

Label-Efficient Skeleton-based Recognition with Stable-Invertible Graph Convolutional Networks

Hichem Sahbi

Sorbonne University, CNRS, LIP6, F-75005, Paris, France

Abstract

Skeleton-based action recognition is a hotspot in image processing. A key challenge of this task lies in its dependence on large, manually labeled datasets whose acquisition is costly and time-consuming. This paper devises a novel, label-efficient method for skeleton-based action recognition using graph convolutional networks (GCNs). The contribution of the proposed method resides in learning a novel acquisition function — scoring the most informative subsets for labeling — as the optimum of an objective function mixing data representativity, diversity and uncertainty. We also extend this approach by learning the most informative subsets using an invertible GCN which allows mapping data from ambient to latent spaces where the inherent distribution of the data is more easily captured. Extensive experiments, conducted on two challenging skeleton-based recognition datasets, show the effectiveness and the outperformance of our label-frugal GCNs against the related work.

1 INTRODUCTION

Skeleton-based recognition is a major task in image processing which consists in analyzing skeletal structures (human body, hands, etc.) by extracting joint positions and modeling their interactions. This task is particularly useful in challenging scenarios like cluttered environments. Early methods rely on handcrafted features [1–4, 6–8, 129], such as joint angles and relative distances, fed into classifiers like support vector machines and hidden Markov models [29–32, 37] as well as manifold learning [33–36]. With the rise of deep learning [114], recurrent neural networks (RNNs), particularly LSTMs and GRUs [11–13, 15, 17, 18], became popular for capturing the temporal dynamics of skeletal sequences. Graph Convolutional Networks (GCNs) have also emerged as powerful learning models, exploiting the inherent graph structures of skeletons to learn spatial relationships between joints, as a part of attention-based models [20, 21, 24, 38, 138]. The latter have been demonstrating impressive performances by effectively modeling long-range dependencies and capturing complex motion patterns.

The success of the aforementioned learning-based methods, for skeleton-based recognition, hinges on the availability of large, diverse datasets of hand-labeled skeleton sequences. However, acquiring such massive data sequences is known to be time and labor demanding. Several solutions address this issue including data augmentation [40], few shot and transfer learning [41], as well as self-supervised learning [42]. Nonetheless, the relative success of these solutions relies upon a strong assumption that knowledge are enough in order to close the *accuracy gap* while actually labeled data are more important. Another category of methods is active learning (AL) [52, 131] which excels at adapting to the “oracle” (expert annotator) in a way that other label-efficient methods do not. Indeed, unlike other methods, AL queries only the *most informative* unlabeled samples for annotation by quantifying and maximizing the impact of

labeling a particular sample on a learning model. Informative datasets are usually selected based on various criteria, notably diversity [49, 51] and uncertainty [42–44, 46, 48] in different contexts [55, 57, 58]. Uncertainty-based methods include margin sampling and entropy-based criteria [62, 64] while diversity-based approaches include coverage maximization [54, 60]. Other strategies consider the representativeness of data, selecting samples that are most similar to the overall data distribution. However, current approaches for identifying these informative subsets often rely on heuristics, lacking rigorous theoretically grounded framework. This limits the optimality of the selected data and can hinder the overall efficiency of AL.

Considering the aforementioned issues, we introduce in this paper a label-efficient GCN for skeleton-based recognition. The contribution of the proposed method resides in a novel principled probabilistic framework that designs unlabeled exemplars (candidate samples for labeling) instead of sampling them from a fixed pool of unlabeled data. These exemplars are obtained as an interpretable solution of an objective function mixing data representativity, diversity and uncertainty. Our proposed framework designs these exemplars using a stable and an invertible GCN that allows mapping input graphs (lying on highly nonlinear manifolds) from ambient (input) to latent spaces where designing these exemplars becomes more tractable; indeed, with the proposed GCNs, data in the latent space follow a standard probability distribution (namely gaussian) whose sampling and search is more tractable compared to the arbitrary distributions in the ambient space. Once designed, the learned exemplars are mapped back to the input space thanks to the invertibility and stability of our designed GCNs. Extensive experiments, conducted on two challenging skeleton-based recognition tasks, show the outperformance of our label-efficient method compared to the related work.

2 PROPOSED DISPLAY MODEL

Our proposed AL solution consists of two principal building blocks: *display* and *learning* models. The *former* aims at designing an acquisition function probing an oracle about the labels of the most informative data, whilst the *latter* seeks to retrain a label-frugal classifier accordingly. These two steps are iteratively applied till reaching enough classification accuracy or exhausting a predefined labeling budget. Let $\mathcal{U} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^p$ be a pool of unlabeled data; at each AL iteration $t \in \{0, \dots, T - 1\}$, a subset \mathcal{D}_t — referred to as *display* — is built from \mathcal{U} (following the model in section 2.1), and used to query the oracle about its labels \mathcal{Y}_t . Then a classifier f_t is trained on $\cup_{k=0}^t (\mathcal{D}_k, \mathcal{Y}_k)$.

Our first contribution (introduced in section 2.1) is based on a novel display model that builds in a *flexible way* displays instead of sampling fixed ones from \mathcal{U} .

2.1 Display model design

The principle of our method consists in designing the most diverse, representative and uncertain data that challenge (the most) the current classifier f_t , leading to a better re-estimate of f_{t+1} in the subsequent iteration $t + 1$ of active learning. We consider a probabilistic framework that *builds* the subsequent display \mathcal{D}_{t+1} (denoted for short as \mathcal{D}) instead of sampling \mathcal{D} from \mathcal{U} . Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ (resp. $\mathbf{D} \in \mathbb{R}^{p \times K}$) be a matrix whose k -th column \mathbf{X}_k (resp. \mathbf{D}_k) corresponds to an element of \mathcal{U} (resp. \mathcal{D}) and $K = |\mathcal{D}|$. In order to obtain the display \mathbf{D} , our proposed framework assigns for each \mathbf{D}_k , a conditional probability distribution measuring the memberships $\{\mu_{ik}\}_i$ as the contribution

of each $\mathbf{x}_i \in \mathcal{U}$ in shaping \mathbf{D}_k . These memberships $\mu = \{\mu_{ik}\}_{ik}$ and the display \mathbf{D} are found by minimizing the following constrained objective function

$$\begin{aligned} \min_{\mu \in \Omega, \mathbf{D}} \quad & \text{tr}(\mu d(\mathbf{X}, \mathbf{D})^\top) + \alpha \sum_{k,k'}^{K,N} \exp\left(-\frac{1}{\sigma} \|\mathbf{D}_k - \mathbf{H}_{k'}\|_2^2\right) \\ & + \beta \text{tr}(\mathbf{D}^\top \mathbf{D}) + \gamma \text{tr}(\mu^\top \log \mu), \end{aligned} \quad (1)$$

being $\Omega = \{\mu : \mu \geq 0; \mathbf{1}_n^\top \mu = \mathbf{1}_K^\top\}$ a convex set that constrains μ to be column-stochastic (i.e., each column as a conditional probability distribution), $^\top$ denotes the transpose, and $\mathbf{1}_K, \mathbf{1}_n$ are two vectors of K and n ones respectively. The first term of Eq. 1 encodes the representativeness of the designed exemplars in \mathbf{D} , aiming to minimize the discrepancy between these exemplars and the original distribution of data in \mathcal{U} . It also serves to constrain the oracle's annotations only on realistically designed exemplars, thereby ensuring relevant annotations and also preventing the selection of trivial or meaningless exemplars. The second term of Eq. 1 captures diversity of \mathbf{D} ; this term seeks to maximize the difference between the N previously and the K currently designed exemplars (resp. matrices \mathbf{H} and \mathbf{D}), and enforces the new ones to be as far as possible from the previous ones.

