite qui est plus résiliente aux occlusions partielles et aux variations d'apparence. En effet, la pose humaine capture les relations spatiales entre les différentes parties du corps, telles que les bras, les jambes et le torse. En se concentrant sur ces relations, il devient possible d'inférer des activités spécifiques sans dépendre uniquement des détails d'apparence qui pourraient être occultés par des objets ou d'autres personnes.

Généralement, les poses extraites pour chaque individu dans la séquence vidéo peuvent être

redondantes et peu pertinentes pour la reconnaissance d’activités, ce qui peut affecter la précision de la classification de l’activité. Pour résoudre ce problème, nous avons adopté un algorithme de regroupement pour sélectionner les poses les plus pertinentes, et nous avons combiné ces poses sélectionnées avec leurs images correspondantes de personnes extraites de l’image originale. Nous avons ensuite créé une structure de grille qui résume les moments clés de chaque personne en utilisant la combinaison d’informations de pose et de contexte. La représentation basée sur la grille s’est avérée bénéfique car elle a permis d’obtenir une vue d’ensemble des activités dans la scène tout en réduisant la complexité du problème de reconnaissance d’activités, en ciblant mieux les parties de la séquence vidéo les plus importantes. Cette approche est décrite en détails dans le chapitre 4 incluant également une évaluation approfondie de cette méthode sur plusieurs jeux de données de reconnaissance d’activités, et une comparaison de nos résultats à ceux d’autres méthodes existantes. Les résultats montrent que notre approche basée sur l’estimation de la pose humaine obtient des performances compétitives par rapport aux méthodes existantes dans ce domaine.

Lors de notre exploration des jeux de données adaptés, nous avons rencontré un défi majeur lié à l’incomplétude des annotations. En effet, la plupart des annotations ne couvraient pas toutes les actions présentes dans une séquence vidéo, et se limitent généralement à l’action principale dans toute la séquence, sans aucune référence aux personnes impliquées. Cette approche ne correspondait pas à nos objectifs de recherche initiaux, qui visaient à analyser et comprendre l’ensemble des actions de chaque individu dans les vidéos. Les méthodes de reconnaissance d’activités existantes [25, 84], lorsqu’elles sont appliquées à ce genre de jeux de données, extraient les caractéristiques globales de la séquence en entrée, et apprennent des schémas qui ne sont pas nécessairement discriminants pour l’action principale ou la plus intense dans une scène. Pour remédier à ce problème, nous avons dû proposer une méthode spécifique pour extraire l’action principale dans une séquence vidéo donnée et reconnaître son activité. Nous avons également évalué cette méthode sur un jeu de données infrarouge afin de valider sa performance et sa robustesse dans des conditions différentes, où les informations visuelles sont limitées ou ambiguës.

Le modèle proposé est capable d’identifier l’acteur principal qui exerce l’action d’intérêt sans utiliser de connaissances préalables ou d’annotations explicites. Notre modèle fonctionne selon un processus en trois étapes. Tout d’abord, nous identifions les personnes clés dans une vidéo qui exécutent une action en nous basant sur le flux optique. En effet, notre modèle évalue chaque mouvement dans la vidéo pour déterminer les acteurs qui ont une incidence sur l’activité principale, et ignorent ceux qui ne contribuent pas significativement à celle-ci. Dans la deuxième étape, nous extrayons les informations de pose pour chaque acteur clé identifié. En tenant compte de ces informations de posture, nous construisons des indices

riches pour compenser l'absence d'informations de texture et de couleur dans le spectre infrarouge. Comme la plupart des modèles d'estimation de pose sont entraînés dans le spectre visible, ils peuvent échouer dans certains cas lorsqu'ils sont appliqués directement au domaine infrarouge. Pour résoudre ce problème, nous proposons un mécanisme de filtrage de pose qui exploite les connaissances apprises à travers les vastes ensembles de données RGB. À partir des séquences de poses filtrées, nous extrayons les poses clés de l'acteur principal pour restreindre l'analyse aux caractéristiques temporelles pertinentes de l'action. Enfin, nous intégrons les caractéristiques les plus représentatives dans une représentation qui encode à la fois les informations infrarouges et les informations de pose, afin d'obtenir une représentation plus pertinente de l'action réalisée. Cette méthode est présentée en détail dans le chapitre 5.

Les chapitres 4 et 5 présentent des méthodes qui reposent sur l'idée que chaque activité exécutée par un acteur peut être décrite par un ensemble fixe de poses clés. Ces poses clés constituent une représentation finale de toutes les postures de la personne tout au long d'une séquence vidéo. Cependant, lorsqu'il s'agit de séquences vidéo de longueurs variables, cette approche peut être limitative et peut exclure des informations importantes pour la reconnaissance de l'activité. Simultanément, diverses recherches dans la littérature se sont concentrées sur la résolution du défi [85, 86] de la modélisation temporelle des actions humaines. Notamment, dans le contexte des méthodes basées sur les réseaux de neurones convolutifs, la capacité de modélisation à long terme s'est révélée limitée en raison de la nature locale de l'opération de convolution. Les noyaux de petite taille inhérents à cette opération ne parviennent qu'à capturer les informations présentes dans leur voisinage immédiat. En contraste, les mécanismes d'attention [18, 87] sont plus avantageux pour la modélisation à plus longue échelle.

Dans cette direction, nous proposons dans le chapitre 6 une approche de reconnaissance d'activités humaines non paramétrique, qui permet de prendre en compte efficacement les séquences vidéos de différentes longueurs et de capturer des informations plus riches et discriminantes de l'action. Notre méthode modélise efficacement les relations temporelles au sein des actions et s'adapte dynamiquement pour capturer toutes les caractéristiques clés dans une séquence vidéo. Au lieu de fournir au modèle d'apprentissage des caractéristiques génériques sur les humains et la scène, nous adoptons un réseau auto-attentif pur qui utilise les représentations les plus discriminantes des actions spécifiques à chaque personne. Ces représentations sont dérivées des poses clés estimées qui décrivent mieux les activités effectuées. En procédant ainsi, nous réduisons considérablement la quantité de biais inductif par rapport à l'utilisation de réseaux de neurones convolutifs.

CHAPITRE 4 ARTICLE 1 : A GRID-BASED REPRESENTATION FOR HUMAN ACTION RECOGNITION

Lamghari, S., Bilodeau, G. A., Saunier, N.

25th International Conference on Pattern Recognition (ICPR), May 5, 2021,
pp. 10500-10507.

doi : 10.1109/ICPR48806.2021.9413136

Abstract

Human action recognition (HAR) in videos is a fundamental research topic in computer vision. It consists mainly in understanding actions performed by humans based on a sequence of visual observations. In recent years, HAR have witnessed significant progress, especially with the emergence of deep learning models. However, most of existing approaches for action recognition rely on information that is not always relevant for this task, and are limited in the way they fuse the temporal information. In this paper, we propose a novel method for human action recognition that encodes efficiently the most discriminative appearance information of an action with explicit attention on representative pose features, into a new compact grid representation. Our GRAR (Grid-based Representation for Action Recognition) method is tested on several benchmark datasets demonstrating that our model can accurately recognize human actions, despite intra-class appearance variations and occlusion challenges.

4.1 Introduction

Human Action Recognition (HAR) is a very popular topic in computer vision. The popularity of this task is mainly due to its use in various real-world applications such as smart video surveillance, autonomous robots, virtual reality, sport video analysis and urban planning [24, 88–90]. The goal of HAR is to identify and classify human actions from video sequences that contain spatial and temporal information related to the performed human action. Actions can be complex (e.g., preparing a meal) or simpler (e.g., walking). In this work, we focus on atomic human actions (e.g., running, dancing, jumping). Despite great progress in the last few years, human action recognition is still a challenging task due to dynamic backgrounds, occlusion, varied people appearance and imaging conditions.

In the last few years, there has been a lot of research based on deep learning to recognize human actions in videos [10, 13, 83, 91, 92]. Since videos are 3D spatio-temporal signals, the

main idea behind the majority of these studies is to extend Convolutional Neural Networks (CNN) to include the temporal information contained in videos. Karpathy et al. [10] proposed several fusion techniques that slightly modify the CNN architectures to operate on stacked video frame inputs. As their results were similar to the results obtained by using individual RGB frames, these techniques were shown to not correctly model the temporal information. In order to operate in the spatio-temporal domain, Ji et al. [67] proposed a 3D CNN model that performs 3D convolutions on stacked video frames to learn spatio-temporal information between consecutive frames. In addition to the fact that 3D CNN perform similarly to 2D CNN, they are computationally expensive to train because they contain many more parameters and do not model long range temporal information. In the same context, Simonyan et al. [13] proposed a two-stream CNN architecture that learns spatial appearance information from RGB frames and motion information between frames using optical flow. To improve this architecture that considers only a single frame as input, Ng et al. [73] and Wang et al. [14] proposed architectures that aggregate the convolutional features at different temporal and spatial positions. However, the streams in these two-stream CNN architectures are independent and there is no shared information between them. These architectures capture only the motion information in short time windows and do not guarantee to keep the most representative features with pooling techniques.

Another line of research incorporates human pose sequences to represent actions as they provide valuable cues for the recognition task. Multiple studies proposed to recognize actions based on 3D poses [28,30,31,93]. However, these methods are less convenient for general cases, because they require special depth sensors. With the research progress in pose estimation over the past few years, some alternative approaches exploit 2D poses to recognize human actions [32,33,94]. Similarly to the previously discussed approaches, 2D pose-based methods still represent actions by randomly learned features. Also, they remain limited in the way they integrate the temporal information that is irrelevant with respect to the dynamic nature of human actions.

To address the aforementioned problems, we propose a novel pose-based approach for human action recognition that learns the temporal discriminative features of actions by integrating them in a compact static representation. Based on key poses, our model consists in integrating the most representative appearance features in a single grid image to obtain a more relevant representation of the performed action. In order to restrict the analysis to only the most likely information related to the action, we only consider the human region in the scene in each frame. By fusing valuable appearance features with representative poses, our grid representation mimics an explicit attention mechanism that allows us to deal with some challenges related to real-world data including occlusions and intra-class appearance

variations. Furthermore, we tested our approach, called GRAR (Grid-based Representation for Action Recognition) on several datasets and found that it yields competitive or state-of-the-art results for individual actions as well as collective activities when integrated into a bottom-up setting.

The contributions of this paper are threefold :

1. We propose a new grid action representation that encodes only discriminative appearance features.
2. We consider an explicit attention mechanism that highlights the representative poses of the action and can handle challenging situations, such as occlusions and intra-class variations.
3. Experiments on three publicly available benchmark datasets demonstrate the effectiveness of our proposed model by achieving competitive results compared to the state-of-the-art.

4.2 Related work

Our proposed work is related to three lines of research : deep learning-related action recognition, pose-based video action recognition and collective activity recognition in videos. In this section, we review notable studies related to these research areas and show how our proposed method differs from them.

4.2.1 Deep Learning-based Action Recognition

In recent years, deep learning methods have shown valuable capabilities in various computer vision applications ranging from image classification to action recognition in videos. CNN [9] are regarded as a powerful class of models for the task of video action recognition. Simonyan et al. [82] proposed to integrate spatial and temporal networks into a two-stream CNN architecture that is trained independently on inputs from static appearance and multi-frame dense optical flow. Similarly, based on a two-stream CNN architecture, Wang et al. [83] introduced a model called TSN that learns video representations based on sparse temporal sampling to encode the long range temporal structure for a better understanding of the dynamics in action videos. A 3D CNN model that extends 2D convolutions to 3D convolutions is proposed by Ji et al. [67] to learn spatio-temporal information between stacked consecutive frames. The problem of 3D CNN is that they contain many more parameters, which makes them computationally expensive. Moreover they do not allow modeling of long range temporal information.

Another widely adopted approach in this context is the use of Recurrent Neural Network (RNN) and its variants (e.g., LSTM and GRU). These networks have demonstrated an impressive performance in modeling long-term dependencies between frames. Donahue et al. [91] applied a long-term recurrent convolutional network to model visual time-series to recognize actions. In a different work, Du et al. [95] introduced a recurrent network based on pose and attention mechanisms, where the spatio-temporal evolution of the human pose is used to guide the process of recognizing human actions in videos. Recently, Li et al. [96] proposed an end-to-end sequence learning framework for action classification that integrates attention via a convolutional LSTM network.

4.2.2 Pose-based Video Action Recognition

Human pose is considered as an appearance clue that can be leveraged to guide the process of action recognition in videos. In this context, Wang et al. [93] proposed to infer the best poses for each frame by extracting spatial-part-sets and temporal-part-sets using a contrast mining algorithm [97], where the output is then fed to a SVM classifier in order to recognize human actions in videos. Nie et al. [86] proposed a similar approach based on a spatial-temporal And-Or graph hierarchical model that decomposes human actions into three levels including poses, spatio-temporal parts and body joints. Later, Zolfaghari et al. [85] integrate poses, motions, and raw images into a three-stream architecture to improve the action recognition performance. Recently, Choutas et al. [33] proposed an approach called PoTion that jointly encodes appearance and motion of semantic keypoints into a clip-level representation serving as input for a shallow CNN.

4.2.3 Collective Activity Recognition

The past few years have witnessed an increasing interest by the research community about collective activity recognition [24], [98], [99], [100], [101]. A notable work was introduced by Choi et al. [1] where they described the activity of a person based on spatio-temporal descriptors in order to infer the high-level collective activity. Recently, multiple deep learning based models have been proposed in this context. A two-stage hierarchical temporal model was introduced by Ibrahim et al. [24] to recognize collective activities. In the first stage, they analyze the temporal dynamics of each person with an LSTM network. Then, they aggregate this information in the second stage with the encoded temporal group dynamic, in order to learn interactions between people that contributes to recognize collective activities. On the other hand, Deng et al. [102] proposed a framework that combines graphical models and deep neural networks. In this model, nodes are representing both people and the scene, which

allows message passing between outputs. Later, StagNet was proposed by Qi et al. [100] where a semantic graph is used to model individual actions as well as their corresponding spatial relations, whereas the temporal interactions are modeled with a structural-RNN architecture.

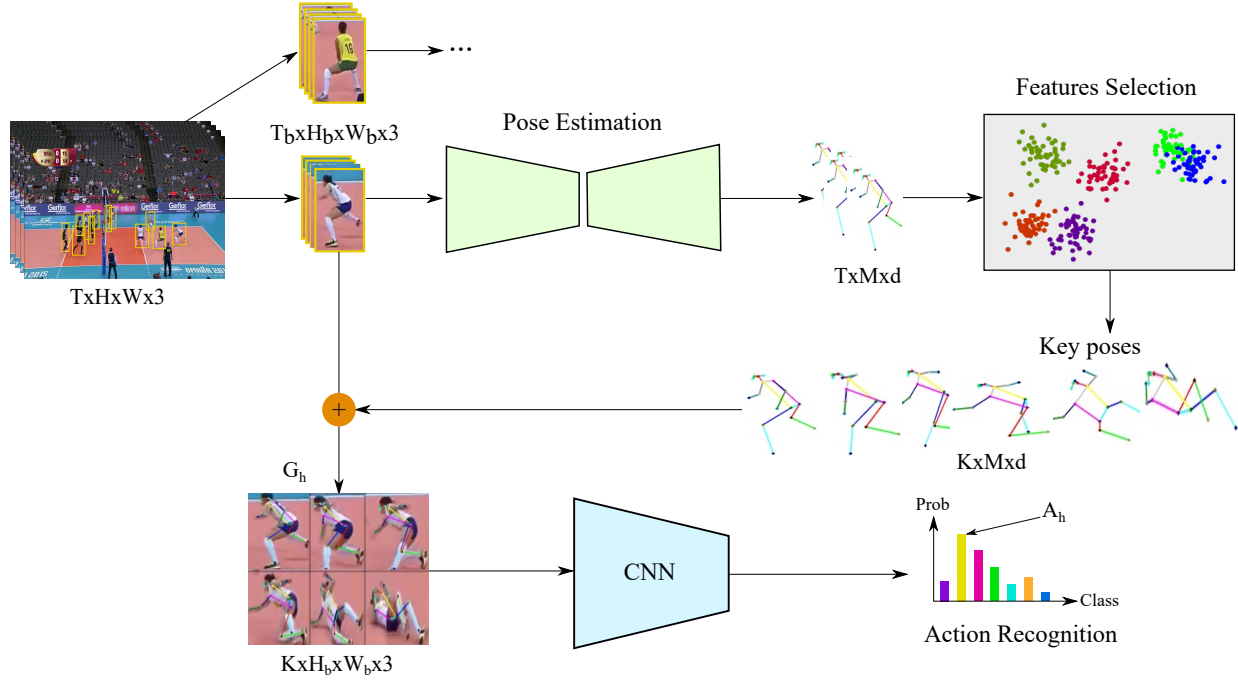


FIGURE 4.1 The pipeline of our proposed GRAR model. From each frame, we extract human poses using a pose estimation method. We select then the most representative key poses for each human based on an unsupervised clustering method. These estimated key poses are combined with their corresponding RGB sources and then concatenated. This results in a compact grid representation G_h , that encodes the relevant RGB and pose information related to the performed action. Finally, we train a deep convolutional neural network on the obtained grid representations to predict the corresponding action category A_h .

Different from these methods, we propose a novel action recognition approach that relies mainly on relevant poses to build an action representation instead of considering random features or frames. Additionally, with an attention guided by key poses, our approach is more robust to occlusion and intra-class appearance variation problems. Following a bottom-up design, our approach can be successfully leveraged to recognize activities based on individual actions.

4.3 Grid-based Representation For Human Action Recognition

In this section, we present the overall design of our novel GRAR model, whose ultimate goal is to recognize efficiently the actions of all persons in a video sequence, by combining

important temporal features associated with poses. The overall pipeline is illustrated in Fig. 5.2.

