Simultaneous segmentation and classification of human actions in video streams using deeply optimized Hough transform

Adrien Chan-Hon-Tong, Catherine Achard, Laurent Lucat

To cite this version:


HAL Id: cea-01818435
https://hal-cea.archives-ouvertes.fr/cea-01818435
Submitted on 3 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Simultaneous segmentation and classification of human actions in video streams using deeply optimized Hough transform

Adrien CHAN-HON-TONG\textsuperscript{a}, Catherine ACHARD\textsuperscript{c,b}, Laurent LUCAT\textsuperscript{a}

\textsuperscript{a}CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Centre d’études de Saclay, Route Nationale, 91400 Gif-sur-Yvette, FRANCE
\{adrien.chan-hon-tong,laurent.lucat\}@cea.fr
\textsuperscript{b}CNRS, UMR 7222, ISIR, F-75005, FRANCE
\textsuperscript{c}Sorbonne Universités, UPMC Univ Paris 06, UMR 7222, ISIR, F-75005, Paris, FRANCE
catherine.achard@upmc.fr

Abstract

Most researches on human activity recognition do not take into account the temporal localization of actions. In this paper, a new method is designed to model both actions and their temporal domains. This method is based on a new Hough method which outperforms previous published ones on honeybee dataset thanks to a deeper optimization of the Hough variables. Experiments are performed to select skeleton features adapted to this method and relevant to capture human actions. With these features, our pipeline improves state-of-the-art performances on TUM dataset and outperforms baselines on several public datasets.

Key words: human actions, segmentation, classification, video streams, Hough

1. Introduction

Human activity recognition is becoming a major research topic (see [1, 30] for reviews). The ability to recognize human activities would enable the development of several applications like gestures based interfaces/interactions or multimedia data mining on web data. However, requirements for human activity recognition systems depends on the targeted applications. For example, on one hand, human action recognition systems for data mining on web data [23] deal with heterogeneous contents, heterogeneous resolutions and severe camera motions, but are not expected to be real time or to deal correctly with all the web database content. On the other hand, systems for gestures based interaction [31] deal with homogeneous videos acquired under controlled setting but need the highest possible accuracy and real time capacity.

Action recognition in natural and unconstrained videos has known a turning point from [33], driven by progresses in object recognition in image. For this task, most of the state-of-the-art approaches [42] are based on Bag Of Words
(BOW) and Support Vector Machine (SVM) pipeline that captures the link between low-level features e.g. spatiotemporal interested points (STIP) or trajectories of interested points and actions through a learning process.

Gestures based interaction is independently reaching a turning point with the release of active camera, such as the Kinect, providing suitable information for human action recognition like depth map, and even, under controlled setting, the skeleton of the subject (see [14] for a review).

Beside gestures based interaction and offline web data annotation, a lot of applications are emerging like ambient intelligence or medical video protection. A typical example is the detection of daily actions to monitor people with a limited autonomy (elderly or disabled person) at home. Indeed, the distribution of their daily actions during a day or a week has a medical relevancy.

This last kind of applications introduces new challenges which are addressed in this paper. In particular, the temporal location of an action is as relevant as the action itself in video based monitoring. It is not the case in human action recognition in web databases, where the goal is to annotate a whole (i.e. segmented) video. For such new applications, the aim is simultaneous classification and temporal localizations of actions. Thus, approaches like BOW+SVM which discard temporal information can not deal with such a challenge. As there is no clear consensus in literature about how this localization should be evaluated, we focus in this paper, on determining at each time what action is performed by an actor among a set of actions. This task is called segmentation in this paper.

Moreover, vision based monitoring systems need to take into account the temporal context. In this way, they oppose gesture based interaction systems where features extracted in a specific time are enough to decide the corresponding action. In particular, the duration of an action is about two orders of magnitude larger in vision based monitoring data than in interface/interaction ones: the maximal size of an action is less than a couple of second for gestures based mouse (e.g. [19]) or 4s in [31] (dance move) against around 30s in [39, 43] (setting a table) and around 2 minutes for [38] (cooking, working at computer).

Among the introduced challenges, daily motions present a higher variability than interface ones where subjects try to perform predefine motions. Thus, direct template matching may not be efficient for daily actions. Also, skeleton extraction is not as accurate for daily action recognition as for gestures based interface. However, considering the large academic efforts to deal with general pose estimation, we expect the skeleton to be available in the future in medical surveillance field, and thus, we still rely on skeleton features in this paper.

To tackle these challenges, we introduce two contributions which are presented after related works (section 2). First, in section 3 and 4, we introduce a new optimized Hough Transform method and present experiments where this new method outperforms previous Hough Transform ones. Then, in section 5, we perform experiments which highlight the relevancy of this method for human
action segmentation in video streams, as it improves state-of-the-art results on several public datasets using skeleton based features. Perspectives opened by these contributions are presented together with conclusions (section 6).

2. Related works

Most related works are based on graphical model like [38, 48, 22, 26]. In [38], a two-layered Maximum Entropy Markov Model (MEMM) is built on pose based features such as proximity of hands with other articulations, positions of articulations, speeds. In [22], recognition is performed by an Action Graph based on velocity features. In [48], Hidden Markov Model (HMM) is learnt on skeleton representations. The ability of latent variables to extract middle level features by clustering similar successive features is currently a major improvement axis [44, 36] for such Conditional Random Field [18] (CRF) pipelines. Such middle level features are also considered in [40] in a naive Bayesian classifier context and lead to performance improvements.

Besides graphical models, works like [15] deal with the segmentation challenge by padding the video with windows on which a BOW+SVM pipeline is applied. By dynamic programming, an optimal padding of the video is computed, thus we call this method OP (for optimal padding). Hough Forest can also perform such segmentation as in [49], where skeleton stream patches are densely extracted and clustered by trees formed with weak binary classifiers based on distances between articulations. Then, each patch votes, and vote agglomeration provides a segmentation after smoothing.

In addition, an important point to mention is that the rising of such methods is slowed down by the few public databases available.

2.1. Available public databases

In [15], a segmentation method is presented but mainly applied on Weizmann [13] and Hollywood [20] datasets where actions are associated with whole (or segmented) videos. This is also the case in [38] which introduces a typical dataset for classification and presents a method for segmentation. The honeybee dataset [28] considered in [15] is not relevant for human action recognition but is relevant for general segmentation: [28] presents an algorithm for segmentation based on Linear Dynamic System (LDS) [32] (a graphical model with continuous state).

This is the opposite situation in [48] where unsegmented videos are provided but the presented algorithm, based on HMM, performs classification but not segmentation.

Considering [7], it appears that there are only few datasets providing unsegmented videos i.e. with different actions in different parts of the video and where the ground truth provides the temporal localization of actions (e.g. one action by frame). This is even worse, if we consider in these datasets only those containing skeleton data and/or focusing on daily actions.

