HAL
open science

Action-centric Polar Representation of Motion Trajectories for Online Action Recognition

Fabio Martínez, Antoine Manzanera, Michèle Gouiffes, Thanh Phuong Nguyen

To cite this version:

Fabio Martínez, Antoine Manzanera, Michèle Gouiffes, Thanh Phuong Nguyen. Action-centric Polar Representation of Motion Trajectories for Online Action Recognition. International Conference on Computer Vision Theory and Applications (VISAPP), 2016, Roma, Italy. 10.5220/0005730404420448 . hal-01671916

HAL Id: hal-01671916
https://hal.science/hal-01671916v1

Submitted on 8 Jan 2018

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.

Action-centric polar representation of motion trajectories for online action recognition

Fabio Martínez [1,2], Antoine Manzanera[2], Michèle Gouiffès [1], Thanh Phuong Nguyen[3]
[1]*LIMSI, CNRS, Université Paris-Saclay, FRANCE*
[2]*U2IS/Robotics-Vision, ENSTA-ParisTech, Université Paris-Saclay, FRANCE*
[3]*LSIS, UMR 7296, Université du Sud Toulon Var, FRANCE*
{*fabio.martinez-carillo, antoine.manzanera*}*@ensta-paristech.fr, michele.gouiffes@limsi.fr, thanh-phuong.nguyen@univ-tln.fr*

Keywords: Action recognition, semi dense trajectories, motion shape context, on-line action descriptors

Abstract: This work introduces a novel action descriptor that represents activities instantaneously in each frame of a video sequence for action recognition. The proposed approach first characterizes the video by computing kinematic primitives along trajectories obtained by semi-dense point tracking in the video. Then, a frame level characterization is achieved by computing a spatial action-centric polar representation from the computed trajectories. This representation aims at quantifying the image space and grouping the trajectories within radial and angular regions. Motion histograms are then temporally aggregated in each region to form a kinematic signature from the current trajectories. Histograms with several time depths can be computed to obtain different motion characterization versions. These motion histograms are updated at each time, to reflect the kinematic trend of trajectories in each region. The action descriptor is then defined as the collection of motion histograms from all the regions in a specific frame. Classic support vector machine (SVM) models are used to carry out the classification according to each time depth. The proposed approach is easy to implement, very fast and the representation is consistent to code a broad variety of actions thanks to a multi-level representation of motion primitives. The proposed approach was evaluated on different public action datasets showing competitive results (94% and 88.7% of accuracy are achieved in KTH and UT datasets, respectively), and an efficient computation time.

1 INTRODUCTION

Action recognition is a very active research domain aimed to automatically segment, detect or recognize activities from video sequences. This domain plays a key role in many different applications such as video-surveillance, biomechanical analysis, human computer interactions, gesture recognition, among others. Action recognition is however very challenging, because of the great variability that is intrinsic to the object, regarding its shape, appearance and motion. The uncontrolled acquisition conditions also complicates the problem, due to illumination changes, different 3d poses, camera movements and object occlusions (Weinland et al., 2011).

Local spatio-temporal features have been widely used to recognize actions by coding salient descriptors, designed to be robust to viewpoint and scale changes. These features have been computed using different strategies, such as: Hessian salient features, Haar descriptors with automatic scale selection or by choosing 3D patches with maximum gradient responses (Ke et al., 2005), (Wang et al., 2009),(Willems et al., 2008). Such descriptors are however dependent on the object appearance, requiring thereby an extensive learning step to represent different motion patterns.

Motion-based approaches using optical flow primitives have also been used to characterize action patterns, with the advantages of being relatively independent from the visual appearance (Cao et al., 2009), (Efros et al., 2003),(Scovanner et al., 2007). For instance, Histograms of Oriented Optical Flow (HOOF) have been proposed as symmetry-invariant motion descriptors, to recognize periodic gestures (Chaudhry et al., 2009) but with the main limitation of losing local relationships of articulated objects. In (Ikizler et al., 2008) a block-based representation that combined HOOF and histograms of contours was proposed to recognize periodic actions. This strategy is

however dependent of a precise definition of the object and requires a complete video description to perform the recognition.