To obtain a more relevant representation of the performed action, first of all, we select the human region in the scene in each frame, instead of taking the entire frame. This restricts the analysis to focus only on the most relevant information related to the person’s action. Given the tracks of every person, we extract the sequence of the 2D human poses at each time step. Once the poses are extracted, we normalize the joint coordinates with respect to the bounding box position at each time t to get their corresponding relative positions. Based on these normalized joint positions, we select the most representative poses of each action using an unsupervised clustering algorithm. This will allow us to keep only the information of interest about the performed action. We will refer to these representative poses as key poses in the rest of the paper.

From the chosen key poses, we create our new static grid representation that integrates only the most discriminative temporal RGB and pose information, which allows our method to deal with periodicity in actions as well as occlusions and intra-class variations problems. Indeed, since we are using key poses, frames where the person is too much occluded will likely be ignored because they will result in unstructured or infrequent poses. Other times, poses can allow to extract information through occlusions (see Fig. 4.2). Finally, we recognize the performed actions by training a convolutional neural network on sets of these grid images. Because actions are represented in a grid image, we can benefit from pre-trained image classification networks.

In the following sections, we describe in details the main components of our proposed model. Since we are interested mainly in human action recognition in this work, we assume that humans are already detected.

4.3.1 Human Pose Estimation

Human actions are highly correlated with their corresponding poses. A 2D human pose is a continuous representation of the body parts in the image space. Let $V_h = \{V_1, V_2, \dots, V_t, \dots, V_T\} \in \mathbb{R}^{T \times M \times d}$ denote the pose sequence for a human h in T frames, where d is the spatial dimensions and M is the total number of human body keypoints (joints). The pose vector V_t is defined as $V_t = \{w_{t1}, w_{t2}, \dots, w_{tm}, \dots, w_{tM}\}$ where w_{tm} is the spatial coordinate of the keypoint m in a single frame at time t . In this work, we use the recently published HighResolution Net (HRNet) pose estimation model [65] for its good performance on small persons. HRNet is a top-down architecture based on repeated multi-scale fusions between parallel multi-resolution sub-networks. For computational purposes, in our proposed method, we use the HRNet-W32

network trained using the *MSE* loss function defined as :

$$\mathcal{L}_{mse} = \frac{1}{M} \sum_{m=1}^M \|C_m - \hat{C}_m\|_2^2 \quad (4.1)$$

where $\hat{C}_m$ and C_m are the predicted and the ground truth confidence map for the m^{th} joint respectively.

Given a sequence of video frames $S \in \mathbb{R}^{T \times H \times W \times 3}$, the model predicts the pose sequence V_h for a human h based on the bounding box sequence $U_h \in \mathbb{R}^{T_h \times H_h \times W_h \times 3}$. As we use the COCO keypoints format [103] for the 2D pose estimation, we consider $M = 17$ and $d = 2$. In some special cases, object detection and tracking methods can fail to estimate good quality bounding boxes. In the case of humans, this corresponds to a bounding box that does not cover all the body joints of the person of interest. To solve this issue, once we get the pose keypoint coordinates, we apply a bounding box refinement process, which consists in modifying the bounding box coordinates $U'_h \in \mathbb{R}^{T_h \times H'_h \times W'_h \times 3}$ to enclose the extreme joints coordinates.

4.3.2 Relevant Features Selection

Frame selection is a challenging task in action recognition in RGB videos. Considering all the frames or choosing some frames randomly to represent an action induce redundant or irrelevant information in the learning process, which is directly reflected on the final classification accuracy. In order to reduce the complexity of our proposed model and increase its generalization ability, we propose to focus on the most relevant information by extracting key poses, a subset of distinctive human poses for each action. For example, the distinctive poses for the action "running" can match frames where the right hand, and both the left knee and foot are all heading forward, in opposite directions to the left hand and the right knee and foot. At first, we make the keypoint sequence V_h invariant to the position in the scene and the scale of the person of interest. Based on the corrected bounding box sequence U'_h , we transform the coordinates of each vector V_t from the frame t coordinate space to the corresponding bounding box coordinate space U'_t . Then, we normalize the obtained coordinates with respect to the dimensions of the considered bounding boxes.

Now that we constructed the $V'_h \in \mathbb{R}^{T \times M \times d}$ sequence for each person in the same normalized coordinate space, we proceed to cluster these keypoint sequences in order to extract the most discriminative poses, that are the key poses. To this end, we employ the well-known Partitioning Around Medoids (PAM) clustering algorithm [104] based on a pairwise dissimi-

larity metric. This method is shown to be more robust to outliers than the sum of squared Euclidean distances used by K-Means. This fact is also demonstrated by our experiments in Section 4.4.

Formally, given an action represented by the V_h' pose sequence, PAM provides us with a set of K pose clusters $(C_1, C_2 \dots C_K)$ along with their reference pose medoids V_h^* , where $V_h^* = \{V_{k_1}, V_{k_2}, \dots, V_{k_K}\} \subset V_h'$ and $V_h^* \in \mathbb{R}^{K \times M \times d}$ (i.e., the most centrally located pose in a cluster), where both the intra-cluster poses similarity and the inter-cluster dissimilarity are maximized. The learning process of extracting key poses for each person performing a specific action is done as follows : at first, we randomly select a set of K poses for each person, then we assign the remaining poses to the cluster with the most similar medoid to them. After that, we select an arbitrary non medoid pose x and compute the cost of swapping the initial medoid with the new candidate medoid. The updated total cost is based on the L1 norm and is defined as :

$$V_{k_1}, V_{k_2}, \dots, V_{k_K} = \underset{i=1}{\operatorname{argmin}} \sum_{i=1}^K \sum_{x \in C_i} \|x - V_{k_i}\|_1 \quad (4.2)$$

After convergence, the obtained poses medoids serve as the key poses, which will be used as the input feature vector for the next module.

4.3.3 Grid Representation Learning

Several research works in the literature have explored multiple fusion techniques to integrate temporal information [10] and complementary modalities [13] in order to improve the final action recognition accuracy. However, these methods remain limited in the way they integrate information in time. In addition, they rely on randomly selected information which results in less representative features. Different from these methods, our GRAR model is based only on relevant RGB information. Without requiring any additional annotations (i.e, pose, skeleton data), the estimated key poses are used as an explicit attention mechanism. By fusing temporal RGB and pose features into a grid image representation, our model efficiently encodes key patterns needed to recognize human actions.

In order to create our new grid structure, we proceed as follows : For each key pose in V_h^* , we get the RGB information of interest I_{h_k} based on its corresponding bounding box. In an explicit manner, we put attention on the key poses by modifying the pixels $p \subset I_{h_k}$ if $p \in v_{h_k}$ so that the pose vector is encoded directly in the selected relevant RGB region. This attention-guided by pose technique allows us to deal with challenging real-world situations such as intra-class variations and occlusions. Fig. 4.2 (b) shows an example where our model can successfully compensate missing information caused by occlusions.

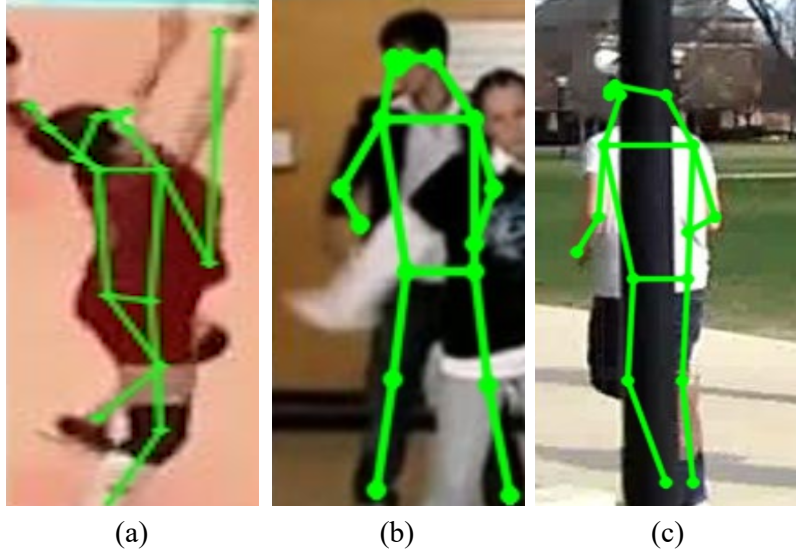


FIGURE 4.2 Examples from the Volleyball and CAE datasets of (a) a pose estimation failure case considered as outlier during key poses selection. (b) and (c) show occluded humans handled with our explicit attention on pose.

Afterwards, we put each fused I_{h_k} in a separated cell forming our main grid structure G_h (see Fig. 4.3 for an example of a grid image). Before grouping the obtained cells, we concatenate each cell with a zeros-valued border. Such operation is necessary to avoid learning unnecessary patterns created by adjacent cells, when the convolution kernel sizes are larger than the inter-cells spacing. Our experiments show that considering a 3 pixel-wide boundary is enough according to the filter sizes of the adopted CNN architecture. It is to note that all of the I_{h_k} cells are not re-scaled, keeping them at their original resolution allows our model to learn jointly the features, as well as their corresponding distance from the camera.

Once the grids are created, we train a CNN to learn human action patterns from these sets of grids. We adopt the Inception-ResNet-v2 model [105] pre-trained on the ImageNet dataset [106] as our backbone CNN architecture. This model is a hybrid Inception network that uses residual connections rather than filter concatenation. We use the categorical cross entropy as our loss function which is given by :

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{w_{y_i}^T G_i + b_{y_i}}}{\sum_{j=1}^n e^{w_j^T G_i + b_j}} \quad (4.3)$$

where G_i refers to the i^{th} grid representation, N is the number of training grids, y_i is the class label of G_i , W is the learned weight matrix and b is the intercept.

Considering only the human region in the scene instead of the whole frame information,

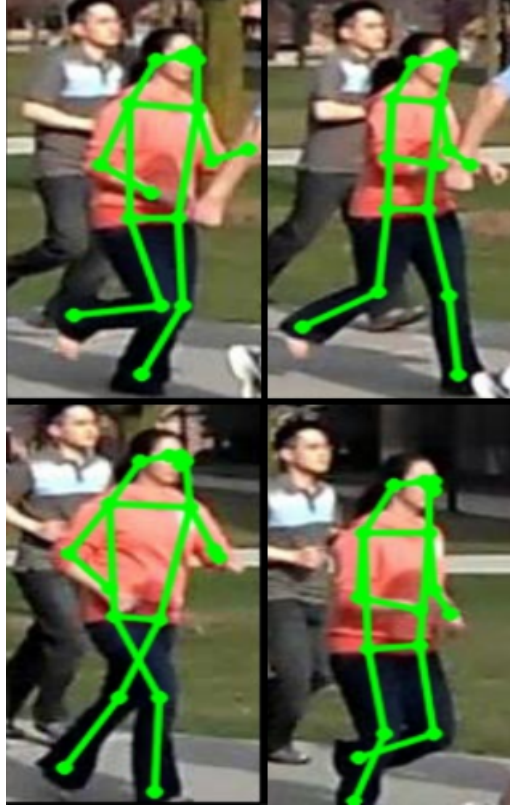


FIGURE 4.3 Example of a grid image from the CAE dataset.

allows us to combine a multitude of cues within the same image representation, without reducing excessively the original image resolution. Excessive downsampling of the image resolution to fall within the CNN input size, often results in losing information valuable for action recognition. Moreover, being based solely on the human region with pose information maximizes the information closely related to the performed actions, and thus allows our model to generalize well to other scenes. For example, in case of a walking action, if our model is trained only with samples of humans walking on a sidewalk, in the inference phase it will be able to recognize that a human walks even if it is on the grass. This can be explained by the fact that the extracted deep features are more related to the human itself than to his environment.

4.4 Experiments

In this section, we evaluate the performance of our proposed model. We first present the considered benchmark datasets and the implementation details. Then we report the results of a series of ablation studies to analyze the impact of each component of our GRAR model

on the recognition performance, followed by a comparison with the state-of-the-art.

4.4.1 Datasets

We evaluate our model on three publicly available datasets : the Collective Activity dataset [1], the Collective Activity Extended dataset [2] and the Volleyball dataset [24].

Collective Activity dataset (CA) [1]

The Collective Activity dataset is a popular dataset for both individual action and group activity recognition. It contains 44 video sequences with a resolution of 640×480 pixels from 5 individual actions and group activity categories (talking, crossing, queuing, waiting, walking). The collective activity label of a scene is defined based on the performed action of the majority of individuals. For train/test split, we follow the same evaluation protocol suggested in [1].

Collective Activity Extended dataset (CAE) [2]

The Collective Activity Extended dataset is an extended version of the original Collective Activity dataset where the "walking" activity was replaced by two new activities "jogging" and "dancing". The reason why "walking" was removed is because in some scenarios this activity is mixed with the Crossing activity. To train our model, we followed the testing scheme mentioned in [102] and use $\frac{2}{3}$ of the videos for training and the rest for testing.

Volleyball dataset [24]

The Volleyball dataset contains 4830 frames collected from 55 YouTube videos and are all about Volleyball games. Each player is labeled with one of these actions : moving, spiking, waiting, blocking, jumping, setting, falling, digging and standing. We adopted the same testing setup used in [24], where $\frac{2}{3}$ of the data is used for training and $\frac{1}{3}$ for testing.

4.4.2 Implementation Details

We implemented our proposed model using the TensorFlow library [107]. We use Inception-ResNet-v2 [105] as our backbone CNN architecture pre-trained on the ImageNet dataset [106]. This network consists in 164 layers, with an image input size of 299×299 . For all the experiments and datasets, we utilized stochastic gradient descent with ADAM [108] and set the optimizer hyperparameters to $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 0.001$. For the CA and CAE datasets, we used 4 key poses and trained the model in 100 epochs with a minibatch size of

16 and an initial learning rate starting from 10^{-3} then reduced with a factor of 0.2 after 10 patience epochs. For the Volleyball dataset, we used 6 key poses and trained the network with a learning rate of 10^{-5} for 130 epochs with a mini batch size of 32.

To track humans in the scene, for the Volleyball dataset, we used the tracker proposed by Cao et al. [109], which is implemented in the Dlib library [110]. For the CA and CAE datasets, we used the tracklets provided by [1]. We used the HighResolution Net (HRNet) algorithm [65] to compute human postures across frames. Specifically, we used the pose-hrnet-w32 architecture trained on COCO dataset. All our experiments were run on a single TITAN Xp NVIDIA GPU.

4.4.3 Ablation Studies

In order to explore the effect of every component of our model on the performance, we conducted extensive ablation studies on the CAE dataset with the following variants.

Clustering Analysis We start by studying the impact of different clustering settings on the recognition performance. The goal here is to find key poses forming the discriminative grid representation. We compare three clustering methods, namely, PAM, K-means and Gaussian Mixture Model (GMM) estimated with the Expectation-Maximization algorithm. For action representation learning, we used the same CNN architecture and parameters settings. The results are reported in table 4.1.

TABLEAU 4.1 Impact of the clustering algorithms and number of key poses on the performance of our model on the CAE dataset. **Boldface : Best result.**

	K-means	PAM	GMM
K = 2	94.0%	94.6%	90.3%
K = 4	94.5%	95.2%	91.7%
K = 6	92.6%	93.9%	91.9%

Compared with K-means and GMM, we can see that PAM gives the highest accuracy 95.2%. This can be explained by the fact that PAM is robust against outliers. In fact, it minimizes the average dissimilarity of human poses in each cluster, rather than minimizing the squared sum of each intra-cluster as adopted by K-means. It is important to highlight that the poses that we use in our model are not manually labeled but instead, are predicted with the HRNet model that can fail sometimes to estimate high quality poses. Fig. 4.2 (a) illustrates an example of a failure case in pose estimation, which is considered in our study as an outlier pose. Compared to PAM and K-means, GMM achieves the worst recognition rate in this

experiment. This can be explained by the fact that the nature of our input pose data is not normally distributed.

Additionally, we evaluated the effect of the number of key poses on the recognition performance. In this experiment, we consider several numbers of key poses, starting from two to six. Intuitively, one can say that the more key poses we consider, the more we gain information about the performed action. However, our evaluation on the datasets demonstrates that above a certain number of key poses, the recognition accuracy starts to decrease. We found that the performance is directly related to the input size of the backbone CNN model that we are using. In fact, by downsampling images to fit the input size of the CNN, some important cues for the recognition task are usually lost. So far, our findings demonstrate that larger downsampling rate makes performance poorer. After extensive experiments, we concluded that the number of key poses must be chosen based on the average size of humans in the scene. Furthermore, to keep the original full resolution for each key pose image, our grid must have a resolution that is close enough to the default CNN input size.

Features selection Next, we study the importance of features selection for human action recognition. We compared random selection against pose-based selection for grid representation. For pose-based selection, we also compared two strategies, using only the poses features (K-Pose, in this case, key poses are drawn in white over a black background) and using only the RGB image corresponding to a key pose (K-RGB). For action representation learning, we employed the same CNN architecture in the three experiments. The results are reported in table 4.2.

Compared with random selection, choosing RGB information based on human pose (K-RGB) gives an improvement of 3.1%. This demonstrates that using the pose estimation to select the RGB data provides us with relevant and discriminative information for human action recognition. On the other hand, using only pose features (K-Pose) with a standard CNN model do not yield good results. Such experience reveals that despite the pose containing valuable cues about the performed action, using it solely does not provide the standard 2D CNN with enough significant features for recognition.

Bounding Boxes Enhancement (EB) Here, we evaluate the effectiveness of making use of the human pose to correct inaccurate bounding boxes used in GRAR pipeline. As illustrated in Table 4.2 at the K-RGB and K-RGB+EB rows, we can see that correcting the human bounding boxes give us 0.6% of improvement. This emphasizes the importance of incorporating human joint features to enhance bounding boxes quality, which is useful not

TABLEAU 4.2 Impact of different modules on the accuracy of GRAR based on CAE dataset. **Boldface : Best result.**

Model Variants	Accuracy
Random	89.2%
Key poses only (K-Pose)	80.5%
Key Frame (K-RGB)	92.3%
Key Frame+Box enhancement (K-RGB+EB)	92.9%
Key Frame+Box enhancement+Pose Attention (K-RGB+EB+PA)	95.2%

only for the action recognition task, but could be useful to other computer vision problems.