As a result, to evaluate our algorithm, we mainly rely in this paper on UTKAD for TUM [39] and University of Texas Kinect-Action Dataset [48].
which are relatively similar datasets: TUM is a multi-sensor dataset that contains especially skeleton streams (fig. 1). Actions are performed by 5 subjects setting the table (e.g. Lowering an object, Opening a drawer). UTKAD is a set of videos acquired with a Kinect that provides skeleton streams. The actions are performed by 10 subjects in a human-machine interaction setting (e.g. clap hands, wave hands). Both datasets are composed of around 20 sequences, about 2 minutes each, containing around 10 kinds of actions.

These datasets contain middle length videos. However, our objective is to offer a pipeline that can be used on stream context for daily action segmentation (e.g. 24 hours of video surveillance).

2.2. Stream constraints

In stream context, segmentation algorithms can not store nor process the entire video to take a decision concerning the human action performed at a specific time. Hence, all algorithms with an unbounded latency like [26] or [28] (LDS) can not be applied. Also, for the same stream constraints, algorithms with a low computational complexity are required. This discards approach like [15] where a SVM classifier [10] is applied to a number of windows which is quadratic in the maximal size of an action, which may be too large for daily actions. Finally, LDS-like algorithms can be degraded: decisions can be taken with bounded latency. However, these degraded versions have also degraded complexity.

For these two reasons, we focus on Hough Transform (HT) which allows both bounded latency and linear complexity (see table 1).

3. Hough transform based segmentation

3.1. Voting process

The main idea of the probabilistic HT for non parametric object (e.g. human activities) introduced in [21] is to perform the detection as a local recognition problem in the voting space built by agglomerating feature votes. The bounded
Method                                      | Bounded latency | small complexity
--------------------------------------------|-----------------|-------------------
LDS e.g. [28]                                | ✗               | ✓                 
Degraded LDS                                 | ✓               | ✗                 
OP e.g. [15]                                 | ✓               | ✗                 
HT e.g. [49]                                 | ✓               | ✓                 

Table 1: Computational ability for algorithms like Linear Dynamic System based smoothing, Optimal Padding and Hough Transform.

latency and linear complexity of HT are induced by this property of local detection.

In the context of temporal detection applied to human action in streams, the HT is composed of three steps:

1: Feature extraction and quantization form codewords
2: Each codeword votes for the possible temporal centers of the possible actions according to specific learned weights
3: All the votes are agglomerated to form the Hough score from which detection decisions are taken

In this article, we offer a new formalism for HT based methods. It consists to model the steps 2 and 3 by introducing a function \( \theta () \) linking codewords, time displacements, and human actions, to vote weights. Let \( c \) be a codeword extracted at time \( t \). It votes with a weight \( \theta (a, \delta_t, c) \) for the hypothesis that the current action is an action of type \( a \) and is centered at time \( t + \delta_t \) (this weight does not depend on the time \( t \) but only on \( a, \delta_t \) and \( c \)). Then, given the set of localized codewords extracted from the video \( \mathcal{V} = \{c, t\} \) (\( c \) for codewords and \( t \) for the corresponding extraction time), the Hough score \( \mathcal{H}_\mathcal{V} \) that an action \( a \) is centered at the time \( t' \) is:

\[
\mathcal{H}_\mathcal{V} (t', a) = \sum_{(c, t) \in \mathcal{V}} \theta (a, t' - t, c) \tag{1}
\]

Salient peaks of \( \mathcal{H}_\mathcal{V} \) form the detections in video \( \mathcal{V} \) (for simplicity, both the video and the set of localized codewords are designed by \( \mathcal{V} \)).

3.2. State-of-the-art on training processes

All HT based methods can be described using the proposed formalism: they differ only in their estimation of the weights \( \theta (a, \delta_t, c) \). Some of these methods consider only statistics on the training database [21], while the other ones introduce an optimization step [25, 46, 50]. In this section, we introduce the main methods of the state of the art and describe them with the proposed formalism.
**Implicit Shape Model:**

In the Implicit Shape Model (ISM) [21], the HT is based on generative weights. Let \( P(a, \delta_t|c) \) be the probability that the action occurring at time \( t \) is \( a \) and is centered at time \( t + \delta_t \), knowing that a codeword \( c \) has been extracted at time \( t \). This probability is estimated with statistics of the training dataset and is supposed to be independent of \( t \) (it depends only on \( a, \delta_t \) and \( c \)). Then, the weights are given by:

\[
\theta_{ISM}(a, \delta_t, c) = P(a, \delta_t|c)
\]

The probability \( P(a, \delta_t|c) \) is estimated by:

\[
P(a, \delta_t|c) \approx \frac{N(a, \delta_t, c)}{N(c)}
\]

where \( N(a, \delta_t, c) \) is the number of times a codeword \( c \) has been extracted at time \( t \) in an action \( a \) centered in \( t + \delta_t \) and \( N(c) \) is the number of occurrences of the codeword \( c \).

This generative ISM approach has the advantage to be meta parameter-free and relatively robust to over-training. However, [25, 46, 50] report that this method produces false positives which can be partially reduced by introducing discriminative parameters.

**Max-Margin Hough Transform:**

In Max-Margin Hough Transform (MMHT) [25], a coefficient is introduced for each codeword to weight the ISM values, resulting in:

\[
\theta_{MMHT}(a, \delta_t, c) = w_c \times \theta_{ISM}(a, \delta_t, c) = w_c \times P(a, \delta_t|c)
\]

The weights \( w_c \) give more or less importance to the different codewords \( c \) according to their discriminative power. They are learnt simultaneously in a discriminative way through an optimization process similar to a SVM.

**Implicit Shape Kernel:**

In Implicit Shape Kernel (ISK) [50], the votes are also based on the ISM generative ones, but some coefficients are introduced to weight the different training examples. Hence, ISK training leads to:

\[
\theta_{ISK}(a, \delta_t, c) = \sum_i w_i \times P_i(a, \delta_t|c)
\]

where \( P_i(a, \delta_t|c) \) is an estimation of the probability \( P(a, \delta_t|c) \) based only on the training example \( i \). The weights \( w_i \) are learnt simultaneously in a discriminative way using a specific kernel-SVM training [50].

MMHT and ISK report experimental improvements over ISM by adding discriminative parameters. This trend is also supported by [46].
ISM + SVM:

In [46], it is shown that learning directly the ISM Hough map $\mathcal{H}(t', a)$ with a linear SVM is equivalent to the ISM combined with a weighting coefficient for each displacement. This results in:

$$\theta_{ISM+SVM}(a, \delta_t, c) = w_\delta_t \times \mathcal{P}(a, \delta_t | c) = w_\delta_t \times \theta_{ISM}(a, \delta_t, c)$$

Hough forest:

To our knowledge, ISM [21] and the presented extensions [25, 46, 50] are the only published methods to estimate the weights of Hough transform. More precisely, these methods define links between codewords and votes. There are, of course, various ways to select the features and the codewords, like the Hough forest (HF) [49] which are major methods of the state of the art. Hough forest uses ISM votes, but the mapping between features (usually data patches) and codewords (a leaf in a weak binary classifier tree) is constructed such that all training features associated with the same codeword are expected to come from training examples with a same label (e.g. same action). Several works (e.g. [49]) report that this automatic feature mapping process associated with ISM votes leads to significant experimental improvements against codewords obtained without learning, e.g. by K means algorithm.