Optical flow fields have also been tracked during the video sequence resulting in local space-time trajectories that describe motion activities over longer duration. Descriptors based on these trajectories currently report very good performance to represent gestures and activities (Kantorov and Laptev, 2014) (Wang et al., 2011) (Jain et al., 2013). For instance in (Wang et al., 2011), local descriptors like HOF (Histograms of Optical Flow), MBH (Motion Boundary Histograms) and HOG (Histograms of Oriented Gradients) were computed around each trajectory and then integrated as space-time volumes centered in each trajectory to describe activities.

In (Jain et al., 2013) and (Wang and Schmid, 2013) the space-time volumes based on the characterization of trajectories were also implemented but using an improved version of the trajectories that takes into account the camera motion correction. These descriptors are however largely dependent on the appearance around the trajectories, which may be sensitive to illumination changes. Additionally, the spatio-temporal volumes are heuristically cut off from a fixed temporal length that may be restrictive to represent non periodic actions in a on-line scheme. Besides, the resulting size of the motion descriptor, composed of many spatio-temporal volumes may be prohibitive in real-time applications.

The main contribution of this work is a compact spatio-temporal action descriptor, based on local kinematic features captured from semi-dense trajectories, that is efficient in recognizing motion activities at each frame. Such features are spatially centered around the potential motion to be characterized and coded as an action polar representation. From point trajectories extracted in each time in the video, potential regions of interest are extracted in each frame using the centers of mass of the activity. Three different types of trajectories are evaluated. Then, around the center of mass of the current motion, an action-centric polar representation is built to capture the spatial distribution of kinematics features. For each region of the polar grid and at each frame, a motion histogram is stored and updated, to represent the regional temporal distribution of the trajectory kinematics. Therefore a temporal series of histograms can be analyzed at different time depths from the current frame. This collection of histograms forms a descriptor that is mapped to Support Vector Machine classifiers that returns an action label.

Unlike most existing methods that arbitrarily cut off the temporal support of their motion representation, the proposed motion descriptor is incrementally computed, which allows an on-line recognition of the actions, i.e., for each frame, a label is assigned to the potential action descriptor. Since the time depth of the histogram is null at the beginning of the sequence to be analyzed, the reliability of the label is low at the beginning and increases with time.

This paper is organized as follows: Section 2 introduces the proposed method, section 3 presents the results and a quantitative evaluation of the method. Section 4 concludes and presents prospective works.

2 The proposed method

Action recognition involves the characterization of particular events and behaviors, embodied by physical gestures in the time-space domain. A potential activity is herein defined as a particular spatial distribution of local motion cues updated along the time. For doing so, an action-centric polar representation is first designed to independently analyze motion trajectories within the different polar regions. Then, in each region are computed local kinematics along the trajectories, which are used to update a temporal histogram in each frame. Finally, the action descriptor is formed by the set of the motion histograms, which may be mapped at each time to a previously trained SVM classifier. A pipeline of the proposed approach is shown in Figure 1.

2.1 Motion cue trajectories

Local cues tracked along trajectories over several frames have demonstrated outstanding action modelling, thanks to their capability to represent foreground dynamics and interaction history. The proposed approach is aimed to recognize potential actions in on-line applications and therefore requires a significant set of trajectories with a suitable trade-off between accuracy and computation time. In this work were considered the following methods to compute trajectories:

Dense trajectories: are extracted from a dense optical flow field which is tracked by using a median filter over the displacement information and using multiple spatial scales. These trajectories have demonstrated good performance in terms of accuracy and speed in different scenarios of action recognition. This method includes a local criterion to remove trajectories with poor motion information, such as: 1) trajectories being beyond certain fixed standard deviation boundaries and 2) trajectories with sudden displacements, defined as