The third term of Eq. 1 acts as an equilibrium criterion measuring the uncertainty associated with exemplars in \mathbf{D} ; in other words, it encourages exemplars to lie on the decision boundaries of the learned classifiers, and it also acts as a regularizer on \mathbf{D} . Minimizing this term effectively identifies exemplars which are inherently ambiguous, and targeting annotations on these highly uncertain data is crucial to reduce model ambiguity and to speedup convergence to well-defined decision functions. Finally, the fourth term corresponds to a regularizer on μ which considers that without any a priori on the three other terms, the conditional probabilities $\mu = \{\mu_{ik}\}_{ik}$ should be flat. All the aforementioned terms are weighted by $\alpha, \beta, \gamma \geq 0$ whose setting is described subsequently.

Proposition 1. *The optimality conditions of Eq. 1 leads to the solution as the fixed-point of*

$$\begin{aligned} \mu^{(\tau+1)} &:= \hat{\mu}^{(\tau+1)} \mathbf{diag}(\mathbf{1}_n^\top \hat{\mu}^{(\tau+1)})^{-1} \\ \mathbf{D}^{(\tau+1)} &:= \hat{\mathbf{D}}^{(\tau+1)} (\mathbf{diag}(\mathbf{1}_n^\top \mu^{(\tau)}) + \beta \mathbf{I})^{-1}, \end{aligned} \quad (2)$$

being $\hat{\mu}^{(\tau+1)}, \hat{\mathbf{D}}^{(\tau+1)}$ respectively

$$\begin{aligned} & \exp\left\{-\frac{1}{\gamma} d(\mathbf{X}, \mathbf{D}^{(\tau)})\right\}, \\ & \mathbf{X} \mu^{(\tau)} - \frac{2\alpha}{\sigma} (\mathbf{D}^{(\tau)} \mathbf{diag}(\mathbf{1}'_N \mathbf{S}) - \mathbf{H} \mathbf{S}), \end{aligned} \quad (3)$$

where \mathbf{S} equates (with $\mathbf{D}^{(\tau)}$ written for short as \mathbf{D})

$$\exp\left\{-\frac{1}{\sigma} (\mathbf{1}_N \mathbf{diag}(\mathbf{D}^\top \mathbf{D})^\top + \mathbf{diag}(\mathbf{H}^\top \mathbf{H}) \mathbf{1}_K^\top - 2\mathbf{H}^\top \mathbf{D})\right\}, \quad (4)$$

here \mathbf{S} is a similarity matrix between \mathbf{D} and \mathbf{H} , $\mathbf{1}_N$ is a vector of N ones, and \mathbf{diag} maps a vector to a diagonal matrix.

In view of space, details of the proof are omitted and follow the optimality conditions of Eq. 1's gradient. More importantly, the solution of μ in Eq. 3 shows that low distances lead to high memberships of the input data in \mathbf{X} to the underlying exemplars in \mathbf{D} , and vice versa, whereas the solution of \mathbf{D} shows that each exemplar \mathbf{D}_k is defined as a combination of two terms: the first one as a normalized¹ linear combination of actual data weighted by their memberships to \mathbf{D}_k

1. thanks to the column-stochasticity of μ .

whilst the second term disrupts further \mathbf{D}_k to make it as different as possible from the previously designed exemplars in \mathbf{H} (depending on the setting of α). Note that $\mu^{(0)}$ and $\mathbf{D}^{(0)}$ are initially set to random values and, in practice, the procedure converges to an optimal solution (denoted as $\tilde{\mu}$, $\tilde{\mathbf{D}}$) in few iterations. This solution defines the subsequent display \mathcal{D}_{t+1} used to train f_{t+1} . Note also that α and β are set to make the impact of the underlying terms equally proportional, and this corresponds to $\alpha = \frac{1}{KN}$ and $\beta = \frac{1}{Kp}$. In Eq. 3, the hyperparameter σ is set proportionally to α in order to absorb the former by the latter, and thereby reduce the total number of hyperparameters. Finally, since γ acts as scaling factor that controls the shape of the exponential function, its setting is iteration-dependent and proportional to the input of that exponential (i.e., $\log(\hat{\mu}^{(\tau+1)})$), so in practice $\gamma = \frac{1}{nK} \|\log(\hat{\mu}^{(\tau+1)})\|_1$.

Now considering the foregoing AL formulation, two variants of the proposed solution are considered in this paper. The first one finds exemplars using the above formulation directly in the ambient (input) space, while the second one finds the exemplars in the latent space, and maps them back to the ambient space thanks to the invertibility and also stability of the learned GCNs (as shown in section 3). As shown subsequently, relying on invertible and stable GCN mapping leads to an extra gain in AL performances as also shown later through experiments.

3 PROPOSED LEARNING MODEL

As introduced, the success of the aforementioned active learning process is highly reliant on the suitability of the display model. In other words, finding suitable displays in the input space should reflect the distribution of the data in the input space. However, for arbitrary input data distributions the display model, in Eq. 1, may hit a major limitation; input data lying on nonlinear manifolds are challenging to parse in order to guarantee that designed displays still lie on these manifolds. In the sequel of this section, we revisit GCNs and we introduce — as a second contribution — a novel design that makes our trained GCNs invertible and stable.