However, in this section, we focus on the optimization of the weights defining the voting process i.e. on the link between codewords and votes which is generic in terms of features and codewords. Thus, the proposed method can be employed with codewords provided by Hough forest approach.

One common point of [25, 46, 50] is that the addition of discriminative parameters to influence ISM votes increases performances of HT. Hence, a natural idea is to learn directly all votes. But before, let us introduce some standard HT post-processing.

3.3. From detection to segmentation

Post-processing for segmentation:

In temporal action detection context, the goal is to find the temporal bounding box of each action, or often simply the center of each action. In segmentation context: the goal is to determine which action is performed by the actor among a set of actions in each frame.

Most of the articles of the HT literature (and particularly those presented before) focus on detection, hence, each codeword naturally votes for centers (temporal center of each action). Thus, these methods need to be adapted to perform segmentation.

Smoothing based segmentation:

To our knowledge, in literature, this adaptation is done by propagating the votes from centers to all frames thanks to a smoothing function. For example, in [49], this is done by applying a strong Gaussian smoothing on $\mathcal{H}$:

$$\mathcal{H}(t'', a) \leftarrow \sum_{t'} \mathcal{H}(t', a) \exp \left( -\frac{1}{2\sigma^2} (t' - t'')^2 \right)$$
Such smoothing can be applied even in a detection context to group close peaks. However, in segmentation context this smoothing should be more intense as each frame must receive a piece of the votes from centers.

Then, the decision about the action at time $t'$ is given by:

$$\hat{a} (t') = \arg \max_a (H (t', a))$$ (2)

As smoothing may have a significant impact on performances in segmentation context, it is interesting to learn an adapted smoothing e.g. to learn a specific variance for each action [49], or more generally, to learn completely some specific smoothing functions (especially not necessarily Gaussian).

**Beyond smoothing:**

However, this smoothing approach has a major drawback: as smoothing is performed separately from voting, it does not simultaneously learn how a codeword should vote and how the smoothing should be done.

Yet, this drawback can be solved by learning already smoothed votes. In this way, votes have just to be accumulated to perform segmentation following equation (2) (without any smoothing or other post-processes). Hence, the entire process, described by equation (1) and (2), is learnt in one step.

Let us point out that smoothed votes are natural for segmentation as they can be seen as votes for presence. In the detection context, each codeword votes for a position corresponding to the center of the action, while in the segmentation context, each codeword votes ideally for the interval where the action takes place (as illustrated in figure 2). This interval can be seen as a smoothed vote. Also, if the votes $P (a \text{ centered at } t' | c \text{ extracted at } t)$ are smoothed by the function $P (t'' \text{ is part of } a | a \text{ is centered at } t')$ then the result is the same as if votes are $P (t'' \text{ is part of } a | c \text{ extracted at } t)$.

In addition, it is better to smooth individually the votes of each codeword and to accumulate the votes than to do the opposite (as mainly done in literature): accumulate and smooth is the particular case of smooth and accumulate when each codeword is smoothed in the same way. In contrary, smooth and accumulate allows to learn a specific smoothing shape for each couple codeword-action.

Hence, from now, codewords will not vote for the center of an action but for the presence of an action. Thus, a codeword $c$ extracted at time $t$ votes with a weight $\theta (a, \delta_t, c)$ for the hypothesis that an action $a$ is present at time $t + \delta_t$. This does not change the form of the equation (1) but the semantic of $\theta$ and of $H$ (which now represents the score that an action $a$ is present at time $t'$). Then, equation (2) provides segmentation.

**Decreasing constraints:**

This semantic change leads to several consequences e.g. on perfectly structured data, all the codewords of an action vote equally for each time in the interval of the action and not only for the center (figure 2). These two ways are equivalent on perfect data as there is a clear correspondence between peaks
intensity and action duration, but the presence representation is more general (because going from interval to center is easy but not the opposite) and thus is more adapted to deal with real data.

Also, a smoothing function is naturally decreasing toward the distance to 0. In our context, the more \( t' \) is far from \( t \) the less a codeword extracted in \( t \) should provide information about \( t' \). Later on in this paper, these constraints are called decreasing constraints and can be written mathematically as:

\[
\forall a, c, \delta_{t_1}, \delta_{t_2} : \begin{cases}
\delta_{t_1} \leq \delta_{t_2} \leq 0 \Rightarrow \theta(a, \delta_{t_1}, c) \leq \theta(a, \delta_{t_2}, c) \\
0 \leq \delta_{t_1} \leq \delta_{t_2} \Rightarrow \theta(a, \delta_{t_1}, c) \geq \theta(a, \delta_{t_2}, c)
\end{cases}
\] (3)

The advantage to vote for presence and not centers is the ability to learn an optimal already smoothed voting function \( \theta() \) as presented in next section.

4. Deeply optimized Hough Transform

We offer a new training process for Hough transform where all votes namely all \( \theta() \) values are directly optimized without using any generative values coming from ISM. Hence, the set of variables to optimize is indexed simultaneously by codewords, displacements and action and not only for one of them like in MMHT or ISM + SVM (see table 4). For this reason, this training process is called Deeply Optimized Hough Transform (DOHT).
\( P(a, \delta_i|c) \) is the probability that an action \( a \) occurring in \( t \) is centered at time \( t + \delta_t \) knowing that a codeword \( c \) has been extracted at time \( t \). \( P_i(a, \delta_i|c) \) is the same probability estimated using only the training example \( i \).

Table 2: The different learning methods of the Hough transform.

<table>
<thead>
<tr>
<th>methods</th>
<th>( \theta )</th>
<th>variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISM [21]</td>
<td>( \theta_{ISM}(a, \delta_t, c) = P(a, \delta_t</td>
<td>w) )</td>
</tr>
<tr>
<td>MMHT [25]</td>
<td>( \theta_{MMHT}(a, \delta_t, c) = w_c \times P(a, \delta_t</td>
<td>c) )</td>
</tr>
<tr>
<td>ISK [50]</td>
<td>( \theta_{ISK}(a, \delta_t, c) = \sum_i (w_i \times P_i(a, \delta_t</td>
<td>c)) )</td>
</tr>
<tr>
<td>ISM+SVM [46]</td>
<td>( \theta_{ISM+SVM}(a, \delta_t, c) = w_{\delta_t} \times P(a, \delta_t</td>
<td>c) )</td>
</tr>
<tr>
<td>DOHT (our)</td>
<td>( \theta_{DOHT}(a, \delta_t, c) = w_{a,\delta_t,c} )</td>
<td>( w_{a,\delta_t,c} )</td>
</tr>
</tbody>
</table>

This process is straightforwardly defined by the goal of selecting the optimal values for \( \theta(a, \delta_t, c) \) that provide correct decisions using equations (1) (2) at testing time. For all training examples \( \mathcal{V} \) and all times \( t' \), the predicted action \( \hat{a} \) should be the real one \( a^* \) (known on the training data) i.e.