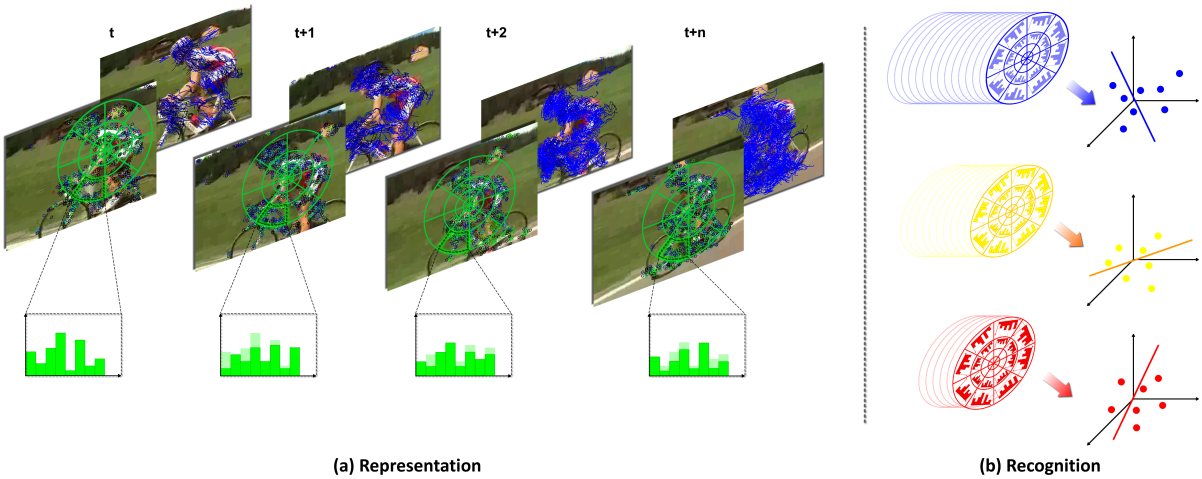

Figure 1: Pipeline of the proposed approach: (a) spatio-temporal representation of the proposed motion descriptor. First, a set of point trajectories are computed from the video. Then, a polar representation is centered in each frame around the potential action defined by the end points of the active trajectories. A histogram is computed in each polar region to estimate the partial distribution of kinematics. The histograms are updated at each time by considering different time intervals, thus enabling frame-level recognition. (b) recognition process using the SVM. Different histogram versions are built at each time corresponding to different time depth. Each one of these action descriptors is mapped to its respective SVM and a recognition label is returned at each frame.

velocity vectors whose magnitude exceeds 70% of the overall displacement of the trajectory (Wang et al., 2011).

Improved Dense trajectories: are a new version of the dense trajectories that takes into account a camera motion correction to globally filter out the trajectories. For doing so, the approach assumes that the background of two consecutive frames are related by a homography. The matching between consecutive frames is carried out using SURF descriptors and optical flow. The trajectory filtering according to speed and standard deviation is also applied (Wang and Schmid, 2013).

Semi-Dense trajectories: are key-point trajectories extracted from a semi-dense point tracking. The selected key-points are tracked using a coarse-to-fine prediction and matching approach allowing high parallelism and dominant movement estimation. This technique has the advantage to produce high density trajectory beams robust to large camera accelerations, allowing statistically significant trajectory based representation, with a good trade-off between accuracy and performance (Garrigues and Manzanera, 2012).

2.2 Kinematic trajectory characterization

A trajectory of duration n is defined as a set of n coordinates $\left\{(x_t, y_t)_{t=t_1}^{t_n}\right\} \in \mathbb{R}^2$ at different times t. Each trajectory contains relevant cues about the dynamic of a particular activity in the video, which can be characterized by kinematic measures computed using finite difference approximations.

In this work were considered different kinematic features like the velocity $\mathbf{v}(t)$, depicted by its direction $\theta(t) = \arg \mathbf{v}(t)$ and modulus (speed) $s(t) = ||\mathbf{v}(t)||$. The curvature κ was also tested in our representation and defined as $\kappa = \frac{\sqrt{\dot{x}_t^2+\ddot{y}_t^2+(\dot{x}_t\ddot{y}_t-\ddot{x}_t\dot{y}_t)^2}}{(\dot{x}_t^2+\dot{y}_t^2+1)^{3/2}}$, where $\dot{x}$ and $\ddot{x}$ are first and second time derivatives computed using finite difference approximations on the trajectories. The features can be used separately or jointly, and the proposed framework is flexible to include any kind of local features.