3.1 A Glimpse on graph convnets

Consider a collection of graphs $\{\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)\}_i$, where \mathcal{V}_i and \mathcal{E}_i represent the nodes and edges \mathcal{G}_i , respectively. For simplicity, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a single graph from this collection. Each graph \mathcal{G} is associated with a signal $\{\psi(v) \in \mathbb{R}^s : v \in \mathcal{V}\}$ and an adjacency matrix \mathbf{A} . Graph Convolutional Networks (GCNs) aim to learn a set of C filters \mathcal{F} that define a convolution operation on the m nodes of \mathcal{G} (where $m = |\mathcal{V}|$) as follows: $(\mathcal{G} * \mathcal{F})_v = g(\mathbf{A} \mathbf{U}^\top \mathbf{W})$. Here, $\mathbf{U} \in \mathbb{R}^{s \times m}$ is the graph signal, $\mathbf{W} \in \mathbb{R}^{s \times C}$ is the matrix of convolutional parameters for the C filters, and $g(\cdot)$ is a nonlinear activation function applied element-wise. In this operation, the input signal \mathbf{U} is projected using the adjacency matrix \mathbf{A} , effectively aggregating the signals from the neighbors of each node v . The entries of \mathbf{A} can be either handcrafted or learned. Hence, $(\mathcal{G} * \mathcal{F})_v$ can be viewed as a two-layer (attention and convolutional) block. The first layer aggregates signals from the neighborhood $\mathcal{N}(\mathcal{V})$ of each node by multiplying \mathbf{U} with \mathbf{A} , while the second layer performs the convolution by multiplying the resulting aggregates with the C filters in \mathbf{W} .

3.2 Invertibility & Stability

In what follows, we formally subsume a given GCN as a multi-layered neural network f whose weights are defined as $\theta = \{\mathbf{W}_1, \dots, \mathbf{W}_L\}$, being L its depth, $\mathbf{W}_\ell \in \mathbb{R}^{d_{\ell-1} \times d_\ell}$ its ℓ^{th} layer weight tensor, and d_ℓ the dimension of ℓ . The output of a given layer ℓ is defined as $\phi^\ell = g_\ell(\mathbf{W}_\ell^\top \phi^{\ell-1})$, $\ell \in \{2, \dots, L\}$, with g_ℓ an activation function; without a loss of generality, we omit the bias in the definition of ϕ^ℓ .

In this section, we are interested in designing invertible and stable networks. Invertibility (bijection) of $f : \mathbb{R}^p \rightarrow \mathbb{R}^q$ guarantees the existence of a *one-to-one* mapping from \mathbb{R}^p to \mathbb{R}^q (with necessarily $p = q$)² so as no distinct network's inputs ϕ_1^1, ϕ_2^1 map to the same output ϕ^L , and for every output ϕ^L , there exists at least one input ϕ^1 such that $f(\phi^1) = \phi^L$. Stability pushes invertibility "one step further" to guarantee that f^{-1} — when evaluated on a given targeted latent distribution (e.g., gaussian) — does not diverge from the ambient (input) distribution.

Definition 1 (Stability). *An invertible network $f : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is called bi-Lipschitzian (or KM-Lipschitzian), if f is K -Lipschitzian and its inverse f^{-1} is M -Lipschitzian.*

In general, making both K and M small for any given nonlinear function is challenging [39]. However, considering our following network f 's design, it becomes possible under specific conditions to make both K and M small (namely close to 1 as a result of our subsequent proposition).

Proposition 2. *Provided that (i) the entrywise activations $\{g_\ell(\cdot)\}_{\ell=2}^L$ are bijective in \mathbb{R}^p , (ii) $l \leq |g'_\ell(\cdot)| \leq u$, and (iii) all the weight matrices in θ orthonormal, then the network f is invertible in \mathbb{R}^p , and KM-Lipschitzian with $K = u^{L-1}$ and $M = (1/l)^{L-1}$.*

Details of the proof are given in the appendix. More importantly, following the above proposition, when f is invertible in \mathbb{R}^p , then one may derive $f^{-1}(\phi^L) = \phi^1$ being $\phi^{\ell-1} = (\mathbf{W}_\ell^\top)^{-1} g_\ell^{-1}(\phi^\ell)$, and when l and u are close to 1, then $K, M \approx 1$ meaning that both f and f^{-1} are 1-Lipschitzian [39] so any slight update of exemplars in the latent space (with the fixed-point iteration in Eq. 2) will also result into a slight update of these exemplars in the ambient space when applying f^{-1} . This eventually leads to stable exemplar design in the ambient space, i.e., they follow the actual distribution of data manifold. As a Lipschitz constant of f is $\prod_\ell \|\mathbf{W}_\ell\|_2 |g'_\ell|$, and for f^{-1} is $\prod_\ell \|(\mathbf{W}_\ell^\top)^{-1}\|_2 |g_\ell^{-1}'|$ (see proof in appendix), the sufficient conditions that guarantee that both f and f^{-1} are Lipschitzian (with $K, M \approx 1$) corresponds to (1) $\|\mathbf{W}_\ell\|_2 \approx 1$, and (2) $l, u \approx 1$ with $l < u$. Hence, by design, conditions (1)+(2) could be satisfied by choosing the slope of the activation functions to be close to one (in practice $u = 0.99$ and $l = 0.95$ corresponding respectively to the positive and negative slopes of the leaky-ReLU), and also by constraining all the weight matrices to be *orthonormal* which also guarantees their invertibility. This is obtained by adding a regularization term, to the cross-entropy (CE) loss, when training GCNs, as

$$\min_{\{\mathbf{W}_\ell\}_\ell} \text{CE}(f; \{\mathbf{W}_\ell\}_\ell) + \lambda \sum_\ell \|\mathbf{W}_\ell^\top \mathbf{W}_\ell - \mathbf{I}\|_F, \quad (5)$$

here \mathbf{I} stands for identity, $\|\cdot\|_F$ denotes the Frobenius norm and $\lambda > 0$ (with $\lambda = \frac{1}{p}$ in practice³); in particular, when $\mathbf{W}_\ell^\top \mathbf{W}_\ell - \mathbf{I} = 0$, then $\mathbf{W}_\ell^{-1} = \mathbf{W}_\ell^\top$ and $\|\mathbf{W}_\ell\|_2 = \|\mathbf{W}_\ell^{-1}\|_2 = 1$. With this formulation, the learned GCNs are guaranteed to be discriminative, invertible and stable.

4 EXPERIMENTS

This section evaluates the performance of baseline and label-frugal GCNs for skeleton-based recognition using the SBU Interaction [1] and First Person Hand Action (FPHA) [37] datasets. The SBU Interaction dataset, captured using the Microsoft Kinect, comprises 282 skeleton sequences of two interacting individuals performing one of eight predefined actions. Each interaction is represented by two 15-joint skeletons, with each joint's 3D coordinates acquired across the video

2. As the output of f depends on the number of classes, a simple trick consists in adding fictitious outputs to match any targeted dimension (similarly for other layers).

3. Note that at frugal data regimes, this optimization problem is easy to minimize as the cross entropy term involves few labeled data, so it is enough to set λ to small values in order to guarantee the minimization of both terms.

frames. Evaluation follows the original train-test split defined in [1]. The FPHA dataset contains 1175 skeleton sequences spanning 45 diverse hand-action categories, performed by six individuals across three different scenarios. The actions exhibit significant variations in style, speed, scale, and viewpoint. Each skeleton consists of 21 hand joints, also represented by sequences of 3D coordinates. Following [37], we evaluate performance using the 1:1 setting, with 600 sequences for training and 575 for testing. For both datasets, we report the average classification accuracy across all action categories.