\[
\forall \mathcal{V}, t': \hat{a}_V(t') = a^*_V(t')
\]

Considering the definition (eq. (2)) of the predicted action \( \hat{a} \), we obtain:

\[
\forall \mathcal{V}, t', a \neq a^*_V(t') : \mathcal{H}_V(t', a) < \mathcal{H}_V(t', a^*_V(t'))
\]

However, dealing with strict inequality is not computationally feasible, thus we impose a margin of 1:

\[
\forall \mathcal{V}, t', a \neq a^*_V(t') : \mathcal{H}_V(t', a) + 1 \leq \mathcal{H}_V(t', a^*_V(t'))
\]

and by replacing \( \mathcal{H} \) by the definition (eq. (1)), we obtain:

\[
\forall \mathcal{V}, t', a \neq a^*_V(t') : \sum_{(c,t) \in \mathcal{V}} \theta(a, t' - t, c) + 1 \leq \sum_{(c,t) \in \mathcal{V}} \theta(a^*_V(t'), t' - t, c)
\]

Hence, the optimal function \( \theta \) is simply the solution of this set of data constraints to which the decreasing constraints (eq. (3)) are added.

However, to manage noisy training data, a soft margin framework is applied as in [8]: some variables \( \xi \) are introduced to allow some data constraints to be unsatisfied but these variables are minimized in order to reduce the number of unsatisfied constraints though a loss function \( L_{data} \). Moreover, to prevent over-fitting, a regularity term \( L_{reg} \) is added to the objective function as in [4]. A coefficient \( C \) mitigates the trade off between the attachment to data and the regularity as in [8, 4]. Thus, our training process consists to solve:

\[
\min_{\theta \geq 0, \xi \geq 0} (L_{reg}(\theta) + C \times L_{data}(\xi))
\]

under constraints:

\[
\left\{ \begin{array}{l}
\forall \mathcal{V}, t', a \neq a^*_V(t') : \sum_{(c,t) \in \mathcal{V}} \left( \theta(a^*_V(t'), t' - t, c) - \theta(a, t' - t, c) \right) + \xi(t') \geq 1 \\
\forall a, c, \delta_{t1}, \delta_{t2} : \\
\delta_{t1} \leq \delta_{t2} \leq 0 \Rightarrow \theta(a, \delta_{t1}, c) \leq \theta(a, \delta_{t2}, c) \\
0 \leq \delta_{t1} \leq \delta_{t2} \Rightarrow \theta(a, \delta_{t1}, c) \geq \theta(a, \delta_{t2}, c)
\end{array} \right.
\]
Figure 3: Overview of the Hough method.

Green indicated meta parameters namely what features are extracted and $K$ and $C$ for the training step. There is no meta parameter during testing time.

There are to our knowledge no specific algorithm to deal with this formulation. The optimization framework eq. (4) seems very similar to SVM, but nearly all decreasing constraints may be saturated (whereas only a few data constraints, the support vectors, are usually expected to be saturated in SVM). This makes optimization challenging. Hence, we offer to use the $L_1$-norm for both $L_{reg}$ and $L_{data}$ as the problem becomes a linear program which is a well-studied problem in literature (eg. [17]) and which can be efficiency solved (for example using the solver CPLEX\(^1\) available freely for academic purposes).

4.1. Overview

The figure 3 gives a complete overview of the pipeline.

This pipeline is the same for all methods from table 4: the only difference

\(^1\)www-01.ibm.com/software/commerce/optimization/cplex-optimizer/
is the cells learning formulation. If the learning formulation is the \textit{DOHT} formulation (eq. (4)), then the pipeline is the \textit{DOHT} pipeline. If the learning formulation consists to use the elementary probability as vote map, this is just \textit{ISM} pipeline. In addition, prediction $\hat{a}$ at frame $t'$ during testing time in video $V$ is simply, as indicated, equal to $\arg\max_a \left( \sum_{(c,t)} \theta(a, t'-t, c) \right)$. There is thus no difficulty to naively implement such pipeline.

For an efficient implementation, the voting process has to be done incrementally as the words are extracted. This can be done by the following pseudo-code:
\begin{verbatim}
//let T be the number of frame in the sequence
//let A be the number of action
//let M be the maximal size of a training action
//let c[t=1..T] be the word extracted at frame t
//let theta[a=1..A,dt=-M..M,c=1..max_words] be the voting weights
//then pred[t'=1..T] is the predicted action at frame t'
pred=prediction(T,A,M,theta,c)
initialize pred[t'=1..T]
initialize H[t'=1..T,a=1..A] to 0
for t from 1 to T do
  for t' from t-M to t+M do
    for a from 1 to A do
      H[t',a]+=theta[a,t'-t,c[t]]
  pred[t-M] = arg max over a of H[t-M,a]
return pred
\end{verbatim}
As illustrated by this pseudo code, the testing time complexity of the Hough pipeline is $O(T \times A \times M)$ and decision in time $t'$ is taken with bounded latency $M$. For \textit{OP}-like algorithm testing time complexity is $O(T \times A \times M^2)$ and for \textit{degraded-LDS} it is $O(T \times A^2 \times M)$ (it is $O(T \times A^2)$ for standard \textit{LDS} but with unbounded latency). So as discussed in table 1, the advantage of the Hough pipeline is the straightforward ability to deal with streams involving long action and large set of actions.

However, the selection of $\theta$ function is crucial: in the next sub-section, we evaluate the different Hough methods i.e. the different ways to select $\theta$ (\textit{ISM}, \textit{HHMT}, \textit{ISM+SVM}, \textit{DOHT}) on public datasets. As \textit{ISK} is only intended for detection and not for segmentation, it is not compared to the others methods.

4.2. Evaluation on honeybee dataset

\textit{honeybee}:

Experiments for evaluating \textit{DOHT} have been conducted on the \textit{honeybee} dataset [28]. The latter dataset provides tracking output of honey bees having 3 kinds of behaviours correlated with their trajectory (figure 4). It is composed of 6 large sequences. This dataset is well designed for action segmentation as each frame is associated with an action. To provide results comparable to [28], algorithms perform a leave-one-out cross validation setting (LOOCV): algorithms learn models using all videos from the dataset except one which is used
Features and codewords:

The raw input signals in this dataset are the sequences of bee 2D positions and orientations \((x_t, y_t, \phi_t)\). As the goal of this set of experiments is to evaluate the voting process optimization, we do not focus on features, and thus, we just consider similar features from [15] which achieves the best published performances on this dataset. We use normalized short temporal series of 2D positions as features. Let us call \(R(\mu)\) the matrix of the 2D rotation of angle \(\mu\) and \(p(t) = (x_t, y_t)\), we consider the vector \((R(-\phi_t)(p_{t-\tau} - p_t), ..., R(-\phi_t)(p_{t+\tau} - p_t))\) as the feature extracted at time \(t\). Multiple \(\tau\) are considered simultaneously. Same features are considered for all \(HT\). All these features are clustered (independently for each \(\tau\)) using standard \(K\) means algorithm. The cluster centres define the codebook and features are mapped to their nearest codeword. In order to provide deterministic results (random free), \(K\) means is initialised with deterministic seeds: the two first seeds are a diameter of the point clouds, then next seed is the point with highest distance to the current set of seeds until \(K\) seeds have been extracted.