2.3 Action-centric polar representation

The motion information naturally available on trajectories and their spatial distribution in the sequence are two fundamental aspects to represent activities. Based on this observation, a motion shape analysis is herein carried out at each frame by measuring kinematic statistics distributed in a polar representation. This spatial representation quantifies the locations of a potential activity in a polar way. Additionally, a rotation-independent region analysis can be deduced to characterize activities and interactions. To obtain such representation, the center-of-mass of the points supporting the current trajectories is first calculated, to center the polar grid with respect to the current ac-

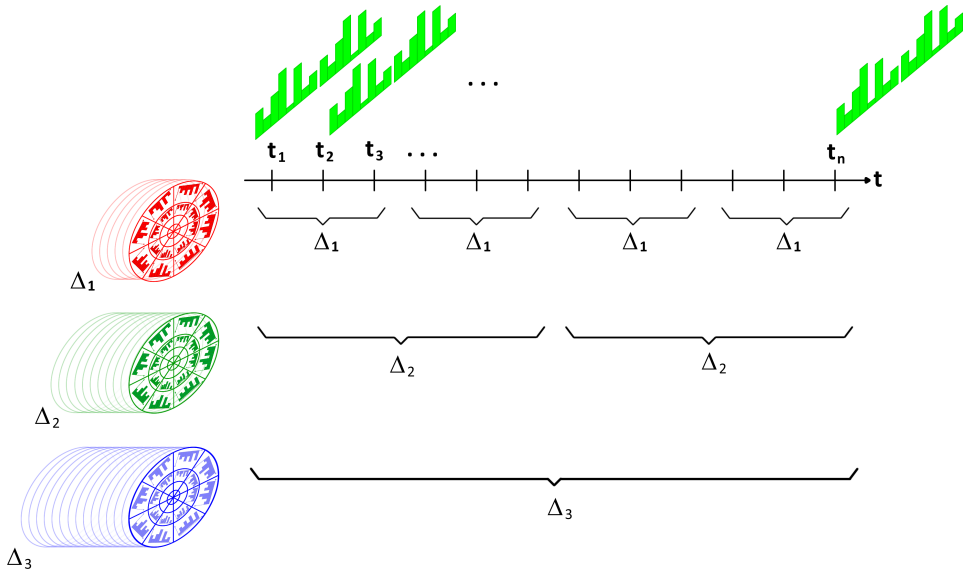

Figure 2: The different time depth histograms allow to enrich the spatio-temporal activity description while remaining efficient and usable in arbitrarily long sequences. During training, individual SVM models are learned by using different time periods for the different versions of descriptors. For the recognition step, the histograms are mapped to the corresponding SVM model according to their time depth. The minimum distance of the combined action descriptors w.r.t the respective hyperplanes allows to define the current label for the action.

tion location. Then, the distribution of the kinematics is measured within each relative region of the polar grid $u(\rho,\theta)$. Particularly, the kinematics measured in each polar region u_k are coded in motion histograms which are updated at each time.

Then, for each polar region u_k, a purely spatial histogram $h_t^k(b) = |\{\mathbf{x} \in u_k; f_t(\mathbf{x}) = b\}|$ codes the distribution of the kinematic feature f (quantized on N bins $b \in \{b_0, \ldots, b_{N-1}\}$) computed at time t on the current support point $\mathbf{x}$ of the trajectory. Then for every time depth Δ_i (see Fig. 2), a time cumulative version of the histogram is initialized every Δ_i frames to the spatial histogram:

$$H_t^{k,i}(b) = h_t^k(b)$$

when the time index t divides the time depth Δ_i. Otherwise it is updated as follows:

$$H_t^{k,i}(b) = \frac{\Delta_i - 1}{\Delta_i} H_{t-1}^{k,i}(b) + \frac{1}{\Delta_i} h_t^k(b)$$

The resulting collection of histograms (see Fig. 2) finally represents the frame level representation of the activity.