Input graphs. We represent each skeleton sequence $\{S^t\}_t$ as a series of 3D joint coordinates $S^t = \{\hat{p}_j^t\}_j$ at each frame t . A joint’s trajectory $\{\hat{p}_j^t\}_t$ tracks its movement across frames. Our input graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ comprises nodes \mathcal{V} with each one $v_j \in \mathcal{V}$ representing a trajectory $\{\hat{p}_j^t\}_t$, and each edge $(v_j, v_i) \in \mathcal{E}$ connects spatially neighboring trajectories. To process each trajectory, we divide its duration into M_c equal temporal chunks (with $M_c = 4$ in practice). Joint coordinates $\{\hat{p}_j^t\}_t$ are assigned to these chunks based on their timestamps, and the average coordinates within each chunk are concatenated to form a trajectory descriptor (denoted as $\psi(v_j) \in \mathbb{R}^s$) of size $s = 3M_c$. This chunking approach preserves temporal information while making the representation independent of frame rate and sequence duration.

Implementation details & baseline GCNs. All GCNs have been trained using the Adam optimizer for 2700 epochs. The batch size is 200 for SBU and 600 for FPHA. A momentum of 0.9 is used, and the global learning rate ν is dynamically adjusted based on the loss Eq. 5’s rate of change. Specifically, ν is decreased by a factor of 0.99 when the rate of loss change *increased*, and increased by a factor of 1/0.99 *otherwise*. Training is performed on a GeForce GTX 1070 GPU with 8 GB memory. No dropout or data augmentation techniques are employed. For SBU, the architecture of our GCN comprises three “mono-head attentions + (8 filters) convolutions” layers followed by one fully connected and a classification layer. The GCN architecture for FPHA is relatively heavier (for a GCN), and differs from SBU in the number of convolutional filters (16 filters instead of 8). Both architectures, on the SBU and the FPHA benchmarks, are accurate (see Tables. 1-2), and our goal is to make them label-efficient while being *as close as possible* to their initial accuracy.

Performances, comparison & ablation. Tables 3-4 show a comparison and an ablation study of our method both on the SBU and the FPHA datasets. According to the observed results, when our display model is run on the ambient space, the accuracy is relatively high, and sometimes overtakes comparative display selections by a noticeable margin. When using the latent space, we observe a further gain of our method. This clearly shows the impact of our model and its extra gain when combined with the latent space. Extra comparison of our method against other display selection strategies also shows a substantial gain. Indeed, our method is compared against different strategies used as display selection (instead of our proposed display model), namely random, diversity [60] and uncertainty [64], all with our GCN learning. From the observed results in tables 3-4, the impact of our method is significant for different settings and for equivalent labeling rates. We also observe that random is already performant (as widely known, see for instance [52] and references therein) mainly when the sample size is relatively large (45%). In contrast, with relatively smaller sizes (15%), random is less performant so more principled selection strategies are required.

Note that random and diversity are not capable of sufficiently refining classifications, whereas uncertainty allows us to refine classifications but without enough diversity. Besides, all these comparative methods suffer at some extent from the rigidity of the selected displays (which are taken from a fixed pool). Our display model, in contrast, allows us to learn flexible exemplars, constrained in the latent space of the proposed invertible and stable GCNs, with a positive impact on performances including at frugal labeling regimes.

Method	Accuracy (%)
Raw Position [1]	49.7
Joint feature [7]	86.9
CHARM [8]	86.9
H-RNN [11]	80.4
ST-LSTM [12]	88.6
Co-occurrence-LSTM [13]	90.4
STA-LSTM [21]	91.5
ST-LSTM + Trust Gate [12]	93.3
VA-LSTM [25]	97.6
GCA-LSTM [18]	94.9
Riemannian manifold. traj [34]	93.7
DeepGRU [15]	95.7
RHCN + ACSC + STUFE [20]	98.7
Our baseline GCN	98.4

TABLE 1

Comparison of our baseline GCN (not label-efficient) against related work on the SBU database.

Method	Color	Depth	Pose	Accuracy (%)
2-stream-color [114]	✓	✗	✗	61.56
2-stream-flow [114]	✓	✗	✗	69.91
2-stream-all [114]	✓	✗	✗	75.30
HOG2-dep [2]	✗	✓	✗	59.83
HOG2-dep+pose [2]	✗	✓	✓	66.78
HON4D [3]	✗	✓	✗	70.61
Novel View [4]	✗	✓	✗	69.21
1-layer LSTM [13]	✗	✗	✓	78.73
2-layer LSTM [13]	✗	✗	✓	80.14
Moving Pose [6]	✗	✗	✓	56.34
Lie Group [29]	✗	✗	✓	82.69
HBRNN [11]	✗	✗	✓	77.40
Gram Matrix [33]	✗	✗	✓	85.39
TF [37]	✗	✗	✓	80.69
JOULE-color [129]	✓	✗	✗	66.78
JOULE-depth [129]	✗	✓	✗	60.17
JOULE-pose [129]	✗	✗	✓	74.60
JOULE-all [129]	✓	✓	✓	78.78
Huang et al. [35]	✗	✗	✓	84.35
Huang et al. [36]	✗	✗	✓	77.57
HAN [24]	✗	✗	✓	85.74
Our baseline GCN	✗	✗	✓	88.17

TABLE 2

Comparison of our baseline GCN (not label-efficient) against related work on the FPHA database.

Labeling rates	Accuracy (%)	Observation
45%	100%	Baseline GCN (not label-efficient)
	<u>89.23</u>	wo display model (random display)
	<u>89.23</u>	+ display model + ambient space (our)
	93.84	+ display model + latent space (our)
	67.69	uncertainty (margin-based)
30%	83.07	diversity (coreset-based)
	80.00	wo display model (random display)
	<u>86.15</u>	+ display model + ambient space (our)
	87.69	+ display model + latent space (our)
	61.53	uncertainty (margin-based)
15%	83.07	diversity (coreset-based)
	<u>69.23</u>	wo display model (random display)
	75.38	+ display model + ambient space (our)
	75.38	+ display model + latent space (our)
	56.92	uncertainty (margin-based)
	66.15	diversity (coreset-based)

TABLE 3

This table shows detailed performances and ablation study on SBU for different labeling rates. Here “wo” stands for “without”. Best results are shown in bold and second best results underlined.

Labeling rates	Accuracy (%)	Observation
45%	100%	Baseline GCN (not label-efficient)
	<u>75.47</u>	wo display model (random display)
	72.52	+ display model + ambient space (our)
	75.65	+ display model + latent space (our)
	63.30	uncertainty (margin-based)
30%	70.26	diversity (coreset-based)
	67.47	wo display model (random display)
	61.21	+ display model + ambient space (our)
	<u>63.65</u>	+ display model + latent space (our)
	56.17	uncertainty (margin-based)
15%	62.08	diversity (coreset-based)
	40.52	wo display model (random display)
	45.21	+ display model + ambient space (our)
	49.21	+ display model + latent space (our)
	41.73	uncertainty (margin-based)
	<u>46.26</u>	diversity (coreset-based)

TABLE 4
Same caption as table 3, but for FPHA.