Results:

The meta parameters of the algorithms before the voting process are the sizes \(\tau\) of the extracted short series, the \(K\) of the \(K\) means algorithms. As we focus on the voting process, we simply make those meta parameters varying on a coarse empirical grid. The only meta parameter of the voting process is \(C\) (eq. 4) which is tuned on the \(2^{-5}, ..., 2^{10}\) grid from [5] (independently for
Method | Accuracy mean on *honeybee*  
--- | ---  
ISM from [21] | 71.9  
MMHT from [25] | 78.8  
ISM + SVM from [46] | 77.5  
DOHT (our) | **86.5**  
Approximated DOHT (our) | 85.1  
Inconsistent DOHT (our) | 79.4  
HDP-HMM [45] | 83.3  
SLDS [28] | 85.9  
PS-SLDS [28] | **87.7**  
naive OP [15] | 84.5  
OP [15] | **89.3**  

Table 3: Results on *honeybee* [28] for a common typical setting.
ISM, MMHT, ISM+SVM are reimplemented. Other results are just extracted from the corresponding papers.

each algorithm). We want to stress that all these meta parameters only concern input data or training step: there is no testing meta parameter in the voting process.

The main result on this dataset is that the maximal performances achieved by ISM, MMHT and ISM+SVM over all meta parameters are lower than the means performances of DOHT. Results for a typical run corresponding to $\tau \in \{1, 3, 6\}$, $K = 10$, and $C = 1$ are presented in table 3.

For these meta parameters, DOHT achieves 86.5% mean accuracy and significantly outperforms ISM, MMHT and ISM+SVM (table 3). In addition, DOHT results are just a little behind state-of-the-art results published in [28, 15]: 86.5% against 89.3% and 87.7%. However, these last results are obtained with algorithms which would lead to overweening computation testing times on larger datasets as discussed in section 2.2 and table 1. This non-linear-testing-time would be especially problematic for PS-SLDS which is close to CRF with latent variables that cluster frames into intervals, leading to a testing time complexity containing both quadratic factor from CRF and OP.

Discussion:

These results on honeybee dataset support the efficiency of the proposed DOHT toward state-of-the-art algorithms with important testing computation times and other HT methods (notice that the testing time of DOHT is the same as the ISM one).

Unfortunately, there is a practical limitation for using DOHT: the training formulation (eq. (4)) can hardly be performed on large datasets due to the required training time. Of course, training is an offline process that can be done one for all, but, even with a 64 processors 2024 Go RAM server, training is not trivial on large datasets. Thus, we offer a practical way to make the learning
4.3. Training speed-up

Identified issues:

Three ways have been identified to reduce the training time: reducing the number of constraints, handling specifically the positivity constraints \((\theta \geq 0)\) and handling specifically the decreasing constraints (eq. (3)).

The number of constraints is reduced by sampling the frame \(t'\) considered to define the constraints.

The positivity constraints are approximated by transforming them into virtual data constraints: for all \(c, \delta_t\), we add to the real data constraints, a data constraint corresponding to a virtual action \(a_{\text{false}}\) present at \(t + \delta_t\) in a stream containing only \(\Upsilon \gg 1\) negative codewords \(c\) at time \(t\). These virtual data constraints are processed as real ones leading to:

\[
\forall a, c, \delta_t : (-\Upsilon \times \theta(a_{\text{false}}, c, \delta_t)) - (-\Upsilon \times \theta(a, c, \delta_t)) \geq 1 - \xi'
\] (5)

These virtual data constraints are satisfied at lower cost if \(\theta(a_{\text{false}}, c, \delta_t) \approx 0\) and for all \(a, c, \delta_t\), \(\theta(a, c, \delta_t) \geq \frac{1}{\Upsilon} \approx 0\), and thus distort optimization toward positive \(\theta()\) (for real actions).

To our knowledge, there is no standard way to deal with the decreasing constraints. A very simple solution could be to discard these constraints, as in this way, the problem is similar to a multiclass SVM. However, this approach, called the Inconsistent DOHT, leads to poor results as shown in table 3. Indeed, as the temporal consistency of votes is completely lost, this lead to a harmful noise over fitting. So, we introduce a trick to make implicit these constraints in a SVM formulation by expressing \(\theta\) using characteristic functions.

Handling decreasing constraints:

Let us introduce the characteristic function: if \(I\) is a time interval then

\[
\forall t : \chi_I(t) = \begin{cases} 
  1 & t \in I \\
  0 & t \notin I 
\end{cases}
\]

Let \(M\) be the maximal length of action in the training data and let \(\mathcal{J}^{\text{full}}\) denote the set of all time intervals containing 0 included in \([-M, M]\). Then, the satisfaction of equation (3) by \(\theta()\) is equivalent to the existence of a function \(w() \geq 0\) satisfying:

\[
\theta(a, \delta_t, c) = \sum_{I \in \mathcal{J}^{\text{full}}} w(a, I, c) \times \chi_I(\delta_t)
\]

especially because equation (3) is satisfy by each \(\chi_I\) if \(0 \in I\). Hence, the decreasing constraints can be omitted during the optimization step by looking for \(w()\) instead of \(\theta()\). However, as there are \(M^2\) intervals in \(\mathcal{J}^{\text{full}}\) (leading to a computational dead-end), DOHT is approximated (an called Approximated DOHT) using a subset \(J \subset \mathcal{J}^{\text{full}}\) of time intervals.
Final training formulation:

Considering our experiments, we recommend this Approximated DOHT with 1-norm for \( L_{\text{data}} \) and with squared 2-norm for \( L_{\text{reg}} \). Thus, our final training problem is a standard multi-class SVM [6, 16]:

\[
\min_{w,\xi} \left( \|w\|_2^2 + C \times \|\xi\|_1 \right)
\]

under constraints: \( \forall \mathcal{V}, t', a \neq a^*_V(t') : \sum_{(c,t) \in \mathcal{V}, I \in J} ((w(a^*_V(t'), I, c) - w(a, I, c)) \chi_I(t' - t)) + \xi(t') \geq 1 \) (6)

where \( \mathcal{V} \) can represent a real data or come from eq. (5). To highlight that this problem is an SVM, let us map the couples \((I,c)\) into integers through a function \( \phi \), then let us introduce the vector \( Q \) such that \( Q_{\phi(I,c)} \) is the number of codewords \( c \) extracted from the video during the interval \( I \) translated by \( t' \), and let the vectors \( W_a \) be such that \( W_{a,\phi(I,c)} = w(a, I, c) \). Then

\[
\sum_{(c,t) \in \mathcal{V}, I \in J} ((w(a^*_V(t'), I, c) - w(a, I, c)) \chi_I(t' - t)) = \langle W_{a^*_V(t')} - W_a, Q \rangle
\]

where \( \langle , \rangle \) is the standard scalar product. Alternatively, the 1-norm for \( L_{\text{reg}} \) et \( L_{\text{data}} \) leads to a \( lp \)-SVM [2].