2.4 SVM Recognition at different temporal scales

Finally, the recognition of each potential activity is carried out by a Support Vector Machine (SVM) classifier, using the set of recursive statistics from the whole polar partition, as a spatio-temporal multiscale motion descriptor. This classifier is well known for being successfully applied to many pattern recognition problems, given its robustness, generalization aptness and efficient use of machine resources. The present approach was implemented by using the *One against one SVM multiclass classification* with a Radial Basis Function (RBF) kernel (Chang and Lin, 2011).

Firstly, for training step, it was learned individual SVM models according to the time periods considered for the different versions of histograms, i.e, a motion characterization from different time depths (see Fig. 2). Each SVM then learns a partial action representation that takes into account different history versions of the motion according to the histograms. A (γ, C)-parameter sensitivity analysis was performed with a grid-search using a cross-validation scheme and selecting the parameters with the largest number of true positives for each time-scale.

In the step of on-line recognition, each motion descriptor is mapped to a specific SVM model according to the time depth. Then, for each one of the K time depths, the distances of the motion descriptor w.r.t the hyperplanes formed by the bank of $\frac{N(N-1)}{2}$ classifiers is stored, with N being the number of action classes. To combine the different time depth classifiers, the distances to the hyperplane are summed for each activity and then, the activity with minimum sum is cho-

sen as the predicted class.

2.5 Datasets

The proposed approach has been evaluated on two well known public human action datasets presenting different levels of complexity, and different types of activity and human interaction. Here is a brief description of these datasets:

KTH: contains six human action classes: *walking, jogging, running, boxing, waving* and *clapping*. Each action is performed by 25 subjects in four different scenarios with different scales, clothes and scene variations. This dataset contains a total of 2 391 video sequences. The proposed approach was evaluated following the original experimental setup which specifies training, validation and test groups (Schuldt et al., 2004) as well as five-fold cross validation suggested in (Liu et al., 2009).

UT-Interaction: contains six different human interactions between different people: *shake-hands, point, hug, push, kick* and *punch* (Ryoo and Aggarwal, 2010). The dataset has a total of 120 videos. Each video has a spatial resolution of 720×480 and a frame rate of 30 fps. A ten-fold leave-one-out cross-validation was performed, as described in (Ryoo and Aggarwal, 2010).

3 Evaluation and Results

The experimental evaluation was designed to assess the different components of the proposed approach, regarding the classification of video-sequences and the on-line recognition. The motion characterization was firstly analyzed according to the three different versions of trajectories (sec 2.1) and several kinematic features coded over them (sec 2.2). This evaluation was carried out to classify complete sequences with one recorded activity. The second analysis was aimed to evaluate the proposed motion descriptor in the task of frame-level recognition. The configuration of the proposed approach was set with a polar grid representation of 8 directions and 4 norm divisions. At each polar grid region, a motion histogram of 32 bins was computed. The resulting size of the frame-level action descriptor was then of 1024 for each considered time depth. Three different time depths Δ_i of 8, 16 and 32 frames were considered. The velocity components and the curvature of the trajectories were considered as kinematic features for the local representation of the trajectories.

Firstly, the proposed motion descriptor was tested to classify actions in complete video-sequences. The label for a global video-sequence was defined following an occurrence criterion of the labels recovered in each frame, i.e., the action that is predicted most often at the frame level is assigned as the sequence label. The motion histograms were computed for each kinematic feature individually. Table 1 shows the results for the different versions of trajectories and for every kinematic feature. In general, the kinematic histograms achieve an appropriate characterization of individual actions and interactions from the datasets, the orientation of the velocity θ providing the best results. A concatenation of the three kinematic features was also tested, but the improvement in accuracy is only around 1%, while the size of the descriptor may turn prohibitive for online applications. The weak improvement provided by concatenating different kinematic histograms is probably due to a strong statistical dependency of these features.