5 CONCLUSION

We introduce in this paper a label-efficient method for skeleton-based action recognition built upon graph convolutional networks (GCNs). The strength of our contribution resides in the design of a new acquisition function as the optimum of an objective function mixing representativity, diversity and uncertainty. We further enhance this design by making our GCNs stable

and invertible thereby transforming input data into latent and more readily learnable spaces. The efficacy and superior performance of our proposed method are demonstrated through extensive experiments on two challenging skeleton-based recognition datasets.

APPENDIX

Sketch of the Proof (Proposition 2). Given a metric space (A, d_A) , where d_A denotes the metric on the set A (by default d_A is taken as ℓ_2 and A as \mathbb{R}^p); considering a subsumed version of our GCNs, and using the Lipschitz continuity, one may write

$$\begin{aligned} d_A(f(\phi_1^1), f(\phi_2^1)) &= d_A(g_L(\mathbf{W}_L^\top \phi_1^{L-1}), g_L(\mathbf{W}_L^\top \phi_2^{L-1})) \\ &\leq u d_A(\mathbf{W}_L^\top \phi_1^{L-1}, \mathbf{W}_L^\top \phi_2^{L-1}) \\ &\leq u \|\mathbf{W}_L\|_A d_A(\phi_1^{L-1}, \phi_2^{L-1}) \\ &\leq u^{L-1} \|\mathbf{W}_L\|_A \dots \|\mathbf{W}_2\|_A d_A(\phi_1^1, \phi_2^1), \end{aligned}$$

being ϕ_1^1, ϕ_2^1 two network inputs. As $\{\mathbf{W}_\ell\}_\ell$ are orthonormal, it follows that $\|\mathbf{W}_\ell\|_A = 1$, and $d_A(f(\phi_1^1), f(\phi_2^1)) \leq K d_A(\phi_1^1, \phi_2^1)$ with $K = u^{L-1}$.

Similarly for f^{-1} , given an output ϕ^L , we have $f^{-1}(\phi^L) = \phi^1$ with $\phi^{\ell-1} = (\mathbf{W}_\ell^\top)^{-1} g_\ell^{-1}(\phi^\ell)$. Hence, considering two network outputs ϕ_1^L, ϕ_2^L one may write

$$\begin{aligned} d_A(f^{-1}(\phi_1^L), f^{-1}(\phi_2^L)) &= d_A((\mathbf{W}_2^\top)^{-1} g_2^{-1}(\phi_1^2), (\mathbf{W}_2^\top)^{-1} g_2^{-1}(\phi_2^2)) \\ &\leq \|(\mathbf{W}_2^\top)^{-1}\|_A d_A(g_2^{-1}(\phi_1^2), g_2^{-1}(\phi_2^2)) \\ &\leq \|(\mathbf{W}_2^\top)^{-1}\|_A (1/l) d_A(\phi_1^2, \phi_2^2) \\ &\leq \prod_\ell \|(\mathbf{W}_\ell^\top)^{-1}\|_A (1/l)^{L-1} d_A(\phi_1^L, \phi_2^L). \end{aligned}$$

As $\{\mathbf{W}_\ell\}_\ell$ are orthonormal, it follows that $\|(\mathbf{W}_\ell^\top)^{-1}\|_A = 1$, and $d_A(f^{-1}(\phi_1^L), f^{-1}(\phi_2^L)) \leq M d_A(\phi_1^L, \phi_2^L)$ with $M = (1/l)^{L-1}$ \square

REFERENCES

- [1] K. Yun, et al. *Two-person interaction detection using body-pose features and multiple instance learning*. In CVPRW, 2012.
- [2] E. Ohn-Bar and Trivedi, M. M. *Hand gesture recognition in real time for automotive interfaces* IEEE TITS, 15(6), 2014.
- [3] O. Oreifej and Z. Liu. *Hon4d: Histogram of oriented 4d normals for activity recognition from depth seq*. In CVPR, 2013.
- [4] H. Rahmani and A. Mian. *3d action recognition from novel viewpoints*. In CVPR, 2016.
- [5] P. Vo and H. Sahbi. "Transductive kernel map learning and its application to image annotation." BMVC. 2012.
- [6] M. Zanfir, et al. *The moving pose: An efficient 3d kinematics descriptor for low-latency action rec and det*. In ICCV, 2013.
- [7] Y. Ji, et al. *Interactive body part contrast mining for human interaction recognition*. In ICMEW, 2014.
- [8] W. Li, et al. *Category-blind human action recognition: A practical recognition system*. In ICCV, 2015.
- [9] J.-F. Hu, et al. *Jointly learning heterogeneous features for rgb-d activity recognition*. In CVPR, 2015.
- [10] Q. Oliveau and H. Sahbi. "Learning attribute representations for remote sensing ship category classification." IEEE JSTARS 10.6 (2017): 2830-2840.
- [11] Y. Du, et al. *Hierarchical recurrent neural network for skeleton based action recognition*. In CVPR, 2015.
- [12] J. Liu, et al. *Spatio-temporal lstm with trust gates for 3d human action recognition*. In ECCV, 2016.
- [13] W. Zhu, et al. *Co-occurrence feature learning for skeleton based act rec using reg deep lstm networks*. In AAAI, 2016.
- [14] H. Sahbi. "Interactive satellite image change detection with context-aware canonical correlation analysis." IEEE GRSL, (14)5, 2017.
- [15] M. Maghoumi and J-J. LaViola. *Deepgru: Deep gesture recognition utility*. In ISVC, 2019.
- [16] H. Sahbi. "Relevance feedback for satellite image change detection." IEEE ICASSP, 2013.