We observe experimentally on honeybee dataset that a small fixed-size subset \( J \) is sufficient for Approximated DOHT to perform as well as the DOHT: using \( J = \{ [-2^{-\alpha}M, 2^{-\beta}M], 0 \leq \alpha, \beta \leq 6 \} \) and the SVM solver from [16], Approximated DOHT speeds up the training time by a factor 50 and achieves equivalent performances than DOHT (85,1% against 86,5% see tab. 3) on honeybee.

This Approximated DOHT combined with a weak sub-sampling of constraints (called DOHT for simplicity later on in this paper) can be trained on large dataset particularly on motion capture stream to segment human activities.

5. Human Action Segmentation in Video Streams

5.1. Normalization of the skeleton signal

In this paper, we consider videos whose skeleton can be extracted to focus on skeleton based features. The raw signal from the video is thus a set of positions for each articulation along the time, in a coordinate system linked to the camera. In order to make the posture independent of the camera configuration, the signals have to be normalized. The main normalization processes of the literature are based on distances, angles or on subjective view.

Angle based normalization. The angles between two vectors are invariant to global rotations and translations. A representative angle normalization has been introduced in [31]. First, the principal component analysis (PCA) is applied on the 3D positions of neck, spine, tail and left/right shoulders and hips providing
two vectors $u$ and $r$. Then, the vector connecting shoulders to elbows, hips to knees and neck to head are decomposed in the spherical system $u, r$ i.e. the vector is describe by the two angles with $u$ (inclination) and $r$ (azimuth). Then, the vector connecting respectively elbows to hands, knees to feet are decomposed in a local spherical system $u', r'$ where $u'$ is the vector connecting respectively shoulders to elbows and hips to knees and $r'$ is the projection of $r$ into the orthogonal plane of $u'$. Finally all angles are unfolded into a 16D vector that describes the 48D skeleton (16 articulations, each corresponding to a 3D point).

Subjective view based normalization. Given in a standard coordinate system linked to a skeleton, the positions of articulations become invariant to global rotations and translations of the skeleton. For example, the positions of all articulations can be expressed in the coordinate system linked to the torso introduced in [31].

In [48], the coordinate system linked to the torso is computed and the sphere containing the person is decomposed into solid angle. Each solid angle corresponds to a histogram bin and each articulation votes for its containing bin, leading to a histogram that characterizes the skeleton.

Distances based normalization. The distances between two articulations are invariant to global rotations and translations. In [43], the distances for all pairs of articulations are considered, introducing some redundancy: 128 distances are computed for a 48D skeleton. Relevant distances are then selected by a learning process. In [49], a Hough forest selects relevant distances throw a sampling into distance based weak binary classifier and entropy based rejection.

Evaluation. In order to evaluate the influence of the normalization, we extract some common features and use them either without normalization or with one of the three different normalizations. The simplest features are considered in this set of experiments: the feature extracted in a frame is simply the full skeleton unfolded into a vector. Each skeleton vector is then transformed into a vector of angles from [31] or a vector of distances from [43] or expressed into the coordinate system from [31] (to evaluate respectively angle, distance or subjective view normalization).

In all cases, after normalization, the skeleton extracted at a specific frame is represented by a fixed size vector. These vectors are then quantified using $K$ means algorithm. Evaluations are performed on UTKAD and TUM datasets with DOHT.

For TUM, [49] being the most relevant paper associated to TUM dataset in our context, we follow the same experimental protocol. For this purpose, the dataset is split into training and testing parts following a predefine scheme (see [49]). Each algorithm provides a predicted action associated with the left hand motion in each frame of the testing sequences. Performances are measured by the total number of correct predictions over the total number of frames in testing videos. If not mentioned, the input of our pipeline is the reduced skeleton with 13-joints.
Normalization of skeleton coupled with DOHT

<table>
<thead>
<tr>
<th>Normalization</th>
<th>Accuracy mean on UTKAD</th>
<th>Accuracy mean on TUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw skeleton</td>
<td>58.3</td>
<td>53.4</td>
</tr>
<tr>
<td>angle skeleton</td>
<td><strong>72.0</strong></td>
<td>61.7</td>
</tr>
<tr>
<td>distance skeleton</td>
<td>69.2</td>
<td>60.2</td>
</tr>
<tr>
<td>subjective skeleton</td>
<td>70.7</td>
<td><strong>62.7</strong></td>
</tr>
</tbody>
</table>

Table 4: Results of different normalizations on the skeleton features on HTKAD and TUM datasets.

In UTKAD, there are sometimes multiple actions in a same frame. In such frame, we remove all actions except the dominant. Then, performances are measured through a LOOCV as for honeybee.

**Results and discussion.** The training meta parameters of this experiment are the $K$ and $C$ (as for honeybee). Maximal performances of these different algorithms are presented in table 4.

All these normalizations achieve roughly equivalent performances but outperform not normalized skeleton. Also, when DOHT is applied on both 3 normalized skeletons, it does not achieve better results than on each normalized skeletons separately, highlighting that these normalizations are not complementary.

On the other side, angle based normalization is less generic than the others since the transformation applied is different for each articulation (normalization is performed incrementally from center articulations to body extremities). Also, distance normalization is not invariant to the global body size of the different subjects contrary to subjective view and angle normalization. So, we decide to apply this subjective view normalization on the feature extracted from the stream. We focus on them in the next sub-section.

5.2. Selected features

In literature, both single frame skeleton features [48] and multiple frames skeleton features [43] are proposed as well as both single articulation [43] or combined [31, 48, 49] approaches. The feature type has a significant impact on e.g. the discriminative power and robustness to occlusions. Typically, a grouped feature may be more discriminative than the corresponding part features but may not be extracted as soon as one part is occluded. In addition, articulation combination leads to a higher feature space and may lead to adverse effects e.g. for dictionary construction using $K$ means algorithm. Finally, when relying on training step, the learning algorithm may be more efficient by selecting relevant combination from individual features.

In order to compete single articulation versus skeleton features, we evaluate DOHT fed by one or the other.
Table 5: Results of skeleton versus articulation features on UTKAD and TUM datasets.

<table>
<thead>
<tr>
<th>input features</th>
<th>Accuracy mean for DOHT on UTKAD</th>
<th>Accuracy mean on TUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>skeleton</td>
<td>70.7</td>
<td>62.7</td>
</tr>
<tr>
<td>articulations</td>
<td>74.8</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Single articulation versus skeleton features:

The skeleton feature is a vector \( s \) where \( s_{3 \times i+j} \) is the coordinate \( j \) (in 3D) of the articulation \( i \) (one vector for each skeleton). These skeleton features are quantified using \( K \) means algorithm.

The single articulation feature corresponds to the 3D position of one articulation \( i \) i.e. the vector \( (s_{3 \times i}, s_{3 \times i+1}, s_{3 \times i+2}) \). These vectors are then quantified independently for each articulation, using \( K \) means algorithm. A unique \( K \) value is considered for all articulations.