Pose Attention (PA) Finally, we study the impact of the introduced pose attention technique, where a key pose is drawn over the corresponding RGB image. We compared the recognition performance of K-RGB+EB against using K-RGB+EB combined with pose-based attention. In both experiments, we considered the enhanced version of the bounding boxes (EB). As indicated in Table 4.2 at the K-RGB+EB and K-RGB+EB+PA rows, we can conclude that putting explicit pose-attention on appearance representation improves the recognition performance by around 2.3%. This indicates that if the estimated pose is of good quality, an attention mechanism based on it makes the model more robust against intra-class variations and occlusions problems, as explained in Fig. 4.2.

4.4.4 Comparison with the State-of-the-Art

In this section, we compare the performance of our GRAR model with respect to several state-of-the-art methods including Learning context [2], Social Cues for activity recognition [111], Hierarchical Deep Temporal Model [24], Structure Inference Machines [102], CERN [98], StagNet [100], Fast collective activity [112], Gaim [113], ARG [114], SSU [99] and SRNN [101].

Results on the Collective Activity dataset

Now that we evaluated multiple variants of our model for individual action recognition, our goal here is to explore the ability of our model to recognize collective activities based on the individual ones. As previously done, we derive the collective activity label of a scene based on the performed action of the majority of individuals. Moreover, we do not use any ground truth annotations about the pose.

Table 4.3 summarizes the state-of-the-art performance on the Collective Activity dataset (CA). Our model with the pose-based grid representation outperforms the compared state-

of-the-art methods. For example, our model achieves $\approx 10\%$ higher accuracy than recent methods based on hierarchical relational networks [24] and recurrent neural networks for activity recognition [102]. This is mostly because we focus primarily on highly discriminative RGB features along with their corresponding poses.

Results on the Collective Activity Extended dataset

The experimental results of human activity recognition on the CAE dataset are shown in Table 4.4. Our pose-based grid model again achieves the state-of-the-art performance with 97.4% collective activity recognition accuracy. This performance demonstrates the effectiveness of choosing relevant RGB information and incorporating key pose features as an explicit attention mechanism in compensating the model weakness when facing ambiguous RGB appearances.

Results on the Volleyball dataset

We further conducted experiments on the Volleyball dataset. Table 4.5 shows the comparison of our proposed model with different recent state-of-the-art methods for individual action recognition. As can be seen, our GRAR model outperforms most of the state-of-the-art methods [24, 98–101] with an accuracy of 82.9%. It is also highly competitive to the ARG method [114], mostly because this latter uses a graph convolutional network that encodes complex actor relation.

4.5 Conclusion

In this paper, we have presented GRAR, a novel pose-based model for human action recognition that uses a grid image of key poses. Our results consistently demonstrate that selecting RGB appearance based on the most discriminative human poses and combining them together in an image leads to considerable improvements. We obtained promising results compared to state-of-the-art approaches on three public benchmark datasets. Our proposed method has several benefits : 1) it is compact, 2) it exploits powerful CNN architectures designed for image classification tasks without requiring any architectural changes, and 3) it is robust against occlusions, intra-class action variations and incorrect human poses estimation.

TABLEAU 4.3 Comparison of the activity recognition performance of state-of-the-art methods versus our model evaluated on CA dataset. **Boldface : Best result.**

Method	Accuracy
Choi et al. [2]	70.9%
Tran et al. [111]	78.7%
Ibrahim et al. [24]	81.5%
Deng et al. [102]	81.2%
Shu et al. [98]	87.2%
Qi et al. [100]	89.1%
Zhang et al. [112]	83.8%
Lu et al. [113]	90.6%
Wu et al. [114]	91.0%
GRAR (Ours)	91.5%

TABLEAU 4.4 Comparison of the activity recognition performance of state-of-the-art methods versus our model evaluated on CAE dataset. **Boldface : Best result.**

Method	Accuracy
Choi et al. [2]	82.0%
Tran et al. [111]	80.7%
Ibrahim et al. [24]	94.2%
Deng et al. [102]	90.2%
Qi et al. [100]	89.7%
Lu et al. [113]	91.2%
Zhang et al. [112]	96.2%
GRAR (Ours)	97.4%

TABLEAU 4.5 Evaluation of action recognition performance of state-of-the-art methods versus our proposed model on the Volleyball dataset. **Boldface : Best result.**

Method	Accuracy
Ibrahim et al. [24]	75.9%
Shu et al. [98]	69.0%
Bagautdinov et al. [99]	82.4%
Qi et al. [100]	81.9%
Biswas et al. [101]	76.6%
Wu et al. [114]	83.1%
GRAR (Ours)	82.9%

CHAPITRE 5 ARTICLE 2 : ACTOR-DRIVEN POSE EMBEDDINGS FOR VIDEO ACTION RECOGNITION

Lamghari, S., Bilodeau, G. A., Saunier, N.

18th IEEE Workshop on Perception Beyond the Visible Spectrum
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
Workshops, August 23, 2022, pp. 399-408.
doi : 10.1109/CVPRW56347.2022.00055.

Abstract

Human action recognition (HAR) in videos is one of the core tasks of video understanding. Based on video sequences, the goal is to recognize actions performed by humans. While HAR has received much attention in the visible spectrum, action recognition in infrared videos is little studied. Accurate recognition of human actions in the infrared domain is a highly challenging task because of the redundant and indistinguishable texture features present in the sequence. Furthermore, in some cases, challenges arise from the irrelevant information induced by the presence of multiple active persons not contributing to the actual action of interest. Therefore, most existing methods consider a standard paradigm that does not take into account these challenges, which is in some part due to the ambiguous definition of the recognition task in some cases. In this paper, we propose a new method that simultaneously learns to recognize efficiently human actions in the infrared spectrum, while automatically identifying the key-actors performing the action without using any prior knowledge or explicit annotations. Our method is composed of three stages. In the first stage, optical flow-based key-actor identification is performed. Then for each key-actor, we estimate key-poses that will guide the frame selection process. A scale-invariant encoding process along with embedded pose filtering are performed in order to enhance the quality of action representations. Experimental results on InfAR dataset show that our proposed model achieves promising recognition performance and learns useful action representations.

5.1 Introduction

Human Action Recognition (HAR) is a fundamental research problem in computer vision that aims to categorize human actions. This task has seen a lot of advances in recent years, making it relevant to a wide range of applications, such as surveillance and security, human robot

interaction, autonomous vehicles, and urban planning [89,90,115]. Human action recognition is a very complex task due to several challenges such as intra-class variations, viewpoint variations, motion velocity variations, background clutter, and human body occlusions.

Thanks to the advances in deep learning, several methods exploiting convolutional neural networks (CNN) for human action recognition in the visible domain have been proposed [82, 83, 91, 116–118]. Compared to classical approaches, such as dense trajectories [119], the two-stream CNN architecture [82] has improved action recognition performance by fusing the output of separate spatial and temporal networks. The first stream network learns spatial salient appearance patterns and the second stream network learns temporal motion features. Wang et al. [83] demonstrated that dividing each video sequence into multiple segments, each processed with a two-stream network, and aggregating their classification scores yields even better recognition results. To enhance human action recognition performances, several other works [32, 35, 85, 120–122] have investigated the integration of human pose sequences or skeletal data information as they provide robust and discriminative representations of actions.

Despite the success of current state-of-the-art methods on visible videos, they still show limited performance when applied in certain real-world situations with low-light conditions such as fog, smoke, or even dark environments. In such scenes, infrared videos are more suitable as they are resistant to background clutter, and less sensitive to lighting conditions. Few works have addressed action recognition in infrared videos [25, 123, 124]. Gao et al. [25] proposed a two-stream CNN framework for infrared action recognition. Jiang et al. [124] passed infrared data and optical flow volumes to a two-stream 3D CNN framework and integrated a discriminative code layer to generate class-based representations.

While these methods have mainly focused on incorporating richer information (e.g., original images, optical flow, motion-history images of optical flow (OF-MHI) [5]) by increasing the number of input streams, they are not robust enough to provide discriminative action representations. This is mainly due to three reasons. First, they do not differentiate multi-person videos from single-person videos when recognizing actions, and learn global scene-level features without focusing on key-actors performing the main action. In fact, videos captured in realistic environments usually contain multiple persons interacting with each other, but only a subset of them at a given time is involved in the main action, as exemplified by the infrared action recognition dataset, InfAR [25]. For instance, a "skip" action in a scene is defined by one human (see Figure 5.1). Thus, integrating spatial and temporal information about other people not involved in the main action, can be uninformative and misleading and prevent the model from learning specific cues to understand such an action. Second, they

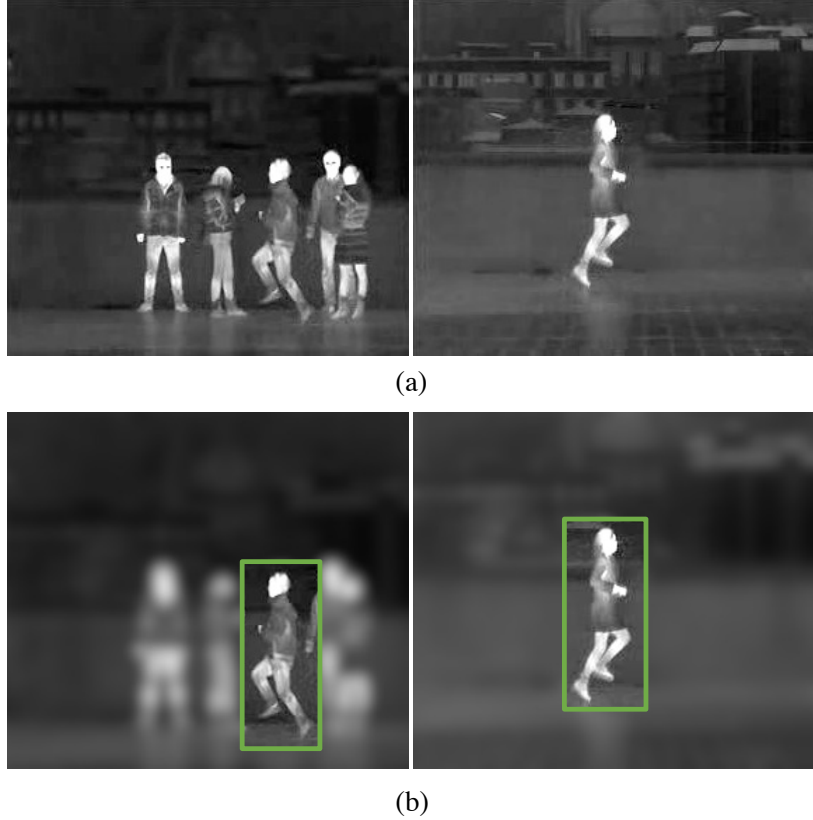


FIGURE 5.1 Example of two actions involving single and multiple persons from the InfAR dataset labeled as "skip". (a) Considering all people in a scene to recognize an action can be uninformative. (b) Identifying the key-actor performing the main action provides accurate guidance for the action recognition model.

learn redundant or random features instead of focusing only on relevant frames. Third, these models often consider the stream networks to be independent and do not allow information sharing between them. They encode only the motion features in short time windows without guaranteeing the preservation of the discriminative cues with pooling techniques.

To address the above shortcomings, we propose Actor-Driven Pose Embeddings for Video Action Recognition (ActAR), a novel model for actor-specific action modeling and recognition in the infrared spectrum. Our model operates in a three-stage fashion. First, we identify key-persons in a video performing an action based on optical flow, and we ignore the others that do not contribute to the main action. This allows our model to focus only on relevant scene information that characterize informative human actions. In the second stage, we extract body pose information for each key-actor. By considering pose information, we provide rich cues to compensate for the missing texture and color information in the infrared spectrum. As most pose estimation models are trained in the visible spectrum, they can fail in some

cases when applied directly to the infrared domain. To solve this problem, we propose a pose embedding filtering mechanism that can leverage complementary discriminative capabilities from large-scale RGB datasets. From the filtered pose sequences, we extract key-poses to restrict the analysis to only the relevant temporal action features. Last, we integrate the most representative features into a compact representation that encodes both the infrared and pose information to obtain a more relevant representation of the performed action. We tested our ActAR approach on the InfAR dataset and found that it achieves good improvement on human action recognition tasks.

Our contributions can be summarized as follows :

- We introduce a unified model for scene understanding by simultaneously addressing two tasks in a single framework : main actor identification and human action recognition.
- Our method operates on raw infrared video sequences and identifies key body pose features for robust and discriminative action representations.
- Extensive experiments on the challenging InfAR dataset demonstrate the effectiveness of our proposed model compared to the state-of-the-art.

5.2 Related work

In this section, we first review existing works on human action recognition in videos, then discuss the importance of pose features for human action recognition.

5.2.1 Human Action Recognition in Videos

Human action recognition in videos aims to classify videos into pre-defined sets of actions. Research methodologies in this field can be subdivided into two main categories : handcrafted features-based and deep learning-based approaches. Early models extracted handcrafted features with a series of operations, such as translations, rotations and computing the trajectories of points to solve HAR related problems. In [47], a viewpoint-free representation for human action from view-invariant motion descriptors was proposed, where only variations in viewpoints related to the central vertical axis of the human body are considered. Later, Wang and Schmid [56] proposed to estimate camera motion to improve the dense trajectories video representation for human action recognition. Cov3DJ [125] uses a covariance matrix of skeletal joint movements as descriptors combined with a classification algorithm to recognize human actions. Despite their success, handcrafted methods capture only local features, making them less discriminative to correctly recognize complex human actions.

Deep-learning based approaches are now dominating this field, as their performance is significantly better compared to handcrafted-based approaches for HAR. In this context, multiple research works have proposed to exploit CNN models for action recognition [82, 116, 118]. The work of [82] is notable in this category as it has successfully integrated CNN for human action recognition by introducing a two-stream method. This method relies on RGB and optical flow and models them into a spatial stream that carries the scene and target information in the video, and a temporal stream that focuses on the target and camera motion. Tran et al. [116] proposed C3D, a method that simultaneously models appearance and motion information using 3D ConvNets and successfully demonstrated that 3D ConvNets are more appropriate for spatio-temporal feature learning than 2D ConvNets. Two-Stream Inflated 3D ConvNet (I3D) was later proposed in [118]. It inflates filters and pooling kernels used for deep image classification ConvNets into 3D, leading to natural spatio-temporal feature extraction from videos.

Since RNNs are suitable for sequential data modeling, various RNN-based methods for human action recognition were proposed. For instance, Srivastava et al. [126] train LSTM with self-prediction to learn salient video representations, while Gammulle et al. [127] created a deep fusion framework by learning spatial features from different layers of CNN and then mapping them with temporal features from LSTM. Later, Du et al. [128] integrated an attention mechanism into a recurrent spatial-temporal network to learn context key features for each timestep prediction of RNN. However, RNN-based methods usually suffer from gradient vanishing and nonparallelism issues making them hard to optimize [129].

Compared with HAR in the visible domain, very few works have addressed this task in the infrared spectrum regardless of its great potential in handling low-light conditions [25, 123, 124]. Gao et al. [25] proposed a two-stream CNN framework for infrared action recognition. Jiang et al. [124] passed infrared data and optical flow volumes to a two-stream 3D CNN framework and integrated a discriminative code layer to generate class-based representations. Later, several multi-stream methods were also proposed. For instance, Liu et al. [84] adopted a three-stream framework over local, global, and spatio-temporal features to learn discriminative action representations. While these methods have mainly focused on incorporating extra information (e.g., original images, optical flow, motion-history images of optical flow (OF-MHI) [5]) by increasing the number of input streams, they are still not robust enough to provide discriminative action representations.

5.2.2 Pose-based Action Recognition

Pose-based human action recognition has recently attracted considerable interest in the literature, as the human body joints position and motion are shown to provide discriminative cues for action recognition [94]. Cheron et al. [32] proposed an action descriptor conditioned on motion and appearance CNN features computed for all body parts, while Zolfaghari et al. [85] combine pose, motion and RGB information using a Markov chain model to classify actions. Similarly, a multitask framework with a single architecture was proposed in [130] for pose estimation and human action recognition. Choutas et al. [33] suggested to encode long-term dependencies of pose keypoints motion and combine it with a shallow CNN to classify actions. Asghari et al. [122] later improved upon this by inferring a fixed-length representation based on the location and body joints heatmap. Unlike previous methodologies, this paper presents a different architecture for HAR that extracts richer key-poses representation.

Besides body pose information, skeleton data consisting in joints and bones representation of the human body, was also considered in the context of human action recognition [120,121,131]. Skeletal body joints information can be acquired using specific sensors such as depth cameras. However, the associated cost and availability of such data makes it limited in practice. Thus, such solutions are outside the scope of this paper.

5.3 Proposed method

In this section, we introduce our proposed Actor-Driven Pose Embeddings for Video Action Recognition (ActAR) method. Specifically, we first present our new key-actor identification module, which defines the main actor performing the action. We then introduce the building blocks of ActAR, including human pose estimation, embedded pose filtering, key-pose selection, and human action learning.

5.3.1 Overview

The proposed approach for human action recognition in videos consists of four components, as illustrated in Figure 5.2. First, the main actors in each sequence are identified using motion information. This helps us focus on regions of interest in each frame, rather than analyzing irrelevant features present in each frame. Second, we extract pose features for each identified actor, as human poses are regarded as a very discriminative and distinctive information to recognize actions. Incorrect pose estimations are filtered through a deep pose embedded filtering mechanism to avoid introducing inaccurate information in the learning process. Based on the filtered pose features, we select the most discriminative key-poses

that are sufficient to describe the main posture states of the actor while performing the action. By representing each actor with a set of key-poses, we optimize the learning process by providing only the relevant key features and disregarding the redundant ones. Third, to make the extracted key-poses amenable to the learning space, we embed them in a compact representation that summarizes the distinctive key-poses of each actor along with the infrared information that provides extra context to the keypoint features. Finally, we train a CNN with the constructed compact representations to learn and recognize human actions in video sequences.