In Table 2 and Table 3 are summarized the average accuracies of different state-of-the-art approaches for automatic classification in KTH and UT-interaction datasets. The proposed approach achieves a good performance for different motion activities under very different scenarios, and a wide spectrum of human motion activities. Furthermore, it has the determining advantage to compute frame level action descriptors with fixed size, allowing the action label prediction of partial sequences in a computationally efficient way. In contrast, the approach proposed by Wang *et. al.* (Wang et al., 2011) uses complete video sequences to compute the motion descriptor. It characterizes in average 30 000 trajectories for each video with descriptors of size 426 for each trajectory. Afterwards, a bag of features is applied to reduce the feature space. Nevertheless, this approach remains limited to off-line applications. Other approaches like (Tabia et al., 2012) and (Zhang et al., 2008) use polar space representations to characterize activities. However they compute their descriptors on entire sequences, thus do not explicitly provide on-line recognition capabilities. Likewise, the local features coded in their descriptors are in most cases appearance-dependent and do not provide a kinematic description.

Regarding the spatial representation, we also tested a log-polar representation that focuses the attention in regions located near the center of mass. This representation is not convenient for actions that are more discriminant in peripheral regions, like waving or pointing. In average, for the orientation of velocity, the log-polar representation achieves an accuracy of 91.19%, and 83% for KTH and UT datasets, respectively.

Figure 3 illustrates the typical action recognition of the proposed method at the frame-level and for dif-

Table 1: Action classification by using different types of trajectories and computing different kinematic features, corresponding to the norm s or the orientation θ of the velocity, and to the curvature κ.

Trajectory	KTH			UT-Interaction		
	θ	s	κ	θ	s	κ
Semi-dense trajectory	92.24	87.13	87.25	85.0	77.1	85.00
Dense trajectory	91.19	90.26	89.33	85.0	71.6	85.00
Improved Trajectory	**94**.00	90.96	90.96	**88.7**	80.0	87.3

Table 2: Action recognition results on KTH dataset for complete video sequences, compared to different state-of-the-art approaches.

KTH	
Method	Acc
Wang *et. al.* (Wang et al., 2011)	94.2
Proposed approach	**94.00**
Laptev *et. al.* (Laptev et al., 2008)	91.8
Motion context (Zhang et al., 2008)	91.33
Polar flow histogram (Tabia et al., 2012)	82.33

Table 3: Action recognition results on UT-interaction dataset for complete video sequences, compared to different state-of-the-art approaches.

UT	
Method	Acc
Proposed approach	**88.7**
Laptev *et. al.* (Kantorov and Laptev, 2014)	87.6
Yu *et. al.* (Yu et al., 2010)	83.3
Daysy (Cao et al., 2014)	71

ferent videos-sequences. The proposed approach recognizes activities in partial video sequences, taking in general around of 30 frames to stabilize on a proper label. This motion descriptor achieves a stable recognition thanks to the integration of the different time depth versions of the histograms, that allows a consistent multiscale spatio-temporal description of the activity.

The proposed approach achieves both fast computation and low memory footprint, thus allowing efficient on-line frame-level recognition. The implementation of the proposed approach was not particularly optimized for these experiments. However the computation of the motion descriptor is very fast, taking in average 0.15 milliseconds for each frame. Additionally, the mapping of each motion descriptor to the SVM model at each time takes in average 8 milliseconds. The experiments were carried out on a single core i3-3240 CPU @3.40GHz.

4 CONCLUSIONS

This paper presented a motion descriptor for the on-line recognition of human actions based on the spatio-temporal characterization of semi-dense trajectories. The proposed approach achieved competitive results on different public datasets while being intrinsically computationally efficient. A polar grid designed over each frame allows to code spatial regions of the activity. In each region, motion histograms are updated at each frame according to the dynamic of the trajectories. These histograms form a motion descriptor that is capable of recognition at each frame and then for partial video sequences. Integrated at different time depths, such descriptors represent multiscale statistics of kinematics features from the trajectories, and they are relatively independent on the visual appearance of the objects. The proposed approach can be extended to multiple actions by computing several polar regions of interest and then characterizing each of them individually.

ACKNOWLEDGEMENTS

This research is funded by the RTRA Digiteo project MAPOCA.

REFERENCES

Cao, T., Wu, X., Guo, J., Yu, S., and Xu, Y. (2009). Abnormal crowd motion analysis. In *Int. Conf. on Robotics and Biomimetics*, pages 1709–1714.

Cao, X., Zhang, H., Deng, C., Liu, Q., and Liu, H. (2014). Action recognition using 3d daisy descriptor. *Mach. Vision Appl.*, 25(1):159–171.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. (2009). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939.