The training meta parameters of this experiment are again \( K \) and \( C \). The maximal performances of these pipelines are presented in table 5.

These experiments support that articulation based features are better than skeleton ones for learning based algorithms. This trend is yet not crucial on UTKAD where actions are correlated with global body motion, but on TUM where action is strongly correlated with the motion of the left hand, the articulation based features strongly outperform skeleton ones with DOHT.

These results support independent articulation features toward combined ones and invite to compete independent frame features toward combined ones.

Single frame versus multiple frame features:

Multiple frame approaches have the ability to capture some temporal property of the action unlike single frame ones, and, have provided an important breakthrough [37, 42] in human action recognition. Particularly, in a BOW+SVM framework [37, 42, 43], actions like stand up or sit down can hardly be distinguished using single frame features as no element of the pipeline takes into account temporal information. So such framework requires multiple frame features like Fourier temporal pyramid [43] or interest point trajectories [37, 42]. But, single frame skeleton features are more robust to skeleton extraction failure, speed variation and to irregular sub-sampling of the frames of the video. This invites to evaluate if multiple frame features are useful in HMM context [48] or DOHT context where the temporal information is already captured by the learning processes.

To evaluate the relevancy of multiple frame features in our context, we introduce several multiple frame features:

**Speeds.** The first multiple frame features are simply the displacement vectors between two successive positions of an articulation i.e. if \( q_t \) is the 3D position of a specific articulation in the normalized coordinate system then \( q_{t+1} - q_{t-1} \) is
Table 6: Results of different features with DOHT on TUM dataset.

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>positions</td>
<td>77.6</td>
</tr>
<tr>
<td>speeds</td>
<td>77.1</td>
</tr>
<tr>
<td>tracklets</td>
<td>80.1</td>
</tr>
<tr>
<td>position + speeds</td>
<td>79.8</td>
</tr>
<tr>
<td>positions + tracklets</td>
<td>80.6</td>
</tr>
<tr>
<td>speeds + tracklets</td>
<td>80.8</td>
</tr>
<tr>
<td>positions + speeds + tracklets</td>
<td>81.5</td>
</tr>
</tbody>
</table>

the feature extracted for this articulation at frame $t$. As for position, $K$ means algorithm transforms these features into codeword.

Sub-trajectories. The second multiple frame features considered are short temporal series of 3D positions of each articulation: let the vector $(q_1, ..., q_T)$ be the normalized trajectory of one articulation, then, we consider the vector $(q_{t-\tau}, ..., q_{t+\tau})$ as a feature extracted at time $t$ with size $\tau$. These features called tracklets are densely extracted i.e. extracted at each frame where available. The hidden target of these features is to capture the gestures performed by the subject. Similar features are also considered in [37, 42] emphasizing the efficiency of such interest point trajectories for human action recognition.

Several sizes $\tau$ may be simultaneously considered. In this case, tracklets with different sizes may be clustered together to make codeword extraction robust to local speed of the action. Such clustering can be performed with $K$ medoids algorithm based on dynamic time wrapping (DTW). DTW [3] provides a classical distance between different sizes series by matching several times in one sequence to one time in the other and vice-versa. $K$ medoids is a variant of $K$ means where only distances are needed and not means which require a common space. Alternatively clustering can be done independently for each tracklet size $\tau$ with $K$ means algorithm. However, considering our experiment, it appears that using different sizes unlikely provides complementary information, so for simplicity we consider one size only.

Results. Experiments to compete single frame versus multiple frame features are performed only on TUM dataset as an irregular sub sampling does not allow using multiple frame features on UTKAD. The meta parameters of the experiment are $C, K$ and $\tau$ for tracklets ($\tau$ concerns only input data and $C, K$ only the training). Due to the size of parameters grid, we can not consider different $K$ for different features, so in an experiment e.g. positions + tracklets a common $K$ is considered for all clustering processes even if positions are 3-D features and tracklets $(3 \times \tau)$-D ones. Maximal performances of pipelines are presented in table 6.

These experiments support that multiple frame features performs better than single frame ones in context of DOHT algorithm. However, these different
Table 7: Results for tracklet features when varying the training meta parameters \( K, C, \tau \) with \textit{DOHT} on \textit{TUM} dataset.

<table>
<thead>
<tr>
<th>( \tau = 4 )</th>
<th>( K : )</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C = 0.25 )</td>
<td>70.5</td>
<td>69.5</td>
<td>67.9</td>
<td></td>
</tr>
<tr>
<td>( C = 0.5 )</td>
<td>72.4</td>
<td>72.1</td>
<td>70.4</td>
<td></td>
</tr>
<tr>
<td>( C = 1 )</td>
<td>74.4</td>
<td>74.2</td>
<td>73.8</td>
<td></td>
</tr>
<tr>
<td>( C = 2 )</td>
<td>75.1</td>
<td>76.3</td>
<td>75.3</td>
<td></td>
</tr>
<tr>
<td>( C = 4 )</td>
<td>74.8</td>
<td>76.5</td>
<td>76.3</td>
<td></td>
</tr>
<tr>
<td>( C = 8 )</td>
<td>74.8</td>
<td>77.2</td>
<td>75.7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \tau = 6 )</th>
<th>( K : )</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C = 0.25 )</td>
<td>70.2</td>
<td>70.0</td>
<td>70.4</td>
<td></td>
</tr>
<tr>
<td>( C = 0.5 )</td>
<td>72.2</td>
<td>73.2</td>
<td>73.7</td>
<td></td>
</tr>
<tr>
<td>( C = 1 )</td>
<td>75.4</td>
<td>75.8</td>
<td>76.3</td>
<td></td>
</tr>
<tr>
<td>( C = 2 )</td>
<td>77.0</td>
<td>78.0</td>
<td>77.4</td>
<td></td>
</tr>
<tr>
<td>( C = 4 )</td>
<td>77.3</td>
<td>79.4</td>
<td>78.3</td>
<td></td>
</tr>
<tr>
<td>( C = 8 )</td>
<td>76.4</td>
<td>78.9</td>
<td>77.6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \tau = 8 )</th>
<th>( K : )</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C = 0.25 )</td>
<td>70.2</td>
<td>71.7</td>
<td>70.5</td>
<td></td>
</tr>
<tr>
<td>( C = 0.5 )</td>
<td>72.8</td>
<td>76.0</td>
<td>73.2</td>
<td></td>
</tr>
<tr>
<td>( C = 1 )</td>
<td>74.9</td>
<td>77.9</td>
<td>76.0</td>
<td></td>
</tr>
<tr>
<td>( C = 2 )</td>
<td>76.6</td>
<td>80.1</td>
<td>78.1</td>
<td></td>
</tr>
<tr>
<td>( C = 4 )</td>
<td>77.1</td>
<td>79.2</td>
<td>78.6</td>
<td></td>
</tr>
<tr>
<td>( C = 8 )</td>
<td>77.3</td>
<td>78.3</td>
<td>77.7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \tau = 10 )</th>
<th>( K : )</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C = 0.25 )</td>
<td>71.4</td>
<td>72.3</td>
<td>70.8</td>
<td></td>
</tr>
<tr>
<td>( C = 0.5 )</td>
<td>73.7</td>
<td>76.1</td>
<td>74.3</td>
<td></td>
</tr>
<tr>
<td>( C = 1 )</td>
<td>76.1</td>
<td>78.2</td>
<td>76.9</td>
<td></td>
</tr>
<tr>
<td>( C = 2 )</td>
<td>77.6</td>
<td>79.3</td>
<td>78.2</td>
<td></td>
</tr>
<tr>
<td>( C = 4 )</td>
<td>76.7</td>
<td>79.4</td>
<td>78.8</td>
<td></td>
</tr>
<tr>
<td>( C = 8 )</td>
<td>76.3</td>
<td>79.0</td>
<td>78.9</td>
<td></td>
</tr>
</tbody>
</table>