5.3.2 Key-Actor Identification

Human action recognition in multi-person videos requires a deep understanding of the scene features. However, in weakly-labeled real-world datasets, scenes can contain multiple persons interacting with each other, but not necessarily all of them are involved in the main action. This makes the task of human action recognition in such situations even more challenging. One solution to this problem would be to focus only on the persons performing the labeled action. Such annotations are expensive to acquire and are not always readily available. Motivated by the need for automated key-actors identification in realistic scenes, we propose a new methodology for actor-specific action recognition that identifies and locates the main person performing the action.

The key-actor identification task requires accurate human body detection and tracking despite all the challenges imposed by realistic videos in the infrared spectrum. To track humans in videos, we apply a multi-object tracker over each frame f of the video sequence S of T frames. We adopted the recent state-of-the-art Bytetrack model [132] to track simultaneously multiple human instances in videos. Given the extracted tracks $[Tr_a, \dots, Tr_A]$, where $Tr_a = [b_{a0}, \dots, b_{aT}]$. The challenge now is to identify the main actor tracks as the other tracks are not informative to the action and may introduce noisy and non-discriminative features for the action recognition model. One robust information that can assist the identification of the main actor is motion feature. At each timestep, and for each track, we extract the optical flow magnitude in the object bounding box $y(Tr_a) = [y_{a0}, \dots, y_{aT}]$. To achieve this, we use the GMA model [133]. This method is able to handle occlusions by aggregating the global motion using a transformer. The aggregated motion features are defined as :

$$\hat{y}_i = y_i + \alpha \sum_{j=1}^N f(\theta(x_i), \phi(x_j)) \sigma(y_i) \quad (5.1)$$

where x denotes the context features, y are the motion features, f refers to a similarity

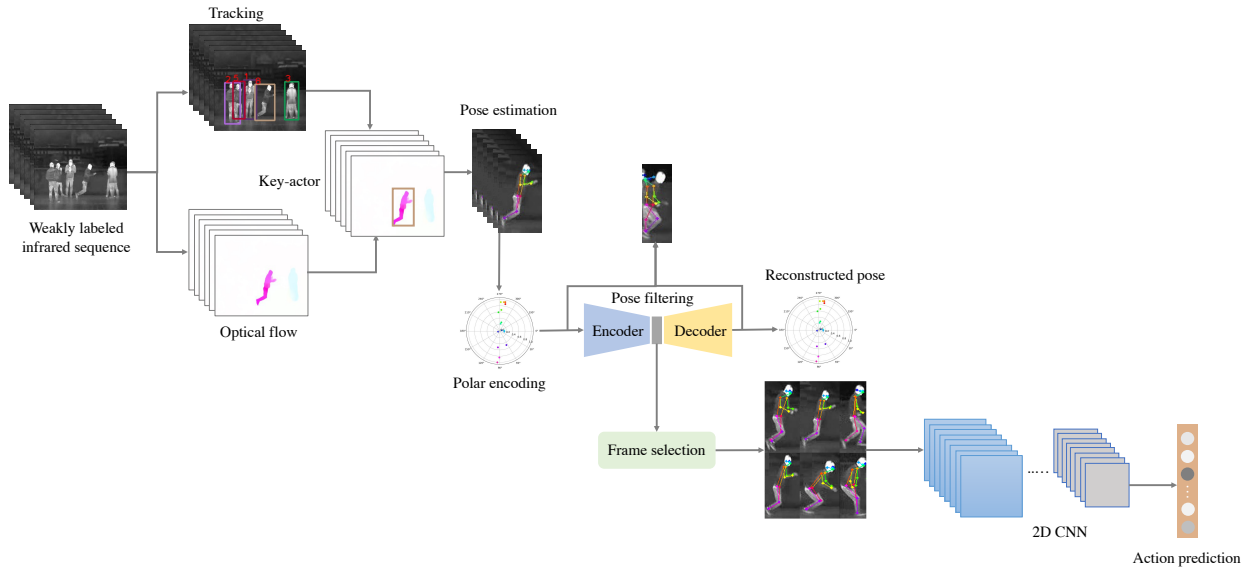


FIGURE 5.2 An overview of the proposed ActAR model. Given an infrared input video, we identify the key-actors based on optical flow. To compensate for the missing texture and color information in the infrared domain, we extract the human body keypoints. We project the estimated keypoints into the polar system, and then we filter out incorrect poses that may mislead the action representation. Next, we select the key-poses based on a deep clustering method to form a compact representation encoding the most relevant infrared and pose information. In the last stage, we train a 2D deep convolutional neural network to predict the human action.

attention function, θ , ϕ and σ are the projection functions for the query, key, and value vectors, respectively, and α is a scalar parameter.

Note that frames with low-score tracking are discarded. To distinguish the main actor, we maximize the motion vector of all actors, and choose the one with strongest motion information, where the key-actor id is defined by :

$$k_{id} = \operatorname{argmax}\left(y(Tr_a), \dots, y(Tr_A)\right) \quad (5.2)$$

Now that we have performed the first precision-level of our model by identifying the temporal tracks of the main actor, we next discuss how to estimate and select discriminative infrared information in the temporal dimension.

5.3.3 Human Pose Estimation

In this work, we mainly guide the task of action recognition from videos by using the evolution of human body poses as a cue. These representations will enable us to deal with the abrupt viewpoint variation and lighting change challenges. These representations are also compact and low-dimensional, which is an essential feature for optimized video analysis. However, for several benchmark datasets, human pose information is not readily available. To address this problem, we estimate 2D human pose features of the identified actors in each video frame using the state-of-the-art human pose detector proposed in [134]. The High-Resolution Net (HRNet) model learns reliable pose estimation of multiple persons in videos using a high-resolution network. It starts with a high-resolution subnetwork, then high-to-low resolution subnetworks are added subsequently, while multiscale fusions are performed repeatedly to allow information sharing between all the parallel multi-resolution subnetworks.

Formally, for a given video sequence $S \in \mathbb{R}^{T \times H \times W \times 3}$ of T frames with width W , and height H , we run the HRNET pose detector on each frame to get human pose keypoints (i.e., body joints). Thus, across all frames, each actor h is represented by a set of K temporal sequence of keypoints denoting the location estimate of each body joint of the main actor at every frame. During the pose estimation process, we optimize the MSE loss function as follows :

$$\mathcal{L}_{mse} = \frac{1}{M} \sum_{m=1}^M \|C_m - \hat{C}_m\|_2^2 \quad (5.3)$$

where C_m and $\hat{C}_m$ refer to the true and the predicted confidence map for the m^{th} body joint, respectively.

This module returns the evolution of each actor body pose across time, which will be used by our model for human action recognition. Next, we encode the extracted pose features to mitigate the effects related to the camera viewpoint changes that may lead to observing the same person with different appearance scales.

5.3.4 Scale-invariant Polar Encoding and Embedded Pose Filtering

The estimated human body poses often contain large variations, such as scale, and bone length ratios [135]. This leads to having the same action represented by different spatial keypoints coordinates and thus can mislead the learning model. To alleviate this, we propose a scale-invariant encoding method to eliminate the influence of such effects. The goal is to let the pose-guided features encode only the human poses related to the action while being invariant to other factors. Specifically, we select the center hip joint as the reference body joint, because it is usually stable in most actions. Then, we calculate the relative positions of all the other joints with respect to the reference joint $\hat{J}^t$:

$$J_m^t = J_m^t - \hat{J}^t \quad (5.4)$$

As the actor body joints are initially defined in the Cartesian coordinates system $(x, y) \in \mathbb{R}^2$, we next project them to the polar coordinate system with a polar angle $\theta \in [0, 2\pi]$. This enables our model to differentiate between keypoint vectors using their angles instead of focusing on their magnitude. Thus, in this new system, the hip center becomes the pole.

In the same context of enhancing the quality of the estimated poses, we have noticed that when using a model trained on RGB data, the predictions on infrared domain may result in incorrect poses in some cases. To overcome this problem, we propose an embedded pose filtering mechanism that leverages latent pose representations from the visible domain to filter poses estimated in the infrared spectrum. To do so, we train a deep autoencoder model to reconstruct ground truth poses from COCO dataset [103]. The trained autoencoder will be also used in the clustering module as we will discuss in the next section. Once the autoencoder has learned the latent pose representations in the RGB domain, the model decides to keep or discard the estimated infrared pose based on the reconstruction error.

5.3.5 Human Pose Features Selection

Since human poses in consecutive frames are often extremely similar, we select individual distinct key-poses that are used as representative poses summarizing the principal actor actions. Focusing on frames with key-poses significantly reduces the complexity and the

computational cost of the human action recognition algorithm. To infer these candidate key-poses, we embed the extracted pose features of every actor across frames into a latent space and cluster the resulting embedded vectors into sets of similar poses. Thus, we adopt a Deep Embedded Clustering method [136] as a base model for our key-pose extraction process. This model consists in a deep autoencoder, followed by a soft clustering layer that assigns each pose sample to a cluster with a specific probability. Specifically, it uses the autoencoder networks to embed input features into a low-dimensional latent space, and infers cluster assignments simultaneously.

Formally, given a polar encoded-keypoints sequence V_h for a target actor, we start by learning unsupervised representations of the input keypoints using a stacked autoencoder (SAE). It has been shown that this type of networks learns discriminative and meaningful latent representations on various datasets [136, 137]. After initializing the network, we remove the decoder to keep just the encoder for deep features extraction. After getting the embedded pose keypoints, we perform standard k-means clustering to obtain the initial cluster centers. And then we optimize the encoder using the following clustering objective :

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (5.5)$$

This objective function calculates the Kullback–Leibler (KL) divergence loss between the auxiliary target distribution p_{ij} and the soft assignments q_{ij} where,

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j q_{ij}^2 / \sum_i q_{ij}} \quad (5.6)$$

and q_{ij} refers to the similarity between embedded pose keypoints V'_h and cluster centroid μ_j measured by the Student's t-distribution [138].

$$q_{ij} = \frac{(1 + \|V'_h - \mu_j\|^2)^{-1}}{\sum_j (1 + \|V'_h - \mu_j\|^2)^{-1}} \quad (5.7)$$

It is to note that minimizing the objective L follows a self-training strategy, as the auxiliary target distribution p_{ij} depends on the soft assignments q_{ij} leading to enhanced clustering predictions.

Finally, to select the actual representative poses of each cluster, we calculate the distance between each pose p_i and the embedded cluster center p_{ic} , and we take the nearest-neighbor pose sample to the cluster center as the representative key-pose for that cluster. We empirically found that the optimal number of clusters is 8, as it provided the best balance between

too few and too many poses clusters. Thus, each actor in a sequence is mainly represented by 8 identified key-poses.

5.3.6 Human Action Representation Learning

To enhance the action recognition performance, various fusion techniques to aggregate temporal information with other modalities have been studied in the literature [34, 82]. However, most of these approaches rely on information that is not always relevant for the human recognition task, and are limited in the way they fuse the temporal information. Unlike these methods, our ActAR model combines actor-specific relevant infrared information with the estimated key-poses into a compact representation. This enables our model to learn key patterns needed to recognize human actions in infrared videos.

To construct the summarized action representation for each key-actor, we follow these steps : we first extract the infrared region of interest I_{hk} that is defined by the bounding box of each key pose in V_h^* . By encoding the pose vector directly for the selected relevant actor region, we compensate for the missing texture and color information in infrared videos.

Next, we aggregate each encoded region I_{hk} into one main grid-like structure for every identified key-actor as illustrated in Figure 5.2. To prevent the model from learning unnecessary action patterns, we combined each encoded region with zeros-valued borders having a 3 pixel width, instead of being completely adjacent, as the the convolution kernel sizes of the adopted CNN architecture can be larger than the inter-region spacing.

After constructing the compact features representation for each key-actor, we train a convolutional neural network to learn the key patterns related to the performed action from the created grid structures. As our backbone CNN architecture, we use the Inception-ResNet-v2 model [105], which has been pre-trained on the ImageNet dataset [106]. The Inception-ResNet-v2 model has a hybrid Inception network architecture that uses residual connections instead of filter concatenation. For the training objective, we adopt the categorical cross entropy loss defined as :

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(w_{y_i}^T G_i + b_{y_i})}{\sum_{j=1}^n \exp(w_j^T G_i + b_j)} \quad (5.8)$$

where G_i denotes the i^{th} grid-like structure, N is the number of training structures, y_i refers to the class label of G_i , W is the learned weight matrix, and b is the intercept.

Actor Identification		Frame Selection		Polar Encoding	Pose Embedding Filtering	Average Precision (%)
All actors	Principal actors	Random	Key frames			
✓		✓				74.44
	✓	✓				79.38
	✓		✓			82.55
	✓		✓	✓		84.26
	✓		✓	✓	✓	85.32

TABLEAU 5.1 Performance contribution of each component in ActAR on InfAR dataset.

5.4 Experiments

After describing the experimental settings, we first present an ablation study and then a comparison with the state-of-the-art.

5.4.1 Experimental Settings

All experiments were performed on a single TITAN Xp GPU. We implemented our proposed model using the TensorFlow library [107]. We use Inception-ResNet-v2 [105] as our CNN backbone pre-trained on the ImageNet dataset [106]. We trained the model using a Stochastic Gradient Descent (SGD) with ADAM [108] and set the optimizer hyperparameters to $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 0.001$. For the key-poses, we used 8 components and trained the network with a learning rate of 10^{-3} for 150 epochs with a mini batch size of 32.

For optical flow estimation we used the GMA method [133] trained on Sintel dataset [139]. For human tracking, we use the ByteTrack model proposed by Zhang et al. [132]. We use the HighResolution Net (HRNet) architecture [65] to compute human poses across frames. Specifically, we used the pose-hrnet-w48 architecture trained on the COCO dataset [103].

We evaluated the proposed Actor-Driven Pose Embeddings model on the Infrared human action recognition dataset, InfAR [25]. Infrared Action Recognition (InfAR) is a challenging video action recognition dataset that provides 600 video clips captured by infrared thermal imaging cameras. Each video clip is annotated with one action from 12 different human action classes. The average video clip length is 4 seconds with a frame rate of 25 and a resolution of 293×256 .

5.4.2 Ablation Study

In this section, we perform an extensive ablation study to investigate the effectiveness of each component of our method on the InfAR dataset. In all ablations, we use average precision as the main evaluation metric. We report the ablation performance of our model in Table 5.1.

We start with a basic model to understand the impact of the actor identification stage. In this experiment, we first consider all persons in the scene as being involved in the main action, and we select random frames for action representation for each person. To infer the scene action classification, we take the most frequent of all person-level actions. With the basic model, we achieve 74.44%. This is expected as the model considers irrelevant information at two levels of learning, namely, the scene and the actor levels.

To avoid incorporating wrong action information induced by persons who do not contribute to the main action in the scene, we integrate in the basic model information solely about key-actors using our proposed optical flow-guided module. Similar to our previous experiment, we randomly select frames to represent the key-actor action. With this model configuration, we outperformed the basic model with a precision of 79.38%. This demonstrates that the same model can get better performance if it is fed with the right features with respect to the key-actor involved in the labeled action.

Next, we upgrade the key-actor model with key-poses-based frames selection. As we previously explained, key-poses are obtained by clustering key-actor body poses using a deep autoencoder model. Clustering performance will be discussed later in this section. Compared with random frames representations, choosing infrared information guided by key-poses to encode action cues provides an improved level of precision in the temporal domain and allows our model to capture more discriminative action features. The obtained results clearly highlight the advantage of incorporating the frame selection module.

We also studied the impact of 2D-keypoint representations on the final action recognition precision. While in the previous experiments, we used actor body joints defined in the Cartesian coordinates system, they are now projected to the polar coordinate system with the center hip as a reference joint since this joint is stable in most actions. Having human poses π -scaled in this new coordinate system significantly benefits the clustering module and enhances the recognition capabilities of our model, as shown in the results. The new keypoints are not only lying on a common space but are also scale-invariant.

As the InfAR dataset is not manually pose-labeled, the keypoints predicted by HRNet can be sometimes incorrect. The integration of all the previous model components with the pose embedding filtering module, i.e., our ActAR model, mitigates the effects of this problem, and thus, reaches the highest action recognition performances. The performance boosts induced by the pose embedding filtering module indicate the advantage of filtering features predicted by trained RGB models in the infrared domain, especially when working with weakly labeled data.

We have seen above that frame selection guided by key-poses clustering can improve the

performance of our model for human action recognition. Now, we want to evaluate the impact of the clustering model for identifying the most informative body poses for each action. We compared three clustering methods : K-means, Gaussian Mixture Model (GMM), and Deep Embedded Clustering (DEC) with different numbers of clusters, varying from 2 to 12. In this experiment, we use the complete version of ActAR with the same parameters settings. We report the clustering ablation results in Figure 5.3. We can see that integrating DEC with ActAR achieves the highest precision. This demonstrates that clustering deep pose features is better than clustering raw keypoints coordinates. On the other hand, ActAR with GMM performs slightly worse than the k-means version in this experiment. This can be explained by the fact that the nature of our infrared input pose data is not naturally normally distributed. Furthermore, we found that the optimal number of key-poses is 8. This indicates that the number of key-poses must be chosen based on the average size of humans in the scene.

5.4.3 Comparison with the State-of-the-Art Methods

We compare the proposed ActAR model with the state-of-the-art methods on the InfAR dataset, which can be grouped into two categories : handcrafted features-based methods (e.g., DT, and iDT) and deep learning-based approaches (e.g., HOF, Two-stream 2D-CNN, Two-stream 3D-CNN, TSTDs, Four-stream CNN, and SCA).