- [17] S. Zhang, et al. *On geometric features for skeleton-based action recognition using multilayer lstm nets*. In WACV, 2017.
- [18] J. Liu, et al. *Skeleton-based human action recognition with global context-aware attention lstm networks*. IEEE TIP, 2017.
- [19] F. Yuan, G-S. Xia, H. Sahbi, V. Prinet. Mid-level features and spatio-temporal context for activity recognition. *Pattern Recognition* 45 (12), 4182-4191
- [20] S. Li, et al. *Global co-occ feature learning and active coordinate sys conversion for skeleton-based act rec*. In WACV, 2020.
- [21] S. Song, et al. *An end-to-end spatio-temporal attention model for human act rec from skeleton data*. In AAAI, 2017.
- [22] N. Bourdis, D. Marraud and H. Sahbi. "Constrained optical flow for aerial image change detection." in IEEE IGARSS, 2011.
- [23] L. Wang, H. Sahbi. Bags-of-daglets for action recognition. In IEEE ICIP 2014.
- [24] J. Liu, et al. *Han: An efficient hierarchical self-attention network for skeleton-based gesture recognition*. arXiv, 2021.
- [25] P. Zhang, et al. *View adaptive recurrent neural nets for high perf human action rec from skeleton data*. In ICCV, 2017.
- [26] N. Bourdis, D. Marraud, and H. Sahbi, Spatio-temporal interaction for aerial video change detection, in IGARSS, 2012, pp. 2253–2256
- [27] C. Feichtenhofer, et al. *Convolutional two-stream network fusion for video action recognition*. In CVPR, 2016.
- [28] H. Sahbi. "Coarse-to-fine deep kernel networks." IEEE ICCV-W, 2017.
- [29] R. Vemulapalli, et al. *Human action recognition by representing 3d skeletons as points in a lie group*. In CVPR, 2014.
- [30] L. Wang and H. Sahbi. *Directed acyclic graph kernels for action recognition*. In ICCV, 2013.
- [31] F. Yuan, et al. *Mid-level features and spatio-temporal context for activity recognition*. In Pattern Recognition, 45(12), 2012.
- [32] A. Mazari and H. Sahbi. *MLGCN: Multi-Laplacian graph conv networks for human action recognition*. In BMVC, 2019.
- [33] X. Zhang, et al. *Efficient temp seq comp and classif using gram matrix embeddings on a riemannian manifold*. In CVPR, 2016.
- [34] A. Kacem, et al. *A novel geometric framework on gram matrix trajectories for human behavior under*. IEEE TPAMI, 2018.
- [35] Z. Huang and L. Van Gool. *A riemannian network for spd matrix learning*. In AAAI, 2017.
- [36] Z. Huang, et al. *Building deep networks on grassmann manifolds*. In AAAI, 2018.
- [37] G. Garcia-Hernando and T-K. Kim. *Transition forests: Learning discriminative temporal transitions for action recognition and detection*. In CVPR, 2017.
- [38] H. Sahbi. *Learning connectivity with graph convolutional networks*. In ICPR, 2020.
- [39] J. Heinonen. *Lectures on Lipschitz Analysis*. Springer, 2005.
- [40] C. Shorten and T.M. Khoshgoftaar. *A survey on image data augmentation for deep learning*. J. Big Data, 6(1), 2019.
- [41] C-A. Brust, et al. *Active learning for deep object detection*. arXiv:1809.09875, 2018.
- [42] K. Wang, et al. *Cost-effective object detection: Active sample mining with switchable selection criteria*. In TNNLS, 2018.
- [43] Y. Gal and Z. Ghahramani. *Dropout as a bayesian approximation: Representing model uncer in deep learning*. ICML 2016.
- [44] D. Yoo I-S. Kweon. *Learning loss for active learning*. In CVPR, 2019.
- [45] N. Bourdis, D. Marraud and H. Sahbi. "Camera pose estimation using visual servoing for aerial video change detection." IEEE IGARSS 2012.
- [46] P. Hemmer, et al. *Deal: Deep evidential active learning for image classification*. In ICMLA, 2020.
- [47] H. Sahbi and N. Boujemaa. "Robust matching by dynamic space warping for accurate face recognition." Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205). Vol. 1. IEEE, 2001.
- [48] A. Culotta and A. McCallum. *Reducing labeling effort for structured prediction tasks*. In AAAI, 2005.
- [49] Y-C. Wu. *Active learning based on div max*. In AMM, 2013.
- [50] Sabrina Tollari, Philippe Mulhem, Marin Ferecatu, Hervé Glotin, Marcin Detyniecki, Patrick Gallinari, H. Sahbi, and Zhong-Qiu Zhao. "A comparative study of diversity methods for hybrid text and image retrieval approaches." In Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 585-592. Springer, Berlin, Heidelberg, 2008.
- [51] S. Agarwal, et al. *Contextual diversity for active learning*. In ECCV, 2020.

- [52] B. Settles. *Active learning literature survey*. University of Wisconsin–Madison, 2009.
- [53] H. Sahbi. A particular Gaussian mixture model for clustering and its application to image retrieval. *Soft Computing* 12 (7), 667-676
- [54] O. Yehuda, et al. *Active learning through a covering lens*. In NeurIPS, 2022.
- [55] R. Caramalau, et al. *Self-supervised Active Learning for Image Classification*. In BMVC, 2022.
- [56] M. Jiu and H. Sahbi. "Laplacian deep kernel learning for image annotation." IEEE ICASSP, 2016.
- [57] Y. Wu, et al. (2019). *Active learning for graph neural networks via node feature propagation*. In arXiv, 2019.
- [58] S-M. Kye, et al. *TiDAL: Learning training dynamics for active learning*. In ICCV, 2023.
- [59] M. Ferecatu and H. Sahbi. "TELECOM ParisTech at ImageClefphoto 2008: Bi-Modal Text and Image Retrieval with Diversity Enhancement." CLEF (Working Notes). 2008.
- [60] Y. Kim and B. Shin. *In defense of core-set: A density-aware core-set selection for active learning*. In KDD, 2022.
- [61] H. Sahbi, Jean-Yves Audibert, Jaonary Rabarisoa, and Renaud Keriven. "Context-dependent kernel design for object matching and recognition." In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8. IEEE, 2008.
- [62] S. Jung, et al. *A simple yet powerful deep active learning with snapshots ensembles*. In ICLR, 2023.
- [63] H. Sahbi, Jean-Yves Audibert, and Renaud Keriven. "Graph-cut transducers for relevance feedback in content based image retrieval." 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007.
- [64] Z. Xu, et al. *Hierarchical point-based active learning for semi-supervised point cloud semantic segmentation*. In ICCV, 2023.
- [65] M. Jiu and H. Sahbi. "Deep representation design from deep kernel networks." *Pattern Recognition* 88 (2019): 447-457.
- [66] P. Simeoni, et al. *Rethinking deep active learning: Using unlabeled data at model training*. In ICPR, 2021.
- [67] C. Wang, et al. *Teaching an active learner with contrastive examples*. In NeurIPS 2021.
- [68] H. Sahbi, S. Deschamps, A. Stoian. Frugal Learning for Interactive Satellite Image Change Detection. IEEE IGARSS, 2021.
- [69] D. Li, et al. *A survey on deep active learning: Recent advances and new frontiers*. In IEEE TNNLS, 2024.
- [70] S. Thiemert, H. Sahbi, and M. Steinebach. "Applying interest operators in semi-fragile video watermarking." *Security, Steganography, and Watermarking of Multimedia Contents VII*. Vol. 5681. SPIE, 2005.
- [71] Liu, M., Liu, H., Hu, Q., Ren, B., Yuan, J., Lin, J., and Wen, J. (2025). 3D Skeleton-Based Action Recognition: A Review. arXiv preprint arXiv:2506.00915.
- [72] Chung, J. L., Ong, L. Y., and Leow, M. C. (2025). A Systematic Literature Review of Optimization Methods in Skeleton-Based Human Action Recognition. *IEEE Access*.
- [73] H. Sahbi. "Imageclef annotation with explicit context-aware kernel maps." *International Journal of Multimedia Information Retrieval* 4.2 (2015): 113-128.
- [74] Zhao, L., Lin, Z., Sun, R., and Wang, A. (2024). A Review of State-of-the-Art Methodologies and Applications in Action Recognition. *Electronics*, 13(23), 4733.
- [75] Zhang, J., Lin, L., Yang, S., and Liu, J. (2024). Self-Supervised Skeleton-Based Action Representation Learning: A Benchmark and Beyond. arXiv preprint arXiv:2406.02978.
- [76] L. Wang and H. Sahbi. "Bags-of-daglets for action recognition." 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014.
- [77] Wang, X., Jiang, X., Zhao, Z., Wang, K., and Yang, Y. (2025). Exploring interaction: Inner-outer spatial-temporal transformer for skeleton-based mutual action recognition. *Neurocomputing*, 636, 130007.
- [78] Chen, Z., Huang, W., Liu, H., Wang, Z., Wen, Y., and Wang, S. (2024). ST-TGR: Spatio-temporal representation learning for skeleton-based teaching gesture recognition. *Sensors*, 24(8), 2589.
- [79] H. Sahbi. "Lightweight Connectivity In Graph Convolutional Networks For Skeleton-Based Recognition." 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021.
- [80] Li, X., Qiu, Y. K., Peng, Y. X., and Zheng, W. S. (2024, May). Patch-Based Privacy Attention for Weakly-Supervised Privacy-Preserving Action Recognition. In 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) (pp. 1-9). IEEE.
- [81] Cunling, B. I. A. N., Weigang, L. Y. U., and Wei, F. E. N. G. (2024). Skeleton-Based Human Action Recognition: History, Status and Prospects. *Journal of Computer Engineering and Applications*, 60(20).