Features are complementary as \textit{DOHT} fed with \textit{positions + speeds + tracklets} perform better than \textit{DOHT} fed by one of the three.

**Discussion:**

Best selected features. To summarise these experiments, it appears that in our context single articulation features are better than combined ones and multiple frame features better than single frame ones. The best evaluated features alone are \textit{tracklets}: the concatenation on a small time window of the 3D positions of one specific articulation densely extracted across videos. However, the best evaluated pipeline is the combination of simple positions + speeds + \textit{tracklets}.

Stability toward meta parameters. In these experiments, we focus on maximal performance achieved when coarsely varying meta parameters e.g. \( K, C, \tau \). The stability toward these meta parameters is yet important in a system perspective. Hence, in table 7, we present the different raw results. We want to stress that all these meta parameters only concern input data (\( \tau \)) or training step (\( K, C \)) but that there are no meta parameter specific to testing voting process.

The interested point is that none of the meta parameters is crucial: even if empirical maximal performance is achieved only with one specific setting, there are plenty of settings which lead to nearly maximal performances. For example, there are 6 setting providing more than the maximal performance minus 1.

One other interesting point is that \( K \) is small. An explanation is that in our training formulation (eq. (6)), the number of variables is \( K \times A \times J \) where \( A \) is the number of activities (typically 10) and \( J \) the number of intervals considered.
for the approximation (typically 64). Thus, increasing $K$ leads to more variables and may lead to over-fitting.

5.3. Comparison to the state-of-the-art

In table 8, our results on human actions are presented together with published comparable ones and with the ISM baseline.

However, considering the few number of comparable results (works dealing with segmentation on TUM and UTKAD datasets), we also offer to compare our algorithm to published close results. For this purpose, we present performance of our pipeline in classification context i.e. on already segmented data. As our algorithm is not designed for such context, these experiments lead to under estimate the performances of our algorithm. However, to our knowledge, the only alternative could have been to reimplement algorithms e.g. [49, 15, 28] and apply it on TUM and UTKAD datasets. But, these works need expert knowledge to select properly hyper parameters, and thus, the reimplementation may lead to under estimate performances of the state of the art. In addition, in classification context, we evaluate another baseline obtained by erasing location information from DOHT. This can be done by erasing the keyword locations (e.g. replacing them by a constant). Thus, the performance differences between this baseline and DOHT can be interpreted as the information provided by taking into account temporal location of extracted keywords from the videos. Alternatively, this can be done in practice by setting $J = \{[-M, M]\}$ where $J$ is the set of intervals from eq. (6). So this baseline corresponds to a $BOW+SVM$ without any normalization (e.g. no $l_1$ normalization, no tfidf normalization). Thus, we call this baseline unscaled SVM. Nearest neighbours algorithm is call NN. Videos from both training and testing sets are segmented according to the ground truth and then given as input to DOHT as it was unsegmented data.

These results are presented in table 9 (to our knowledge no other results on these datasets have been published). Our algorithm achieves 82.4% means accuracy on UTKAD toward 91.5% for the state-of-the-art. This is a promising result considering that our algorithm is designed for another context.

At first sight, another comparison could have be done on CAD-60 [38], especially because [38] is an algorithm which deals with segmentation. However there are three problems. First, even if [38] presents a segmentation algorithm,

<table>
<thead>
<tr>
<th>algorithm</th>
<th>TUM</th>
<th>TUM 27-joints</th>
<th>UTKAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>speed features + HF [49]</td>
<td>80.3</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td>all features + HF [49]</td>
<td>77.6</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>ISM</td>
<td>58.3</td>
<td>58.3</td>
<td>62.7</td>
</tr>
<tr>
<td>DOHT</td>
<td>81.5</td>
<td>83.0</td>
<td>74.8</td>
</tr>
</tbody>
</table>

Table 8: Comparison with published results on HTKAD and TUM datasets. Results are just extracted from the corresponding papers and do not come from reimplementation.
<table>
<thead>
<tr>
<th>algorithm</th>
<th>TUM</th>
<th>UTKAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM [48]</td>
<td>90.9</td>
<td></td>
</tr>
<tr>
<td>SVM [9]</td>
<td>90.4</td>
<td></td>
</tr>
<tr>
<td>sparse SVM [41]</td>
<td>90.9</td>
<td></td>
</tr>
<tr>
<td>ISM</td>
<td>65.5</td>
<td>64.1</td>
</tr>
<tr>
<td>unscaled SVM</td>
<td>80.6</td>
<td>79.8</td>
</tr>
<tr>
<td>DOHT</td>
<td>88.5</td>
<td>82.4</td>
</tr>
</tbody>
</table>

Table 9: Results on already segmented data on several datasets.
Results are just extracted from the corresponding papers and do not come from reimplementation.

CAD-60 data are already segmented. Then, results on this dataset are reaching a saturation point: even ISM achieves 90.3% of precision and 86.0% of recall in new person setting (see [38]). Finally but more problematic, algorithm from [38] is designed to decide as neutral the frames which are not interesting for any action but the ground truth (yet provided by [38]) does not take this into account. So, algorithm from [38] only achieves only 69.0% precision and 57.3% recall that is much lower than the baseline ISM.

Considering these comparisons, our pipeline improves state-of-the-art performances in human action segmentation context and achieves promising results on human action classification task. This supports the pertinence of our pipeline for human action segmentation.

6. Conclusion

In this paper, we offer two contributions. First, we offer a new Hough transform method based on discriminative parameters only. This Deeply Optimized Hough Transform is designed for segmentation by performing a voting process based on smoothed weights. It significantly outperforms generative based Hough transform on public datasets.

The second contribution of this paper is the application of this method to human action segmentation. Combined with the selected skeleton based features, this method slightly improves state-of-the-art performances on unsegmented data and achieves honorable performances on segmented ones on public datasets highlighting the relevancy of this new method for human action segmentation.

In future works, we will adapt and apply our method to deal with human action spatiotemporal segmentation in classic videos. This aim introduces two important challenges: performing segmentation in 3D (picture space + time) and managing a large number of features.
References