The results of our model and existing methods are reported in Table 5.2 using the average precision evaluation metric for human action predictions. We observe that our method achieves the best action recognition precision. In fact, ActAR surpasses handcrafted features-based methods [56,119] that uses dense trajectories with a considerable margin of 13.49% and deep learning-based methods [25,84,124,140,141] that uses different multi-stream architectures with a margin ranging from 1.07% to 8.66%. Our results support our hypothesis that our main actor-driven ActAR model with discriminative pose-based features is beneficial for the human action recognition task. The results also reveal the impact of the embedded pose filtering module on the recognition performance, especially when extracting features from infrared videos through models that are trained on labeled RGB data.

5.5 Conclusion

In this paper, we have introduced a new actor-driven model for effective human action recognition in infrared videos with single and multiple persons. Extensive evaluation experiments demonstrate the effectiveness of ActAR compared to existing methods, and highlights the impact of each component of the proposed model. The proposed actor identification and frame

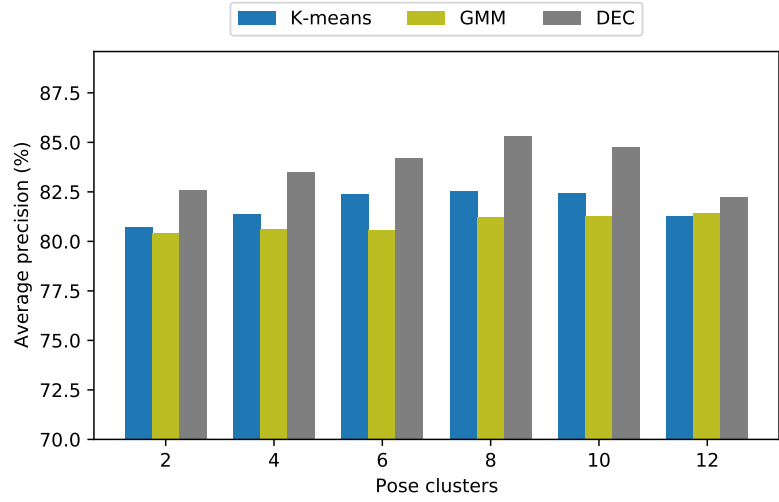


FIGURE 5.3 Impact of different pose clustering algorithms on the performance of ActAR on InfAR dataset.

Model	Average precision(%)
HOF [25]	68.58
DT [119]	68.66
iDT [56]	71.83
Two-stream 2D-CNN [25]	76.66
Two-stream 3D-CNN [124]	77.50
CDFAG [142]	78.55
TSTDDs [84]	79.25
Four-stream CNN [140]	83.50
SCA [141]	84.25
ActAR	85.32

TABLEAU 5.2 Comparison of the average recognition precision of ActAR with state-of-the-art methods on InfAR dataset.

selection modules effectively encode discriminative infrared information with key-actors body poses, and combining them together in a compact representation leads to considerable improvements. This further shows that identifying key-actors responsible for the action is crucial for efficient human action recognition. Therefore, our model can be easily generalized to any multi-person sequence, as it does not require key-person annotations for training. Finally, The embedded pose filtering module also shows that infrared spectrum can benefit from RGB models and data by getting better quality features.

CHAPITRE 6 ARTICLE 3 : TRANSFORMER-BASED HUMAN ACTION RECOGNITION WITH DYNAMIC FEATURE SELECTION

Lamghari, S., Bilodeau, G. A., Saunier, N.
20th Conference on Robots and Vision, June 8, 2023.

Abstract

Human action recognition in videos is an important task of computer vision that aims to automatically recognize and classify human actions in video sequences. However, accurately recognizing human actions can be challenging due to the complexity and variability of human motion and appearance. In this paper, we propose ActiViT, a novel approach for human action recognition in videos based on a Transformer architecture. Unlike existing methods that rely on convolutional or recurrent layers, our model is entirely based on the Transformer encoder, enabling us to leverage valuable information in action image patches features. We demonstrate that by dynamically selecting key patches guided by specific human poses, our model learns informative features useful for distinguishing between different actions. Our experimental results on real-world datasets convincingly demonstrate the effectiveness of our model and the importance of selecting discriminative key poses for action recognition.

6.1 Introduction

Human action recognition (HAR) is an important field of research in computer vision, which consists of detecting and identifying the actions performed by humans in video sequences automatically. HAR has proven to be useful for a wide range of practical applications, including video surveillance [143], sports analysis [24], urban planning [144], etc. Human actions can be classified into two main categories : simple and complex. Simple actions involve one or few atomic sub-actions, while complex ones are composed of multiple sub-actions that are performed in a specific order to achieve a goal. Recognizing both simple and complex actions is crucial for building robust human action recognition systems. However, recognizing individual and collective actions whether being simple or complex of several people in the same video remains a difficult and challenging task due to various factors present in realistic environments, such as varying imaging conditions, the dynamism in backgrounds, diverse motion and appearance of people. These factors can have substantial impact on the recognition performance. Thus, developing effective techniques to overcome these challenges is essential for advancing the field of human action recognition in videos.

In recent years, several deep learning-based HAR approaches have been proposed [82, 83], with many of them relying on Convolutional Neural Networks (CNN) to learn visual deep features of actions, allowing for more precise HAR. However, these approaches are limited in their ability to capture motion information within short time intervals and may not encode the most significant features.

On the other hand, with the significant advancements of Transformer architectures in Natural Language Processing tasks [18], researchers have recently explored their application in computer vision tasks such as image classification [19], where their self-attention mechanism has shown effectiveness in modeling temporal relationships and dependencies in images. However, despite this progress, the potential of Transformer in HAR is not yet thoroughly investigated.

In a different research direction, some researchers have integrated 3D human poses for action recognition to provide a more comprehensive representation of the action being performed in a video sequence [28, 31]. However, these methods are not practical, as they require specialized depth sensors. As an alternative, some approaches use 2D human poses, which are simpler and easy to acquire, to leverage the inherent body posture and gesture information [32, 33]. Although these methods have shown promising performance, they are still limited by their dependency on randomly learned features to represent human actions.

In this paper, we propose ActiViT, a novel Transformer-based approach for human action recognition in videos. Our method effectively models the temporal relationships within actions and dynamically adapts to capture all the key features in a video sequence. Instead of providing the learning model with generic features about the humans and the scenes, we adopt a pure Transformer network that takes in the most discriminative embedded representations of actions that are specific to each person in a video sequence. These representations are derived from the estimated key body poses that better describe the performed actions. By doing this, we significantly reduce the amount of inductive bias compared to using CNN [19]. Furthermore, our non-parametric pose clustering approach enables our model to effectively handle sequences of varying lengths and learn valuable cues from them. Extensive experimental results on three benchmark datasets demonstrate our model capability in improving the performance of human action recognition, particularly in scenarios where actions may be complex or occur in challenging imaging conditions.

Our main contributions can be summarized as follows :

- We present a novel Transformer-based approach for efficient human action recognition that captures key spatial and temporal action representative features ;
- We distill human pose features to identify the most discriminative action image patches

- dynamically, which further enhances the performance of our action recognition model ;
- Our model can effectively handle sequences of varying lengths, making it more versatile and practical for real-world action recognition scenarios ;
- Our experimental results on real-world datasets showcase the superior capabilities of our model in accurately predicting human actions compared to state-of-the-art methods.

6.2 Related Work

Deep learning-based approaches have achieved good performance in recognizing human actions. Popular approaches have proposed to use convolutional neural networks to capture the spatial and temporal information from videos. For instance, Simonyan et al. [82] proposed a two-stream CNN architecture that integrates spatial and temporal networks. The architecture is trained independently on both static appearance inputs and multi-frame dense optical flow inputs. Karpathy et al. [34] utilized two separate streams to speed up the model computation by feeding low-resolution RGB frames and high-resolution center crops to each stream. In the same research direction, Wang et al. [83] proposed TSN, a model based on a two-stream CNN architecture that utilizes sparse temporal sampling to encode the long-range temporal structure for improved dynamic understanding of actions. Other studies have also extended the two-stream network to extract long-term video-level information for human action recognition. To capture spatio-temporal information between consecutive frames in videos, Ji et al. [145] proposed a 3D CNN model that uses 3D convolutions. However, one issue with 3D CNN is that they tend to have a large number of parameters that can make them computationally expensive. Additionally, they may not be well-suited for modeling long-range temporal dependencies in videos.

Since CNN-based methods do not capture effectively long actions evolution and dynamics in time, other methods have proposed to learn action patterns using Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) to benefit from the recurrent connections in their hidden layers. A notable work in this category was proposed by Donahue et al. [91], which uses an 2D CNN network for frame-level feature extraction, which are then processed by an LSTM network to generate a final action label. Srivastava et al. [126] proposed an approach that uses an LSTM encoder to map an input video to a fixed-length representation. This representation was then decoded using an LSTM decoder for unsupervised video reconstruction and prediction.

With the remarkable success that Transformer networks have achieved for many NLP tasks, Vision Transformers have emerged and they have been proven to be superior to CNN in

modeling sequential data. In particular, Vision Transformer has been applied for various computer vision tasks such as image classification and video recognition. The self-attention mechanism in Transformer enables capturing long-range dependencies in the input image or video frame. A notable work by Vaswani et al. [18], proposed an attention-based model Transformer that incorporates a multi-head self-attention layer to learn useful sequence representations. Later, Dosovitskiy et al. [19] proposed ViT, one of the first models that showcased the effectiveness of using a Transformer for image classification. As human actions typically involve long spatio-temporal relationships, recent action recognition methods have attempted to leverage Vision Transformer for that purpose. For instance, Girdhar et al. [20] introduced a Transformer based architecture incorporating a CNN to learn specific features from the spatiotemporal context of the person action being classified.

There have also been other research studies that proposed to integrate various modalities into the HAR learning process, such as depth or skeleton data, to improve action recognition performance. For instance, Zolfaghari et al. [85] proposed a method for action recognition that combined RGB and depth information using a multi-stream CNN architecture. Nie et al. [86] presented a comparable approach based on a hierarchical model of spatial-temporal And-Or graphs, which decomposes human actions into three levels : poses, spatio-temporal parts, and body joints. Choutas et al. [33] introduced an action classification approach that captures the long-term dependencies of motion in pose keypoints and integrates it within a shallow CNN.

In contrast to existing approaches, we propose a human action recognition framework that utilizes the valuable information derived from key pose and that learns high quality action embedding using a pure Transformer encoder, without the need to rely on convolutional or recurrent networks.

6.3 Proposed Method

In this section, we first provide an overview of the proposed method, followed by a description of the specific components of our method including pose estimation, key pose identification, object embedding and the Transformer model for action learning.

6.3.1 Overview

The goal of our method is to recognize human actions in realistic scenes that involve multiple humans by learning highly discriminative features guided by key poses. As illustrated in Figure 6.1, our proposed approach consists of four main stages : human pose estimation,

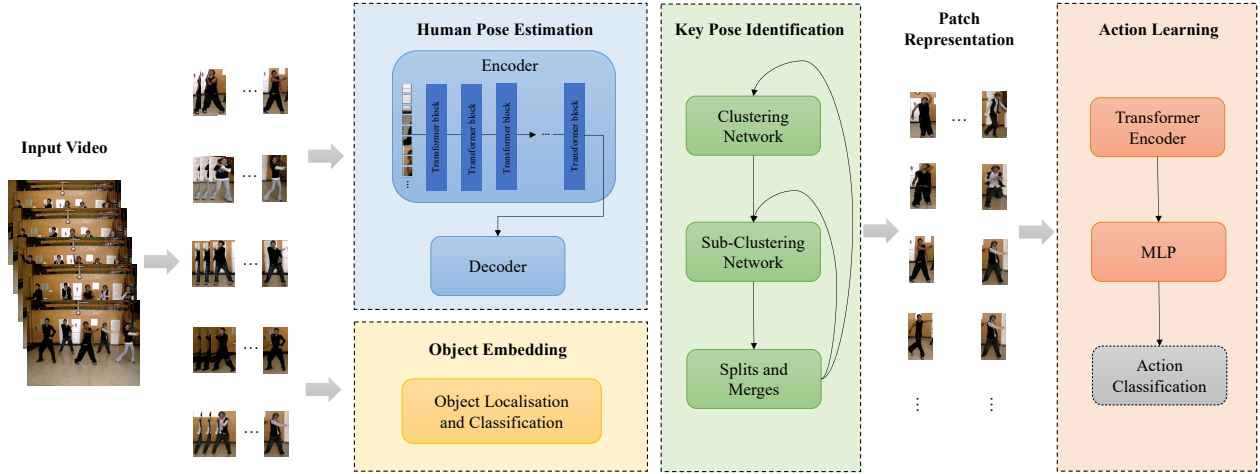


FIGURE 6.1 The pipeline of our proposed ActiViT model. The model comprises four stages. In the first stage, we estimate the pose keypoints of all individuals in each frame of a video sequence. The second stage uses a non-parametric approach to cluster the extracted body poses, enabling dynamic key pose selection. In the third stage, we localize and recognize objects appearing in the scene that individuals interact with. Finally, we feed the resulting key image patches, along with their corresponding objects embeddings, to the transformer model to learn discriminative features for improved action recognition.

key pose identification, object embedding, and Transformer-based action representation learning. In the first stage, we extract the pose keypoints of persons in order to capture the patterns formed by the body joints. The second stage of our approach involves fine-grained and dynamic selection of key poses. This selection process enables our model to focus on the most informative body poses that are essential in distinguishing between different actions. In the third stage, we localize and detect objects in scenes to better capture the contextual information of actions. Finally, we learn the performed action by feeding image patches corresponding to the selected key poses along with their corresponding objects to a Transformer network. By doing so, we exploit the Transformer attention mechanism to further learn and capture the essential features that discriminate human actions.

6.3.2 Human Pose Estimation

Human pose contains valuable information about the body joint movements, which is important for many video analysis tasks, including HAR. Representations that leverage human body pose have demonstrated a considerable potential in enhancing the performance of HAR. Therefore, we focus on estimating body joint keypoints for each person from raw RGB frames. This step is important to improve the ability of our model to encode the spatial and temporal characteristics of human actions.

To extract the body joint keypoints, we use a pose estimator that takes as input bounding boxes of persons and then predicts the position of their body joints. We use a recent model called ViTPose [146], as our method for pose estimation. ViTPose follows the widely adopted top-down approach for human pose estimation. It relies on a plain Vision Transformer and a lightweight decoder for feature map extraction and pose estimation, respectively.

Formally, we assume that an input video sequence contains T frames, each containing a set of N humans $H_f = \{h_1, h_2, \dots, h_N\}$. To extract the body poses of each human h_i , we feed its corresponding bounding box in frame t to the ViTPose model. As a result, we get a set of 2D pose vectors denoted as $P_{h_i} = \{P_{h_i}^1, P_{h_i}^2, P_{h_i}^3, \dots, P_{h_i}^T\}$, where T is the total number of frames in a video sequence. Every $P_{h_i}^t$ is a d -dimensional vector that contains the spatial coordinates of the body pose keypoints of human h_i .

6.3.3 Key Pose Identification

After estimating the body poses represented by a set of keypoint coordinates for each person in a scene, it is very likely that some poses will be irrelevant and redundant. Feeding the action learning network with all of these representations or with a randomly selected subset can have a direct effect on its performances. To solve this problem, it is important to select a subset of body poses that better discriminate the performed action. This also means that the computational complexity of the learning model can be simplified, since fewer input poses will need to be considered. To achieve better action recognition performance, we need to ensure that the selected key poses are truly the ones that capture the most relevant information about the actions. Moreover the number of key poses to consider can vary from one video to another and is highly dependent on the performed action nature and complexity. Thus, it is crucial to consider a scalable pose selection algorithm that adapts automatically to the nature of the poses.

To do so, we propose a dynamic key pose identification mechanism based on deep clustering, that groups the set of estimated poses for each human into different clusters while minimizing the intra-cluster similarity and maximizing the inter-cluster similarity. Specifically, we use DeepDPM [147], a non-parametric deep clustering algorithm that infers the optimal number of clusters during training by splits and merges [148]. The clustering procedure consists of clustering and sub-clustering networks. In the first part, soft clustering assignments are generated for each input body pose using a deep clustering network. Based on the Expectation–Maximization (EM) algorithm in Bayesian Gaussian Mixture Models (GMM), the loss function of the clustering stage is defined as [147] :

$$\mathcal{L}_{cl} = \sum_{i=1}^N KL(r_i || r_i^E), \quad (6.1)$$

where KL refers is the Kullback-Leibler divergence, $r_i \in [0, 1]$ is the soft clustering assignment of pose $P_{h_i}^t$, and R_i^E are the learning probabilities from the Expectation step. In the second part, the soft cluster assignments are transformed into hard cluster assignments by attributing each pose to the cluster with highest membership value. These hard cluster assignments are then used to generate soft sub-cluster assignments using K -sub-clustering networks. The resulting sub-cluster assignments are the basis for making split and merge decisions to adapt and infer the optimal number of clusters dynamically. For the sub-clustering stage, the following isotropic loss is used :

$$\mathcal{L}_{sub-cl} = \sum_{k=1}^K \sum_{i=1}^{N_k} \sum_{j=1}^2 \tilde{r}_{i,j} || P_{h_i}^t - \tilde{\mu}_{k,j} ||_{\ell_2}^2 \quad (6.2)$$

where N_k is the number of body poses that are hard assigned to cluster k , $\tilde{r}_{i,j}$ is the soft assignment of pose $P_{h_i}^t$ to sub-cluster j , and $\tilde{\mu}_{k,j}$ denotes the mean of sub-cluster j belonging to cluster k . Splits and merges are then used to change the number of clusters K adaptively.

To select the poses that are most informative for action recognition, we compute the cosine distance between all pose vectors and their corresponding cluster mean, and then the nearest pose to the center will be selected as a candidate action-representative pose. It is to note that since the clustering approach is non-parametric, the number of key poses for each human action can be varied from one sequence to another. By doing so, we ensure that all the important cues of the action are captured, regardless of the sequence length, or action complexity.