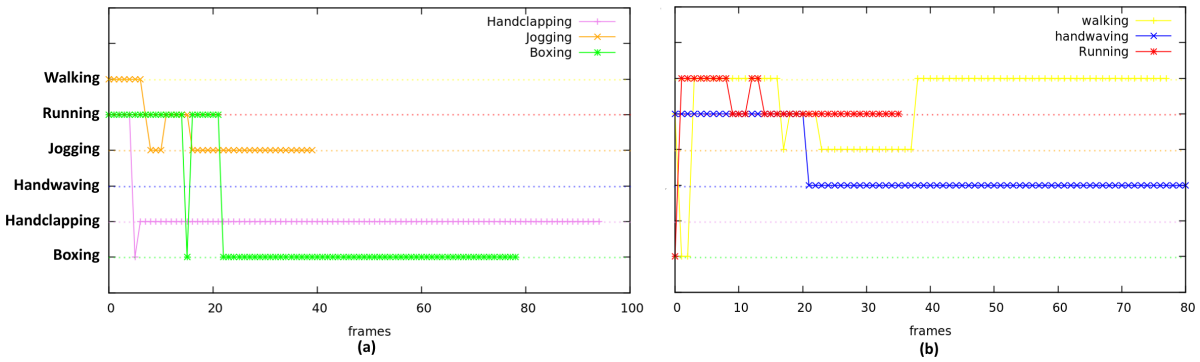

Figure 3: On-line action recognition for different videos recording several human motion activities. The frame-level recognition is carried out by mapping the motion histograms to the different SVM models at each time. Then action label with minimal distance to the hyperplanes is assigned to the motion descriptor.

Efros, A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing Action at a Distance. In *Int. Conf. on Computer Vision*, Washington, DC, USA.

Garrigues, M. and Manzanera, A. (2012). Real time semi-dense point tracking. In Campilho, A. and Kamel, M., editors, *Int. Conf. on Image Analysis and Recognition (ICIAR 2012)*, volume 7324 of *Lecture Notes in Computer Science*, pages 245–252, Aveiro, Portugal. Springer.

Ikizler, N., Cinbis, R., and Duygulu, P. (2008). Human action recognition with line and flow histograms. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4.

Jain, M., Jegou, H., and Bouthemy, P. (2013). Better exploiting motion for better action recognition. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 2555–2562, Washington, DC, USA. IEEE Computer Society.

Kantorov, V. and Laptev, I. (2014). Efficient feature extraction, encoding and classification for action recognition.

Ke, Y., Sukthankar, R., and Hebert, M. (2005). Efficient visual event detection using volumetric features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 166–173 Vol. 1.

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*.

Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos "in the wild". *IEEE International Conference on Computer Vision and Pattern Recognition*.

Ryoo, M. S. and Aggarwal, J. K. (2010). UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA).

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local svm approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, pages 32–36, Washington, DC, USA. IEEE Computer Society.

Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. pages 357–360.

Tabia, H., Gouiffes, M., and Lacassagne, L. (2012). Motion histogram quantification for human action recognition. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2404–2407. IEEE.

Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3169–3176, Washington, DC, USA. IEEE Computer Society.

Wang, H. and Schmid, C. (2013). Action Recognition with Improved Trajectories. In *ICCV 2013 - IEEE International Conference on Computer Vision*, pages 3551–3558, Sydney, Australia. IEEE.

Wang, H., Ullah, M. M., Kliser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *BMVC*. British Machine Vision Association.

Weinland, D., Ronfard, R., and Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241.

Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg. Springer-Verlag.

Yu, T.-H., Kim, T.-K., and Cipolla, R. (2010). Real-time action recognition by spatiotemporal semantic and structural forest. In *Proceedings of the British Machine Vision Conference*, pages 52.1–52.12. BMVA Press. doi:10.5244/C.24.52.

Zhang, Z., Hu, Y., Chan, S., and Chia, L.-T. (2008). Motion context: A new representation for human action recognition. *Computer Vision–ECCV 2008*, pages 817–829.