- [82] H. Sahbi. "CNRS-TELECOM ParisTech at ImageCLEF 2013 Scalable Concept Image Annotation Task: Winning Annotations with Context Dependent SVMs." CLEF (Working Notes). 2013.
- [83] Habib, M. K., Yusuf, O., and Moustafa, M. (2025). Skeleton-Based Real-Time Hand Gesture Recognition Using Data Fusion and Ensemble Multi-Stream CNN Architecture. *Technologies*, 13(11), 484.
- [84] Liu, X., and Gao, B. (2025). Individual contribution based spatial-temporal attention on skeleton sequences for human interaction recognition. *IEEE Access*.
- [85] Sahbi, H., and N. Boujemaa. "Robust face recognition using dynamic space warping." International Workshop on Biometric Authentication. Springer, Berlin, Heidelberg, 2002.
- [86] Peng, K., Fu, J., Yang, K., Wen, D., Chen, Y., Liu, R., ... and Roitberg, A. (2024, September). Referring atomic video action recognition. In European Conference on Computer Vision (pp. 166-185). Cham: Springer Nature Switzerland.
- [87] Li, M., Wu, Y., Sun, Q., and Yang, W. (2024). Two-Stream Proximity Graph Transformer for Skeletal Person-Person Interaction Recognition With Statistical Information. *IEEE Access*.
- [88] H. Sahbi, D. Geman. A Hierarchy of Support Vector Machines for Pattern Detection. *Journal of Machine Learning Research* 7 (10).
- [89] Bukht, T. F. N., Jalal, A., and Rahman, H. (2024, November). Enhanced Human Interaction Recognition Framework using Pyramid Matching and Deep Neural Network. In 2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE) (pp. 1-6). IEEE.
- [90] Kumar, R., and Kumar, S. (2024). A survey on intelligent human action recognition techniques. *Multimedia Tools and Applications*, 83(17), 52653-52709.
- [91] S. Thiemert, H. Sahbi, and M. Steinebach. "Using entropy for image and video authentication watermarks." *Security, Steganography, and Watermarking of Multimedia Contents VIII*. Vol. 6072. SPIE, 2006.
- [92] Sun, J., Huang, L., Wang, H., Zheng, C., Qiu, J., Islam, M. T., ... and Black, M. J. (2024). Localization and recognition of human action in 3D using transformers. *Communications Engineering*, 3(1), 125.
- [93] Chen, H., Zendehdel, N., Leu, M. C., Moniruzzaman, M., Yin, Z., and Hajmohammadi, S. (2024, July). Repetitive action counting through joint angle analysis and video transformer techniques. In International Symposium on Flexible Automation (Vol. 87882, p. V001T08A003). American Society of Mechanical Engineers.
- [94] H. Sahbi. Coarse-to-fine support vector machines for hierarchical face detection. Diss. PhD thesis, Versailles University, 2003.
- [95] Purkar, S., Patil, S., Kale, V., and Kadam, B. D. (2024, June). Video activity classification: a comparative analysis and deep learning based implementation. In 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS) (pp. 1-6). IEEE.
- [96] N. Boujemaa, F. Fleuret, V. Gouet, and H. Sahbi. "Visual content extraction for automatic semantic annotation of video news." In the proceedings of the SPIE Conference, San Jose, CA, vol. 6. 2004.
- [97] Askari, F., Yared, C., Ramaprasad, R., Garg, D., Hu, A., and Clark, J. J. (2024). Video interaction recognition using an attention augmented relational network and skeleton data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3225-3234).
- [98] Shin, J., Hassan, N., Miah, A. S. M., and Nishimura, S. (2025). A comprehensive methodological survey of human activity recognition across diverse data modalities. *Sensors*, 25(13), 4028.
- [99] H. Sahbi. "Misalignment resilient cca for interactive satellite image change detection." 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016.
- [100] Banerjee, B., and Baruah, M. (2024). Attention-Based Variational Autoencoder Models for Human-Human Interaction Recognition via Generation. *Sensors*, 24(12), 3922.
- [101] Khean, V., Kim, C., Ryu, S., Khan, A., Hong, M. K., Kim, E. Y., ... and Nam, Y. (2024). Human Interaction Recognition in Surveillance Videos Using Hybrid Deep Learning and Machine Learning Models. *Computers, Materials and Continua*, 81(1).
- [102] T. Napoléon and H. Sahbi. "From 2D silhouettes to 3D object retrieval: contributions and benchmarking." *EURASIP Journal on Image and Video Processing* 2010 (2010): 1-17.
- [103] Wang, H., Cheng, Q., Yu, B., Zhan, Y., Tao, D., Ding, L., and Ling, H. (2024). Free-form composition networks for egocentric action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10), 9967-9978.
- [104] Sachdeva, K., Sandhu, J. K., and Sahu, R. (2024, February). Exploring video event classification: leveraging two-stage neural networks and customized CNN models with UCF-101 and CCV datasets. In 2024 11th international conference on computing for sustainable global development (INDIACoM) (pp. 100-105). IEEE.