6.3.4 Object Embedding

While human body pose can provide valuable cues for recognizing a performed action, incorporating contextual information about the action environment can further enhance the accuracy of action recognition. To this end, we propose to incorporate surrounding objects embeddings to provide contextual information about the scenes. First, we detect and localize objects in each frame of the input video. For this, we use the popular object detector Faster R-CNN [45] that is built based on a deep convolutional neural network and uses a two-stage detection approach. In the first stage, a set of object proposals are generated by the network, which are candidate regions in the scene that may contain objects. In the second stage, the resulting proposals are classified as either objects or background.

After detecting the set of objects present in a scene, we associate each object with the person of interest in the scene by calculating the Euclidean distance between their corresponding bounding box centroids. The resulting objects associated with the person are then combined with the corresponding key-pose based representations for each person. This joint representation can then be used as input to the Transformer network to learn and predict the performed action by the human. By incorporating object embeddings into the learning process, we can capture the contextual environment in which actions occur, which can be particularly beneficial for accurate action recognition.

6.3.5 Transformer-based Action Representation Learning

Learning good features that can discriminate the performed action is crucial for accurate action recognition. Several attempts have been made to incorporate diverse modalities features. However, existing methods are still restricted in how they incorporate temporal information, in addition to their reliance on randomly selected features, leading to uninformative action representations. To solve this problem, we propose an efficient learning model based on Vision Transformer networks to benefit from the effectiveness of self-attention in learning fine-grained action-related features.

Formally, we represent the selected K key poses for each human h_i in a scene as $\hat{P}_{h_i} = \{\hat{P}_{h_i}^1, \hat{P}_{h_i}^2, \dots, \hat{P}_{h_i}^K\}$. For every pose in $\hat{P}_{h_i}$, we retrieve its corresponding RGB image patch by referring to its bounding box. Once all the person key poses have been represented by their corresponding action image patches along with embedded objects, we group them together to form a sequence $\hat{A}_{h_i}$ (see Figure 6.2) serving as input to the Transformer encoder. We map the input sequence to dimension D , using the following linear projection :

$$A_0 = [A_{cls}; \hat{A}_{h_i}E] + E_{pos} \quad (6.3)$$

Following Dosovitskiy et al. [19], both a trainable embedding token, A_{cls} , and positional embeddings, E_{pos} , are added to the set of embedded patches. After generating the sequence of embedded representations, we train the Transformer encoder to learn and recognize human action patterns from these sets of representations. Specifically, we use the pure Vision Transformer model with a multi-head attention mechanism, and the cross-entropy loss as the training objective. The structure of the transformer encoder [18] comprises a series of layers consisting of multi-head self-attention and MLP blocks. Additionally, layernorm (LN) is applied prior to each block, while residual connections are incorporated after each block.

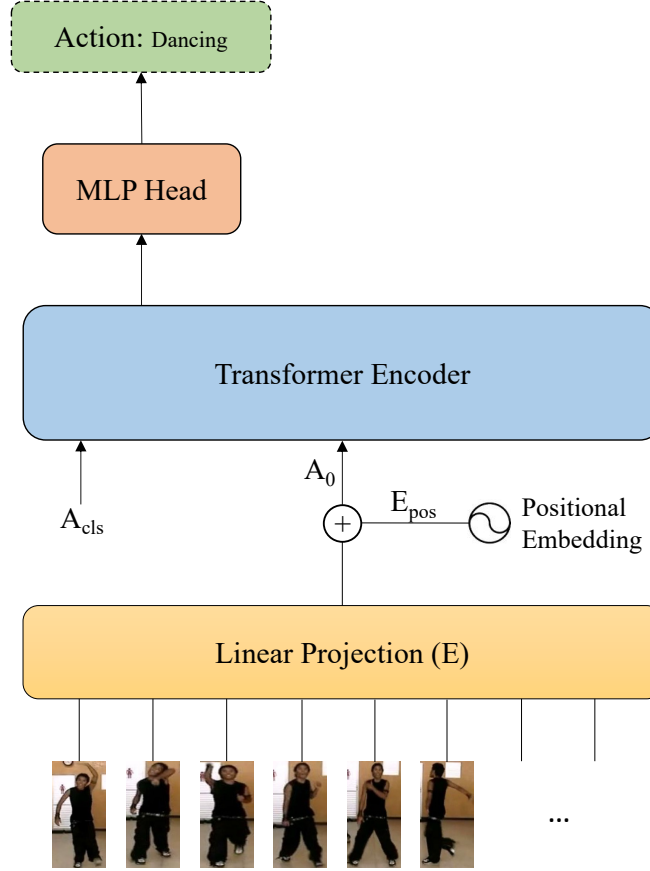


FIGURE 6.2 Overview of the action Vision Transformer model. E_{pos} is a positional embedding vector and A_{cls} is a learnable embedding token.

6.4 Experiments

In this section, we present our experiments on three benchmark datasets, Collective Activity [1], Collective Activity Extended [149] and UCLA Courtyard [26]. We will describe these datasets in detail, and then present the results of our ablation study experiments, followed by a comparison with state-of-the-art approaches.

6.4.1 Datasets

In our experiments, we evaluated our proposed model on the following real-world datasets :

- Collective Activity dataset (CA) [1] : is a commonly used dataset for recognizing human activities. It comprises 44 video sequences, each with a resolution of 640x480 pixels, that belong to one of the five categories of individual actions and group activities. The performed action of the majority of individuals determines the nature of the

group activity label in a scene. To evaluate our model on this dataset, we employed the same protocol recommended by Choi et al. [1].

- Collective Activity Extended dataset (CAE) [2] : is an augmented version of the Collective Activity dataset where two new activities "dancing" and "jogging" were added, and the activity "walking" was removed.
- UCLA Courtyard dataset [26] : consists of a set of videos of 106 minutes from a bird-eye viewpoint, with a resolution of 2560x1920 pixels and a frame rate of 30 fps. It is labeled with six group activities, and ten individual actions.

6.4.2 Implementation Details

We implemented our proposed model using the PyTorch framework [150]. To select the key poses, we trained the DeepDPM model with a learning rate of $2e-3$ for 200 epochs with a batch size of 128. For pose estimation, we used the ViTPose [146] model on the three datasets, specifically, the huge model trained on MS COCO [103], and we set the dimensionality of keypoints to $d = 17$. For object detection, we used Faster R-CNN [45] with a ResNeXt-101-FPN backbone trained on COCO. For the Transformer encoder, we trained the ViT model [19] for a total of 400 epochs with a batch size of 64 and a learning rate of $3e-5$, using the AdamW optimizer [151]. All our experiments were performed using an Nvidia RTX 3090 GPU.

6.4.3 Ablation studies

In this section, we provide and discuss the results of a series of ablation experiments we conducted to evaluate the importance of various components in our human action recognition model. Specifically, we examined the impact of the following modules on the recognition performance.

Parametric vs Non-Parametric Key Pose Identification

To validate the effectiveness of our proposed dynamic key pose identification mechanism, we compare the performance of our model when considering a fixed number of key poses versus dynamically selecting key poses. For a non-parametric key pose identification, we use the DeepDPM clustering model [147] to get the optimal number of action representative key poses. While for the parametric approach, we use the Deep Embedded Clustering (DEC) model [136] with a fixed number of clusters.

The results of this experiment are reported in Figure 6.3. We can observe that using dynami-

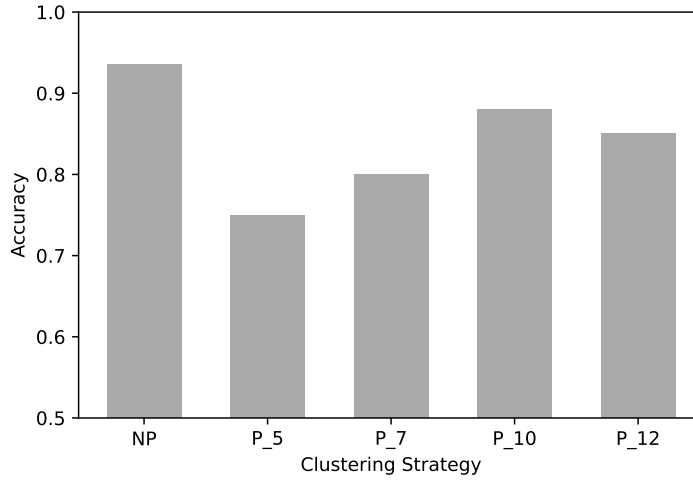


FIGURE 6.3 Impact of the clustering strategy on the performance of our model on the CAE dataset. "NP" refers to our model using the non-parametric approach, and "P_5", "P_7", "P_10", "P_12", refer to the parametric approach with fixed numbers of poses $\in \{5, 7, 10, 12\}$.

cally selected key poses leads to better performance compared to using fixed numbers of key poses. This means that inferring the optimal number of key poses from the input sequence is better to capture the variations of each action, leading to a more accurate representation of every performed action. In fact, each action is naturally described by a specific number of key poses that reflect the main body gestures, which can vary from one action to another. For instance, the action "dancing" is better discriminated with multiple poses, while the action "waiting" can be accurately identified using fewer poses. Therefore, dynamically selecting key poses also allows our model to handle sequences of varying lengths, making it more versatile in real-world scenarios where videos may have different durations or contain actions with varying levels of complexity.

RGB vs Pose-based Patch Selection

In the same direction, and to further understand the effectiveness of our proposed key pose identification approach, we evaluated the impact of selecting action image patches by clustering body poses versus clustering raw RGB frames, both using the non-parametric clustering approach. In Table 6.1, we summarize the results of this experiment. As shown in the table, our proposed pose-based selection approach achieves an accuracy of 92.8%, compared to 88.6% when using raw RGB frames for clustering and selection. This suggests that body poses carry informative features that better help distinguishing between actions compared to

TABLEAU 6.1 Impact of the integration of different modules on the recognition performance of our model on the CAE dataset.

Model variant	Accuracy (%)
RGB-based dynamic patch selection	88.6
Pose-based dynamic patch selection	92.8
Pose-based dynamic patch selection with object embedding	93.5

raw RGB frames. In addition, pose features can compensate for information loss in situations where RGB appearances are ambiguous, further highlighting the advantages of clustering pose features for accurate action recognition. It should be noted that our approach does not rely on ground truth pose annotations, instead, we estimate the pose from video sequences. This makes our approach more applicable for real-world datasets where pose annotations are not available.

Incorporating Object Embedding

In this study, we investigate the importance of integrating object embeddings on the action recognition performance of ActiViT. To do so, we evaluate our model without and with detected objects features. In the first experiment, we do not consider the object embedding module in our pipeline and we only use the pose-based patches as input to the Transformer encoder. In the second experiment, we follow the same process depicted in Figure 6.1, where we feed to the Transformer encoder the set of pose-based patches with their nearest detected object. The results of both experiments are reported in Table 6.1. This demonstrates the importance of integrating object features for human action recognition. In other experiments, we observed a significant improvement in the performance of ActiViT especially on the UCLA Courtyard dataset when object features were included. This highlights the importance of considering contextual information in datasets like UCLA Courtyard where some actions involve objects like "riding-bike" and "riding-skateboard".

6.4.4 Comparison with the State-of-the-Art

Once we assessed the impact of each component on the performance of our model through the ablation study, we selected the optimal version that includes all components (i.e., pose estimation, key pose identification, object embedding, Transformer-based action representation learning), and we compared the performance of our model with the existing state-of-the-art approaches for human action recognition.

TABLEAU 6.2 Comparison of the performance of our model with state of the art models on the CA Dataset. Boldface : Best results.

Model	Accuracy (%)
Choi et al. [2]	70.9
Tran et al. [111]	78.7
Ibrahim et al. [24]	81.5
Deng et al. [102]	81.2
Lu et al. [113]	90.6
Shu et al. [98]	87.2
Qi et al. [100]	89.1
Zhang et al. [112]	83.8
ActiViT (Ours)	90.8

TABLEAU 6.3 Comparison of the performance of our model with state of the art models on the CAE Dataset. Boldface : Best results.

Model	Accuracy (%)
Choi et al. [2]	82.0
Tran et al. [111]	80.7
Deng et al. [102]	90.2
Qi et al. [100]	89.7
Lu et al. [113]	91.2
Ehsanpour et al. [152]	89.4
Liu et al. [153]	90.8
ActiViT (Ours)	93.5

For comparison, we considered the following approaches : Learning context [2], StagNet [100], Fast collective activity [112], Gaim [113], CERN [98], Social cues [111], Structure inference machines [102], Hierarchical deep temporal model [24], Activities joint learning [152], Deep fully connected model [153], Cost-sensitive multiscale activity [26], Contextual activity model [154], Large scale max-margin classification [155], and Monte Carlo tree search activity [156].

In table 6.2 and 6.3, we report the results of our proposed model on the Collective Activity and the Collective Activity Extended datasets, respectively, using the accuracy metric, in comparison to existing action recognition approaches. Our proposed method outperforms existing methods on both CA and CAE with an accuracy of 90.8% and 93.5%, respectively. This represents a significant improvement over previous methods utilizing recurrent neural networks [98, 100, 102], hierarchical relational networks [24], and graph convolutional networks [113]. These results demonstrate the effectiveness of our four-stage method in learning discriminative action features that are useful for predicting human actions in videos.

TABLEAU 6.4 Comparison of the performance of our model with state of the art models on the UCLA Courtyard Dataset. Boldface : Best results.

Model	Average Precision	False Positive rates
Amer et al. [26]	68.1	14.1
Tran et al. [154]	71.4	-
Hariharan et al. [155]	74.9	-
Amer et al. [156]	79.3	12.7
Qi et al. [100]	81.5	11.6
ActiViT (Ours)	76.5	13.2

We also provide the results of ActiViT on UCLA Courtyard dataset compared with state-of-the-art approaches. Table 6.4 summarizes the results in terms of the average precision, and false positive rates metrics. We can observe that our proposed model obtains competitive results. Specifically, we find that the dynamic action image patch selection driven by non-parametric pose clustering leads to further improvements, despite the challenges imposed by this dataset, such as class imbalance and the high-level of interactions causing occlusions. We also notice that Qi et al. [100] achieve the best performance on the UCLA dataset. However, this model works differently from ours, as it employs a semantic graph architecture which facilitated the spatio-temporal representations learning of the interactions between individuals in this dataset. Compared to some existing methods that require complex computations as the sequence length increases, our approach has the advantage of handling variable-length sequences effectively, through the use of our fine-grained key pose selection process and the Transformer attention mechanism.

6.5 Conclusion

In this article, we presented a novel method for human action recognition in videos. It consists of four modules : human pose estimation, key pose identification, object embedding and Transformer-based action representation learning. By guiding the dynamic selection of action-representative patches with the estimated poses and combining them with objects features, our model is able to effectively learn informative features that are useful in discriminating between different actions. The results of our experiments on three benchmark human action recognition datasets demonstrate the effectiveness of our proposed method in recognizing human actions in challenging environments.

CHAPITRE 7 DISCUSSION GÉNÉRALE

Dans ce chapitre, nous présentons une analyse approfondie des avantages, des limitations, ainsi que des défis liés à chacune des méthodes proposées de reconnaissance d'activités humaines dans les vidéos. De plus, nous explorons l'utilité des composants de ces méthodes et les possibilités d'amélioration visant à rendre ces approches applicables à des scénarios réalistes.

7.1 Reconnaissance des activités humaines dans le spectre visible

La méthode de reconnaissance d'activités humaines proposée, GRAR, a démontré des performances prometteuses sur plusieurs ensembles de données, témoignant de sa robustesse face aux problèmes d'occlusion et de variations d'apparence au sein d'une même classe. Nous avons également prouvé l'impact significatif de l'intégration de la pose pour caractériser et extraire les éléments clés de l'activité exécutée, par rapport à une reconnaissance d'activité sans informations de pose. Toutefois, il est important de noter certaines limitations potentielles. Premièrement, dans certaines situations, il est possible que la pose soit mal estimée. Par exemple, si les conditions d'éclairage sont défavorables, si les sujets sont éloignés de la caméra ou si les mouvements sont rapides et complexes, la précision de la détection de la pose peut être compromise. Ceci pourrait conduire à une sélection inappropriée des caractéristiques et donc avoir un impact sur la précision de la reconnaissance d'activités humaines.

Parmi les modules clés de GRAR, l'exploitation d'une grille de représentation des informations spécifiques à chaque activité est importante. Cette approche permet de mettre davantage l'accent sur les caractéristiques clés de chaque activité guidées par la pose, plutôt que de traiter toutes les trames indépendamment. Cependant, il convient de noter que la dimension de cette représentation est fixe, ce qui signifie que toutes les activités sont représentées par le même nombre de caractéristiques clés. Cette approche peut présenter certaines limitations, notamment en fonction de la durée et de la complexité des activités. Prenons l'exemple d'une activité complexe telle que "danser". Le nombre de poses temporelles clés nécessaires pour décrire cette activité est plus élevé que celui requis pour décrire une activité simple comme "courir". L'hypothèse selon laquelle un nombre fixe de poses clés peut décrire chaque activité peut donc ne pas être suffisamment flexible pour capturer les informations spatiales et temporelles discriminantes des différentes activités. En conséquence, il est possible que certaines formes ou caractéristiques importantes puissent être omises dans la représentation, ce qui pourrait compromettre la qualité des résultats obtenus par GRAR. Afin de pallier

cette limitation, il est nécessaire d’explorer des approches plus flexibles pour représenter les activités dans le temps. C’est donc en réponse à ce défi que nous avons proposé la méthode décrite dans le Chapitre 6. Cette approche non-paramétrique vise à représenter les activités de manière plus flexible et complète, en tenant compte de leur durée et de leur complexité respectives. Cependant, elle nécessite une plus grande quantité de données d’entraînement pour atteindre des performances de reconnaissance d’activités optimales.

7.2 Reconnaissance des activités avec identification d’acteur principal

La méthode présentée dans le chapitre 5 a été développée dans le but de permettre une reconnaissance fiable des activités, même lorsque les données sont partiellement annotées, où les étiquettes sont attribuées par séquence plutôt que par acteur. C’est pour cette raison que nous avons proposé d’intégrer un module d’identification de l’acteur principal basé sur le flot optique. En plus de cela, les résultats ont démontré que cette approche est également applicable à la reconnaissance d’activités humaines dans le spectre infrarouge, ouvrant ainsi des perspectives d’utilisation dans des environnements réels et diversifiés.

Afin d’apprendre les caractéristiques spatiales des activités, la méthode ActAR utilise un réseau de neurones convolutif, en particulier le modèle Inception-ResNet-v2, pré-entraîné sur des données RGB. Lorsqu’il est appliqué aux données infrarouges, cette approche tire parti de l’apprentissage préalable réalisé sur les données RGB, qui contiennent des informations visuelles riches. Les caractéristiques apprises à partir de ces données peuvent fournir des informations complémentaires sur la texture aux données infrarouges, permettant ainsi une reconnaissance précise des activités

Tant dans le spectre infrarouge que dans le spectre visible, il reste toutefois que dans certaines situations, identifier l’acteur principal dans de telles conditions est un défi, notamment lorsqu’il y a un niveau élevé de mouvement dans la scène observée. Dans de telles situations, le mouvement intense peut entraîner des perturbations dans les informations fournies par le flot optique, ce qui complique l’identification précise de l’acteur principal. Cependant, malgré ce défi, notre méthode offre des résultats prometteurs dans la reconnaissance d’activités humaines, même dans des conditions pareilles.

Contrairement au spectre visible, où les détails visuels et les caractéristiques de l’environnement peuvent fournir des indices supplémentaires pour l’identification des acteurs, les caméras infrarouges peuvent avoir une résolution spatiale plus faible et limitée. Cela signifie que les informations contextuelles spécifiques à l’environnement ou les objets, peuvent être moins perceptibles. Par conséquent, il peut parfois être difficile de différencier clairement l’acteur

principal des autres individus et éléments présents dans la scène.

7.3 Modélisation temporelle des activités à court et à long terme

Le chapitre 6 présente une méthode de reconnaissance d'activités humaines dynamiques dans les vidéos. Nous avons démontré que cette méthode peut modéliser efficacement les dépendances à court et à long terme des caractéristiques de transition associées aux activités exercées. De plus, nous avons démontré la pertinence de l'incorporation des encodages des objets. En les combinant avec les patchs RGB clés des régions d'intérêt spécifiques décrivant l'activité de chaque individu, nous avons pu obtenir une représentation sémantique et précise des caractéristiques de l'activité humaine. Cette combinaison a permis de capturer les informations contextuelles importantes, et d'identifier plus précisément les dépendances entre les différents éléments de la scène. Par conséquent, cette combinaison a amélioré la précision et la richesse de la compréhension des activités réalisées par les individus.

Bien que notre méthode proposée intègre les encodages des objets pour améliorer la représentation sémantique et la compréhension des activités, ceci peut être difficile dans certaines situations. La résolution de la séquence vidéo peut affecter la capacité de la méthode à capturer tous les détails fins, y compris ceux nécessaires pour estimer la pose ainsi que les caractéristiques des objets présents dans la scène. En particulier, les petits objets peuvent être négligés lors de la détection car ils sont moins visibles et peuvent être confondus avec le bruit d'arrière-plan. En plus, la complexité des scènes peut également influencer la détection d'objets, car la présence de nombreux objets similaires peut introduire de l'ambiguïté, ce qui peut entraîner des erreurs de reconnaissance d'activités effectuées par les individus.

Tout comme les méthodes précédentes, cette approche présente également une dépendance à l'égard de la qualité de la pose humaine pour extraire les patchs clés. Si la détection de la pose est imprécise ou comporte des erreurs, cela peut entraîner une sélection inadéquate des patchs clés et une représentation erronée des caractéristiques de l'activité humaine. Cependant, cette limitation est partiellement atténuée grâce à l'utilisation d'un mécanisme de sélection de patchs clés non paramétrique dans notre méthode. Ce mécanisme permet de compenser les détections incorrectes ou imprécises en choisissant les patchs clés les plus pertinents et discriminatifs. Par conséquent, même en présence d'erreurs dans la détection de la pose humaine, notre méthode est en mesure de sélectionner les patchs clés les plus cohérents et représentatifs pour décrire l'activité humaine de manière précise. Néanmoins, il est important de souligner que malgré ces améliorations, une détection précise et fiable de la pose demeure essentielle pour assurer des performances optimales.

7.4 Utilité des différents éléments

Dans cette section nous allons discuter de l'utilité des composants de chacune des méthodes, ainsi que de leurs types d'applications spécifiques.

L'utilisation de la région d'intérêt définie par la boîte englobante de chaque individu, plutôt que l'ensemble de la trame, pour représenter les activités de chaque individu offre plusieurs avantages. Tout d'abord, elle permet de réduire la quantité d'informations à traiter par le réseau de neurones convolutif ou le réseau auto-attentif, ce qui peut optimiser le processus d'inférence. De plus, en excluant les régions qui ne contiennent pas d'individus, le modèle peut éviter d'apprendre des caractéristiques sur des éléments non pertinents et mieux résister au bruit ou aux variations mineures dans la scène. Le modèle peut également mieux capturer les informations discriminantes liées à l'activité observée, améliorant ainsi la précision et la qualité de ses prédictions.

Nous optimisons davantage cette approche en nous concentrant exclusivement sur les poses clés, plutôt que de traiter chaque instant de la séquence vidéo. Ceci permet de réduire encore plus la quantité de données à traiter et simplifie le processus d'inférence. Au lieu d'analyser chaque image de la séquence vidéo, nos modèles proposés se focalisent uniquement sur les moments clés qui représentent les informations essentielles qui décrivent chaque activité.

Pour le cas des applications liées aux systèmes embarqués, ayant des contraintes de ressources matérielles et de puissance de calcul limitées, nous recommandons l'utilisation de l'algorithme PAM pour la sélection des poses clés. Ceci permet une extraction efficace et optimisée des caractéristiques les plus représentatives des activités humaines dans les vidéos. Ainsi, il permet de réduire les exigences en termes de puissance de calcul tout en maintenant de bonnes performances de reconnaissance.

Afin d'améliorer significativement les performances de reconnaissance d'activités humaines, l'adoption de l'algorithme DeepDPM pour la sélection des poses clés s'avère particulièrement bénéfique. En effet, cette approche permet une capture dynamique et précise des caractéristiques distinctives des activités à court et à long terme. De cette manière on peut obtenir une compréhension plus approfondie des activités et des comportements des individus dans les environnements publics, ce qui permet par exemple de prendre des décisions éclairées pour optimiser les ressources et améliorer la planification urbaine.

Dans les situations où l'on souhaite reconnaître l'activité la plus intense dans une séquence vidéo ou lorsque les annotations disponibles sont limitées ou incertaines, l'utilisation du flot optique pour identifier l'acteur principal dans la scène est particulièrement avantageuse. En se focalisant sur l'acteur principal, cette approche permet de différencier clairement ses

actions des autres mouvements environnants. Ceci facilite la reconnaissance de l'activité la plus significative, même en présence de plusieurs individus dans la scène, contrairement à une analyse globale de l'ensemble de la scène.

CHAPITRE 8 CONCLUSION

Dans le cadre de cette thèse, nous avons présenté de nouvelles méthodes pour l'identification et la reconnaissance d'activités humaines dans des séquences vidéo. Ces méthodes ont été spécifiquement développées pour relever les défis inhérents aux environnements réalistes tel que les espaces publics. Nous avons exploité des techniques avancées d'apprentissage automatique, en particulier l'apprentissage profond, afin d'atteindre cet objectif.

8.1 Synthèse des travaux

La première méthode proposée de reconnaissance d'activités humaines est une méthode basée sur l'estimation de la pose des individus présents dans la scène observée. Cette approche permet de déterminer la position des différentes articulations du corps, fournissant ainsi des informations précieuses pour comprendre et reconnaître les activités humaines. Étant donné que les poses estimées résultantes sont souvent redondantes et que toutes ces poses ne sont pas nécessairement indicatives de l'activité en cours, un algorithme de regroupement est utilisé ensuite pour extraire des poses clés à partir des poses estimées. Les poses clés identifiées, avec leurs boîtes englobantes correspondantes en RGB, sont ensuite utilisées pour créer une représentation en grille compacte. Cette représentation encode toutes les informations spatio-temporelles essentielles à la reconnaissance d'activités.

La deuxième méthode proposée de reconnaissance d'activités humaines est une approche complète de la reconnaissance d'activités humaines, visant à résoudre les défis liés aux informations non pertinentes résultant de la présence de plusieurs personnes actives qui ne contribuent pas à l'activité d'intérêt. Cette méthode permet d'identifier et de localiser l'acteur principal dans une scène donnée en exploitant le flux optique. Les poses clés de l'acteur identifié sont ensuite extraites afin de guider le processus de sélection des caractéristiques les plus représentatives de l'activité d'intérêt. En outre, cette méthode est également applicable au spectre infrarouge, permettant ainsi une reconnaissance efficace des activités humaines. Elle est capable d'identifier automatiquement les acteurs clés impliqués dans l'activité, sans nécessiter de connaissances préalables ou d'annotations explicites.

La troisième méthode proposée pour la reconnaissance d'activités humaines consiste à extraire des représentations discriminantes de la dynamique des activités. Cette approche utilise un algorithme de regroupement non-paramétrique pour modéliser de manière efficace les relations temporelles au sein des activités, qu'elles soient de courte ou de longue durée. De

plus, elle s'adapte dynamiquement afin de capturer toutes les caractéristiques clés d'une activité. Les représentations obtenues sont ensuite fournies à un réseau auto-attentif qui est capable d'apprendre les caractéristiques spatio-temporelles liées aux dépendances et interactions complexes entre les différents éléments de l'activité humaine dans l'espace et le temps.

8.2 Recommandations pour travaux futurs

La reconnaissance d'activités humaines dans les vidéos est en constante évolution, offrant de nombreuses possibilités d'amélioration et de développement, grâce aux avancées significatives que ce domaine a connu au fil des années. Dans ce qui suit, nous présentons différentes directions de recherche qui pourraient contribuer à améliorer la précision et la fiabilité de la reconnaissance d'activités humaines dans les vidéos.

Les méthodes décrites dans cette thèse peuvent être adaptées pour reconnaître également les activités de groupe, en passant du niveau individuel à un niveau collectif. Cela peut être réalisé en intégrant les informations d'interaction, le contexte spatial ainsi que la direction de chaque individu. En considérant ces aspects supplémentaires, il sera possible de capturer les relations complexes entre les membres du groupe et d'identifier les schémas d'interactions plus spécifiques, permettant ainsi d'améliorer la précision et la compréhension globale des activités de groupe.

Les futurs travaux dans le domaine de la reconnaissance d'activités humaines peuvent tirer parti des informations multimodales présentes dans les vidéos, qui contiennent à la fois du contenu visuel et auditif. Par exemple, dans les espaces publics, en complément d'analyse des mouvements visuels, le son de cris ou de bruits de conflit pourrait indiquer la présence de problèmes de sécurité dans l'endroit surveillé, bien plus que si l'on analysait seulement des images vidéo. De ce fait, pour améliorer la précision de la reconnaissance, il est possible d'exploiter ces informations en intégrant les différentes modalités d'apparence, de mouvement et de son, et en adaptant le réseau auto-attentif de vision pour permettre un apprentissage multimodale intégrant des mécanismes d'auto-attention spatio-temporelle. En combinant les informations visuelles et auditives, il devient possible de capturer les relations complexes entre les différentes modalités et d'extraire des caractéristiques pertinentes pour la reconnaissance d'activités.

Dans les méthodes proposées, l'extraction des caractéristiques des activités a été basée sur l'information de la pose en 2D plutôt que de la pose en 3D en raison de sa simplicité d'extraction. Cependant, récemment, la reconstruction de la pose en 3D à partir des informations en 2D, comme la pose en 2D ou une image RGB, a gagné en importance [157, 158]. L'intégration

de cette information permettrait de bénéficier des avantages de la représentation en 3D, tels que la capture plus précise des mouvements et des relations spatiales entre les différentes parties du corps humain, ce qui pourra améliorer considérablement la précision et la qualité de la reconnaissance d'activités humaines.

RÉFÉRENCES

- [1] W. Choi, K. Shahid et S. Savarese, “What are they doing ? : Collective activity classification using spatio-temporal relationship among people,” dans *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, p. 1282–1289.
- [2] ———, “Learning context for collective activity recognition,” dans *CVPR 2011*. IEEE, 2011, p. 3273–3280.
- [3] O. Barnich et M. Van Droogenbroeck, “Vibe : A universal background subtraction algorithm for video sequences,” *IEEE Transactions on Image processing*, vol. 20, n^o. 6, p. 1709–1724, 2011.
- [4] OpenCV, “Optical flow,” https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html, consulté le : 22 Mai 2023.
- [5] A. F. Bobick et J. W. Davis, “The recognition of human movement using temporal templates,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, n^o. 3, p. 257–267, 2001.
- [6] L. Gorelick *et al.*, “Actions as space-time shapes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, n^o. 12, p. 2247–2253, 2007.
- [7] A. H. Shabani, D. A. Clausi et J. S. Zelek, “Evaluation of local spatio-temporal salient feature detectors for human action recognition,” dans *2012 Ninth Conference on Computer and Robot Vision*. IEEE, 2012, p. 468–475.
- [8] H. Wang *et al.*, “Action recognition by dense trajectories,” dans *CVPR 2011-IEEE Conference on Computer Vision & Pattern Recognition*. IEEE, 2011, p. 3169–3176.
- [9] Y. LeCun *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, n^o. 11, p. 2278–2324, 1998.
- [10] A. Karpathy *et al.*, “Large-scale video classification with convolutional neural networks,” dans *CVPR 2014*, 2014.
- [11] M. Asadi-Aghbolaghi *et al.*, “Deep learning for action and gesture recognition in image sequences : A survey,” dans *Gesture Recognition*, 2017.
- [12] D. Tran *et al.*, “Learning spatiotemporal features with 3d convolutional networks,” *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 4489–4497, 2015.
- [13] K. Simonyan et A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” dans *NIPS*, 2014.

- [14] L. Wang *et al.*, “Temporal segment networks : Towards good practices for deep action recognition,” dans *ECCV*, 2016.
- [15] L. Wiest, “Recurrent neural networks - combination of rnn and cnn,” <https://wiki.tum.de/display/lfdv/Recurrent+Neural+Networks+-+Combination+of+RNN+and+CNN>, consulté le : 29 Avril 2019.
- [16] J. Donahue *et al.*, “Long-term recurrent convolutional networks for visual recognition and description,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, p. 677–691, 2015.
- [17] S. Sharma, R. Kiros et R. R. Salakhutdinov, “Action recognition using visual attention,” dans *ICLR 2015*, 2015.
- [18] A. Vaswani *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [19] A. Dosovitskiy *et al.*, “An image is worth 16x16 words : Transformers for image recognition at scale,” dans *The International Conference on Learning Representations*, 2021.
- [20] R. Girdhar *et al.*, “Video action transformer network,” dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 244–253.
- [21] J. Yang *et al.*, “Recurring the transformer for video action recognition,” dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 14 063–14 073.
- [22] Y. Zhang *et al.*, “Vidtr : Video transformer without convolutions,” dans *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, p. 13 577–13 587.
- [23] I. Laptev, B. Caputo *et al.*, “Recognizing human actions : a local svm approach,” dans *null. IEEE*, 2004, p. 32–36.
- [24] M. S. Ibrahim *et al.*, “A hierarchical deep temporal model for group activity recognition,” dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, p. 1971–1980.
- [25] C. Gao *et al.*, “Infar dataset : Infrared action recognition at different times,” *Neurocomputing*, vol. 212, p. 36–47, 2016.
- [26] M. R. Amer *et al.*, “Cost-sensitive top-down/bottom-up inference for multiscale activity recognition,” dans *The 12th European Conference on Computer Vision–ECCV, Florence, Italy*, 2012.

- [27] “L’évolution démographique,” 2020. [En ligne]. Disponible : <https://www.un.org/fr/un75/shifting-demographics>
- [28] Y. Du, W. Wang et L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, p. 1110–1118.
- [29] C. Zimmermann *et al.*, “3d human pose estimation in rgb-d images for robotic task learning,” dans *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, p. 1986–1992.
- [30] A. Shahroudy *et al.*, “Ntu rgb+ d : A large scale dataset for 3d human activity analysis,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 1010–1019.
- [31] J. Liu *et al.*, “Spatio-temporal lstm with trust gates for 3d human action recognition,” dans *European conference on computer vision*. Springer, 2016, p. 816–833.
- [32] G. Chéron, I. Laptev et C. Schmid, “P-cnn : Pose-based cnn features for action recognition,” dans *Proceedings of the IEEE international conference on computer vision*, 2015, p. 3218–3226.
- [33] V. Choutas *et al.*, “Potion : Pose motion representation for action recognition,” dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, p. 7024–7033.
- [34] A. Karpathy *et al.*, “Large-scale video classification with convolutional neural networks,” dans *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, p. 1725–1732.
- [35] S. Lamghari, G.-A. Bilodeau et N. Saunier, “A grid-based representation for human action recognition,” dans *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, p. 10 500–10 507.
- [36] M. Piccardi, “Background subtraction techniques : a review,” dans *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, vol. 4. IEEE, 2004, p. 3099–3104.
- [37] A. Elgammal, D. Harwood et L. Davis, “Non-parametric model for background subtraction,” dans *European conference on computer vision*. Springer, 2000, p. 751–767.
- [38] B. K. Horn et B. G. Schunck, “Determining optical flow,” *Artificial intelligence*, vol. 17, n°. 1-3, p. 185–203, 1981.
- [39] J. L. Barron, D. J. Fleet et S. S. Beauchemin, “Performance of optical flow techniques,” *International journal of computer vision*, vol. 12, n°. 1, p. 43–77, 1994.