- [105] X. Li and H. Sahbi. "Superpixel-based object class segmentation using conditional random fields." 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011.
- [106] Mansouri, A., Elzaar, A., Madani, M., and Bakir, T. (2024). Design and hardware implementation of cnn-gcn model for skeleton-based human action recognition. WSEAS Transactions on Computer Research, 12, 318-327.
- [107] Roy, K. (2024). Multimodal Score Fusion with Sparse Low-rank Bilinear Pooling for Egocentric Hand Action Recognition. ACM Transactions on Multimedia Computing, Communications and Applications, 20(7), 1-22.
- [108] H. Sahbi. Kernel PCA for similarity invariant shape recognition. Neurocomputing 70 (16-18), 3034-3045.
- [109] Tse, T. H. E., Feng, R., Zheng, L., Park, J., Gao, Y., Kim, J., ... and Chang, H. J. (2025, April). Collaborative Learning for 3D Hand-Object Reconstruction and Compositional Action Recognition from Egocentric RGB Videos Using Superquadrics. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 7, pp. 7437-7445).
- [110] Yang, J., Liang, J., Pan, H., Cai, Y., Gao, Q., and Wang, X. (2025). A Unified Framework for Recognizing Dynamic Hand Actions and Estimating Hand Pose from First-Person RGB Videos. Algorithms, 18(7), 393.
- [111] M. Jiu and H. Sahbi. "Semi supervised deep kernel design for image annotation." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.
- [112] Karim, M., Khalid, S., Lee, S., Almutairi, S., Namoun, A., and Abohashrh, M. (2025). Next Generation Human Action Recognition: A Comprehensive Review of State-of-the-Art Signal Processing Techniques. IEEE Access.
- [113] Zhang, Y., Zhang, F., Zhou, Y., and Xu, X. (2024). ACA-Net: adaptive context-aware network for basketball action recognition. Frontiers in Neurorobotics, 18, 1471327.
- [114] M. Jiu and H. Sahbi. "Deep kernel map networks for image annotation." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.
- [115] Wang, R., Wang, Z., Gao, P., Li, M., Jeong, J., Xu, Y., ... and Lu, C. (2025). Real-Time Video-Based Human Action Recognition on Embedded Platforms. ACM Transactions on Embedded Computing Systems, 24(5s), 1-24.
- [116] Zhu, A., Ke, Q., Gong, M., and Bailey, J. (2024). Part-aware unified representation of language and skeleton for zero-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18761-18770).
- [117] H. Sahbi and N. Boujemaa. "From coarse to fine skin and face detection." Proceedings of the eighth ACM international conference on Multimedia. 2000.
- [118] Gunasekara, S. R., Li, W., Yang, J., and Ogunbona, P. O. (2024). Asynchronous joint-based temporal pooling for skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 35(1), 357-366.
- [119] Geng, P., Lu, X., Li, W., and Lyu, L. (2024). Hierarchical aggregated graph neural network for skeleton-based action recognition. IEEE Transactions on Multimedia.
- [120] H. Sahbi and F. Fleuret. Kernel methods and scale invariance using the triangular kernel. Diss. INRIA, 2004.
- [121] Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y., and Sun, Z. (2025). Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 29248-29257).
- [122] Chen, Y., Chen, D., Liu, R., Zhou, S., Xue, W., and Peng, W. (2024). Align before adapt: Leveraging entity-to-region alignments for generalizable video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18688-18698).
- [123] H. Sahbi and F. Fleuret. Scale-invariance of support vector machines based on the triangular kernel. Diss. INRIA, 2002.
- [124] Yang, Y., Zhang, J., Zhang, J., and Tu, Z. (2024). Expressive keypoints for skeleton-based action recognition via skeleton transformation. arXiv preprint arXiv:2406.18011.
- [125] Qu, H., Yan, R., Shu, X., Gao, H., Huang, P., and Xie, G. S. (2025). MVP-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. IEEE Transactions on Multimedia.
- [126] H. Sahbi, J-Y. Audibert, R. Keriven. Context-dependent kernels for object classification. IEEE transactions on pattern analysis and machine intelligence 33 (4), 699-708.
- [127] Wang, X., Yan, Y., Hu, H. M., Li, B., and Wang, H. (2024). Cross-modal contrastive learning network for few-shot action recognition. IEEE Transactions on Image Processing, 33, 1257-1271.
- [128] Zhou, L., Lu, Y., and Jiang, H. (2024). Fease: Feature selection and enhancement networks for action recognition. Neural Processing Letters, 56(2), 87.

- [129] L. Wang and H. Sahbi. "Nonlinear cross-view sample enrichment for action recognition." European Conference on Computer Vision. Springer, Cham, 2014.
- [130] Wang, B., Chang, F., Liu, C., Wang, W., and Ma, R. (2024). An efficient motion visual learning method for video action recognition. Expert Systems with Applications, 255, 124596.
- [131] H. Sahbi and S. Deschamps. *Adversarial label-efficient satellite image change detection*. In IEEE IGARSS 2023.
- [132] Zhang, R., and Yan, X. (2024, April). Video-language graph convolutional network for human action recognition. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7995-7999). IEEE.
- [133] H. Sahbi, P. Etyngier, J-Y. Audibert, R. Keriven. Manifold learning using robust graph laplacian for interactive image search. In CVPR 2008.
- [134] Wanyan, Y., Yang, X., Dong, W., and Xu, C. (2024). A comprehensive review of few-shot action recognition. arXiv preprint arXiv:2407.14744.
- [135] Xie, J., Meng, Y., Zhao, Y., Nguyen, A., Yang, X., and Zheng, Y. (2024, March). Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 38, No. 6, pp. 6225-6233).
- [136] M. Ferecatu and H. Sahbi. "Multi-view object matching and tracking using canonical correlation analysis." 2009 16th IEEE International Conference on Image Processing (ICIP). IEEE, 2009.
- [137] Saha, A., Gupta, S., Ankireddy, S. K., Chahine, K., and Ghosh, J. (2024). Exploring explainability in video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8176-8181).
- [138] H. Sahbi. *Learning laplacians in chebyshev graph convolutional networks*. In IEEE ICCV W2021.
- [139] Le, H., Lu, C. K., Hsu, C. C., and Huang, S. K. (2025). Skeleton-based human action recognition using LSTM and depthwise separable convolutional neural network. Applied Intelligence, 55(5), 298.
- [140] Xiao, J., Xiang, T., and Tu, Z. (2025). Adaptive prototype model for attribute-based multi-label few-shot action recognition. arXiv preprint arXiv:2502.12582.
- [141] H. Sahbi, L .Ballan, G. Serra, A. Del-Bimbo. Context-dependent logo matching and recognition. IEEE Transactions on Image Processing 22 (3), 1018-1031.