- [40] C. Stauffer et W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” dans *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2. IEEE, 1999, p. 246–252.
- [41] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” dans *Scandinavian conference on Image analysis*. Springer, 2003, p. 363–370.
- [42] M. Shibata, Y. Yasuda et M. Ito, “Moving object detection for active camera based on optical flow distortion,” *IFAC Proceedings Volumes*, vol. 41, n°. 2, p. 14 720–14 725, 2008.
- [43] N. Dalal et B. Triggs, “Histograms of oriented gradients for human detection,” dans *international Conference on computer vision & Pattern Recognition (CVPR’05)*, vol. 1. IEEE Computer Society, 2005, p. 886–893.
- [44] C. L. Zitnick et P. Dollár, “Edge boxes : Locating object proposals from edges,” dans *European conference on computer vision*. Springer, 2014, p. 391–405.
- [45] S. Ren *et al.*, “Faster r-cnn : Towards real-time object detection with region proposal networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [46] J. Redmon *et al.*, “You only look once : Unified, real-time object detection,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 779–788.
- [47] D. Weinland, R. Ronfard et E. Boyer, “Free viewpoint action recognition using motion history volumes,” *Computer vision and image understanding*, vol. 104, n°. 2-3, p. 249–257, 2006.
- [48] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, n°. 2, p. 91–110, 2004.
- [49] I. Laptev, “On space-time interest points,” *International journal of computer vision*, vol. 64, n°. 2-3, p. 107–123, 2005.
- [50] H. Wang *et al.*, “Evaluation of local spatio-temporal features for action recognition,” dans *BMVC 2009-British Machine Vision Conference*. BMVA Press, 2009, p. 124–1.
- [51] R. Messing, C. Pal et H. Kautz, “Activity recognition using the velocity histories of tracked keypoints,” dans *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, p. 104–111.
- [52] P. Matikainen, M. Hebert et R. Sukthankar, “Trajectons : Action recognition through the motion analysis of tracked features,” dans *2009 IEEE 12th international conference on computer vision workshops, ICCV workshops*. IEEE, 2009, p. 514–521.

- [53] M. Jain, H. Jegou et P. Bouthemy, “Better exploiting motion for better action recognition,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, p. 2555–2562.
- [54] B. D. Lucas et T. Kanade, “An iterative image registration technique with an application to stereo vision,” dans *IJCAI’81 : 7th international joint conference on Artificial intelligence*, vol. 2, 1981, p. 674–679.
- [55] H. Wang *et al.*, “Dense trajectories and motion boundary descriptors for action recognition,” *International journal of computer vision*, vol. 103, n^o. 1, p. 60–79, 2013.
- [56] H. Wang et C. Schmid, “Action recognition with improved trajectories,” dans *Proceedings of the IEEE international conference on computer vision*, 2013, p. 3551–3558.
- [57] A. Krizhevsky, I. Sutskever et G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” dans *Advances in neural information processing systems*, 2012, p. 1097–1105.
- [58] O. Russakovsky *et al.*, “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, n^o. 3, p. 211–252, 2015.
- [59] P. Turaga *et al.*, “Machine recognition of human activities : A survey,” *IEEE Transactions on Circuits and Systems for Video technology*, vol. 18, n^o. 11, p. 1473, 2008.
- [60] Y.-G. Jiang *et al.*, “High-level event recognition in unconstrained videos,” *International journal of multimedia information retrieval*, vol. 2, n^o. 2, p. 73–101, 2013.
- [61] Y. LeCun *et al.*, “Handwritten digit recognition with a back-propagation network,” dans *Advances in neural information processing systems*, 1990, p. 396–404.
- [62] R. Girshick *et al.*, “Rich feature hierarchies for accurate object detection and semantic segmentation,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, p. 580–587.
- [63] K. Simonyan et A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” dans *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.
- [64] C. Szegedy *et al.*, “Going deeper with convolutions,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 1–9, 2015.
- [65] J. Wang *et al.*, “Deep high-resolution representation learning for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, n^o. 10, p. 3349–3364, 2020.
- [66] S. Ioffe et C. Szegedy, “Batch normalization : Accelerating deep network training by reducing internal covariate shift,” dans *ICML*, 2015.

- [67] S. Ji *et al.*, “3d convolutional neural networks for human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, p. 221–231, 2010.
- [68] E. Mansimov, N. Srivastava et R. R. Salakhutdinov, “Initialization strategies of spatio-temporal convolutional neural networks,” *CoRR*, vol. abs/1503.07274, 2015.
- [69] L. Sun *et al.*, “Human action recognition using factorized spatio-temporal convolutional networks,” *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 4597–4605, 2015.
- [70] Y. Zhou *et al.*, “Mict : Mixed 3d/2d convolutional tube for human action recognition,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 449–458, 2018.
- [71] B. Zhang *et al.*, “Real-time action recognition with enhanced motion vector cnns,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2718–2726, 2016.
- [72] K. Chatfield *et al.*, “Return of the devil in the details : delving deep into convolutional nets,” dans *Proceedings of the British Machine Vision Conference 2014*. British Machine Vision Association, 2014.
- [73] J. Y.-H. Ng *et al.*, “Beyond short snippets : Deep networks for video classification,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 4694–4702, 2015.
- [74] L. Wang, Y. Qiao et X. Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 4305–4314, 2015.
- [75] C. Feichtenhofer, A. Pinz et R. P. Wildes, “Spatiotemporal residual networks for video action recognition,” dans *NIPS*, 2016.
- [76] Z. Wu *et al.*, “Deep learning for video classification and captioning,” *CoRR*, vol. abs/1609.06782, 2016.
- [77] A. H. Waibel *et al.*, “Phoneme recognition using time-delay neural networks,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, p. 328–339, 1989.
- [78] L. Yao *et al.*, “Describing videos by exploiting temporal structure,” *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 4507–4515, 2015.
- [79] B. Singh *et al.*, “A multi-stream bi-directional recurrent neural network for fine-grained action detection,” dans *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

- [80] A. Dosovitskiy *et al.*, “An image is worth 16x16 words : Transformers for image recognition at scale,” *arXiv preprint arXiv :2010.11929*, 2020.
- [81] G. Bertasius, H. Wang et L. Torresani, “Is space-time attention all you need for video understanding?” dans *ICML*, vol. 2, n°. 3, 2021, p. 4.
- [82] K. Simonyan et A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” dans *Advances in neural information processing systems*, 2014, p. 568–576.
- [83] L. Wang *et al.*, “Temporal segment networks : Towards good practices for deep action recognition,” dans *European conference on computer vision*. Springer, 2016, p. 20–36.
- [84] Y. Liu *et al.*, “Global temporal representation based cnns for infrared action recognition,” *IEEE Signal Processing Letters*, vol. 25, n°. 6, p. 848–852, 2018.
- [85] M. Zolfaghari *et al.*, “Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection,” dans *Proceedings of the IEEE International Conference on Computer Vision*, 2017, p. 2904–2913.
- [86] B. Xiaohan Nie, C. Xiong et S.-C. Zhu, “Joint action recognition and pose estimation from video,” dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, p. 1293–1301.
- [87] S. Woo *et al.*, “Cbam : Convolutional block attention module,” dans *Proceedings of the European conference on computer vision (ECCV)*, 2018, p. 3–19.
- [88] U. Gaur *et al.*, “A “string of feature graphs” model for recognition of complex activities in natural videos,” dans *2011 International Conference on Computer Vision*. IEEE, 2011, p. 2595–2602.
- [89] M. Sudha *et al.*, “Approaches and applications of virtual reality and gesture recognition : A review,” *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 8, n°. 4, p. 1–18, 2017.
- [90] L. Xia *et al.*, “Robot-centric activity recognition from first-person rgb-d videos,” dans *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, p. 357–364.
- [91] J. Donahue *et al.*, “Long-term recurrent convolutional networks for visual recognition and description,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, p. 2625–2634.
- [92] Y. Zhou *et al.*, “Mict : Mixed 3d/2d convolutional tube for human action recognition,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 449–458.

- [93] C. Wang, Y. Wang et A. L. Yuille, “An approach to pose-based action recognition,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, p. 915–922.
- [94] H. Jhuang *et al.*, “Towards understanding action recognition,” dans *Proceedings of the IEEE international conference on computer vision*, 2013, p. 3192–3199.
- [95] W. Du, Y. Wang et Y. Qiao, “Rpan : An end-to-end recurrent pose-attention network for action recognition in videos,” dans *Proceedings of the IEEE International Conference on Computer Vision*, 2017, p. 3725–3734.
- [96] Z. Li *et al.*, “Videolstm convolves, attends and flows for action recognition,” *Computer Vision and Image Understanding*, vol. 166, p. 41–50, 2018.
- [97] G. Dong et J. Li, “Efficient mining of emerging patterns : Discovering trends and differences,” dans *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, p. 43–52.
- [98] T. Shu, S. Todorovic et S.-C. Zhu, “Cern : confidence-energy recurrent network for group activity recognition,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 5523–5531.
- [99] T. Bagautdinov *et al.*, “Social scene understanding : End-to-end multi-person action localization and collective activity recognition,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 4315–4324.
- [100] M. Qi *et al.*, “stagnet : An attentive semantic rnn for group activity recognition,” dans *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, p. 101–117.
- [101] S. Biswas et J. Gall, “Structural recurrent neural network (srnn) for group activity analysis,” dans *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, p. 1625–1632.
- [102] Z. Deng *et al.*, “Structure inference machines : Recurrent neural networks for analyzing relations in group activity recognition,” dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, p. 4772–4781.
- [103] T.-Y. Lin *et al.*, “Microsoft coco : Common objects in context,” dans *European conference on computer vision*. Springer, 2014, p. 740–755.
- [104] L. Kaufman et P. J. Rousseeuw, “Clustering by means of medoids. statistical data analysis based on the l1 norm,” *Y. Dodge, Ed*, p. 405–416, 1987.
- [105] C. Szegedy *et al.*, “Inception-v4, inception-resnet and the impact of residual connections on learning,” dans *Thirty-first AAAI conference on artificial intelligence*, 2017.

- [106] J. Deng *et al.*, “Imagenet : A large-scale hierarchical image database,” dans *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, p. 248–255.
- [107] M. Abadi *et al.*, “Tensorflow : A system for large-scale machine learning,” dans *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, p. 265–283.
- [108] D. P. Kingma et J. Ba, “Adam : A method for stochastic optimization,” *arXiv preprint arXiv :1412.6980*, 2014.
- [109] C. Cao *et al.*, “Look and think twice : Capturing top-down visual attention with feedback convolutional neural networks,” dans *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, p. 2956–2964.
- [110] T. Lan *et al.*, “Discriminative latent models for recognizing contextual group activities,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, n^o. 8, p. 1549–1562, 2011.
- [111] K. N. Tran *et al.*, “Social cues in group formation and local interactions for collective activity analysis,” dans *VISAPP (1)*, 2013, p. 539–548.
- [112] P. Zhang *et al.*, “Fast collective activity recognition under weak supervision,” *IEEE Transactions on Image Processing*, vol. 29, p. 29–43, 2019.
- [113] L. Lu *et al.*, “Gaim : Graph attention interaction model for collective activity recognition,” *IEEE Transactions on Multimedia*, vol. 22, n^o. 2, p. 524–539, 2019.
- [114] J. Wu *et al.*, “Learning actor relation graphs for group activity recognition,” dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, p. 9964–9974.
- [115] M. Lu, Y. Hu et X. Lu, “Driver action recognition using deformable and dilated faster r-cnn with optimized region proposals,” *Applied Intelligence*, vol. 50, n^o. 4, p. 1100–1111, 2020.
- [116] D. Tran *et al.*, “Learning spatiotemporal features with 3d convolutional networks,” dans *Proceedings of the IEEE international conference on computer vision*, 2015, p. 4489–4497.
- [117] C. Feichtenhofer, A. Pinz et A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 1933–1941.
- [118] J. Carreira et A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” dans *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 6299–6308.

- [119] H. Wang *et al.*, “Action recognition by dense trajectories,” dans *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. USA : IEEE Computer Society, 2011, p. 3169–3176. [En ligne]. Disponible : <https://doi.org/10.1109/CVPR.2011.5995407>
- [120] S. Yan, Y. Xiong et D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” dans *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [121] L. Shi *et al.*, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, p. 12 026–12 035.
- [122] S. Asghari-Esfeden, M. Sznai et O. Camps, “Dynamic motion representation for human action recognition,” dans *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, p. 557–566.
- [123] Y. Zhu et G. Guo, “A study on visible to infrared action recognition,” *IEEE Signal Processing Letters*, vol. 20, n^o. 9, p. 897–900, 2013.
- [124] Z. Jiang, V. Rozgic et S. Adali, “Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks,” dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, p. 115–123.
- [125] M. E. Hussein *et al.*, “Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations,” dans *Twenty-third international joint conference on artificial intelligence*, 2013.
- [126] N. Srivastava, E. Mansimov et R. Salakhudinov, “Unsupervised learning of video representations using lstms,” dans *International conference on machine learning*. PMLR, 2015, p. 843–852.
- [127] H. Gammulle *et al.*, “Two stream lstm : A deep fusion framework for human action recognition,” dans *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, p. 177–186.
- [128] W. Du, Y. Wang et Y. Qiao, “Recurrent spatial-temporal attention network for action recognition in videos,” *IEEE Transactions on Image Processing*, vol. 27, n^o. 3, p. 1347–1360, 2017.
- [129] R. Pascanu, T. Mikolov et Y. Bengio, “On the difficulty of training recurrent neural networks,” dans *International conference on machine learning*. PMLR, 2013, p. 1310–1318.

- [130] D. C. Luvizon, D. Picard et H. Tabia, “2d/3d pose estimation and action recognition using multitask deep learning,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 5137–5146.
- [131] Y. Wei *et al.*, “3d single-person concurrent activity detection using stacked relation network,” dans *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, p. 12 329–12 337.
- [132] Y. Zhang *et al.*, “Bytetrack : Multi-object tracking by associating every detection box,” *arXiv preprint arXiv :2110.06864*, 2021.
- [133] S. Jiang *et al.*, “Learning to estimate hidden motions with global motion aggregation,” dans *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 9772–9781.
- [134] K. Sun *et al.*, “Deep high-resolution representation learning for human pose estimation,” dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 5693–5703.
- [135] J. Liu *et al.*, “Normalized human pose features for human action video alignment,” dans *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 11 521–11 531.
- [136] J. Xie, R. Girshick et A. Farhadi, “Unsupervised deep embedding for clustering analysis,” dans *International conference on machine learning*. PMLR, 2016, p. 478–487.
- [137] Q. V. Le, “Building high-level features using large scale unsupervised learning,” dans *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, p. 8595–8598.
- [138] L. Van Der Maaten, “Learning a parametric embedding by preserving local structure,” dans *Artificial intelligence and statistics*. PMLR, 2009, p. 384–391.
- [139] D. J. Butler *et al.*, “A naturalistic open source movie for optical flow evaluation,” dans *European conference on computer vision*. Springer, 2012, p. 611–625.
- [140] J. Imran et B. Raman, “Deep residual infrared action recognition by integrating local and global spatio-temporal cues,” *Infrared Physics & Technology*, vol. 102, p. 103014, 2019.
- [141] X. Chen *et al.*, “Infrared action detection in the dark via cross-stream attention mechanism,” *IEEE Transactions on Multimedia*, 2021.
- [142] Y. Liu *et al.*, “Transferable feature representation for visible-to-infrared cross-dataset human action recognition,” *Complexity*, vol. 2018, 2018.

- [143] G. Vallathan *et al.*, “Suspicious activity detection using deep learning in secure assisted living iot environments,” *The Journal of Supercomputing*, vol. 77, 2021.
- [144] A. Sheikh-Mohammad-Zadeh, N. Saunier et E. Waygood, “Developing an objective framework to evaluate street functions,” *Sustainability*, vol. 14, n°. 12, 2022.
- [145] S. Ji *et al.*, “3d convolutional neural networks for human action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, n°. 1, 2012.
- [146] Y. Xu *et al.*, “ViTPose : Simple vision transformer baselines for human pose estimation,” dans *Advances in Neural Information Processing Systems*, 2022.
- [147] M. Ronen, S. E. Finder et O. Freifeld, “Deepdpm : Deep clustering with an unknown number of clusters,” dans *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [148] J. Chang et J. W. Fisher III, “Parallel sampling of dp mixture models using sub-cluster splits,” *Advances in Neural Information Processing Systems*, vol. 26, 2013.
- [149] Y. Kong et Y. Fu, “Human action recognition and prediction : A survey,” *International Journal of Computer Vision*, vol. 130, n°. 5, 2022.
- [150] A. Paszke *et al.*, “Pytorch : An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.
- [151] S. J. Reddi, S. Kale et S. Kumar, “On the convergence of adam and beyond,” dans *The International Conference on Learning Representations*, 2018.
- [152] M. Ehsanpour *et al.*, “Joint learning of social groups, individuals action and sub-group activities in videos,” dans *The 16th European Conference on Computer Vision–ECCV, Glasgow, UK*, 2020.
- [153] J. Liu *et al.*, “Deep fully connected model for collective activity recognition,” *IEEE Access*, vol. 7, 2019.
- [154] K. N. Tran *et al.*, “A group contextual model for activity recognition in crowded scenes.” dans *VISAPP (2)*, 2015.
- [155] B. Hariharan *et al.*, “Large scale max-margin multi-label classification with priors,” dans *The 27th International Conference on Machine Learning (ICML-10)*, 2010.
- [156] M. R. Amer *et al.*, “Monte carlo tree search for scheduling activity recognition,” dans *The IEEE international conference on computer vision*, 2013.
- [157] C.-H. Chen et D. Ramanan, “3d human pose estimation= 2d pose estimation+ matching,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 7035–7043.

- [158] G. Pavlakos *et al.*, “Learning to estimate 3d human pose and shape from a single color image,” dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 459–468.