REPRESENTACIÓN DE PATRONES ESPACIO-TEMPORALES PARA EL RECONOCIMIENTO DE ACCIONES

GUSTAVO ADOLFO GARZÓN VILLAMIZAR

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2019

REPRESENTACIÓN DE PATRONES ESPACIO-TEMPORALES PARA EL RECONOCIMIENTO DE ACCIONES

GUSTAVO ADOLFO GARZÓN VILLAMIZAR

Trabajo de investigación para optar al título de
Magíster en Ingeniería de Sistemas e Informática

Director:
Fabio Martínez Carrillo
Ph.D en Ingeniería de Sistemas y Computación

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2019

DEDICATION

To my family, for such infinite tolerance and affection. To my wife, for giving color to my life and for that immense love that never runs out. To my son, for his little laugh that illuminates our home. To Daniel and Mario, for not losing touch, for the good advices and for balancing my points of view. To BivL^2ab members, for their support and such epic work atmosphere.

ACKNOWLEDGMENTS

The completion of this work was possible thanks to:

Professor Fabio Martínez Carrillo, whose infinite patience, motivation, advice and endless manuscript reviewing sessions, made these results possible. But more importantly: for showing himself as a human being.

Professor Lola Bautista, for the timely and helpful guidance to BivL^2ab members and the valuable feedback on our projects.

Gladys Noriega, for her hard work and the countless consultations to which she gave solution.

Lady Gutierrez, for the extensive collaboration in academic and administrative matters.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

pag.

LIST OF APPENDICES

RESUMEN

TÍTULO: REPRESENTACIÓN DE PATRONES ESPACIO-TEMPORALES PARA EL RECONOCIMIENTO DE ACCIONES [*]

AUTOR: GUSTAVO ADOLFO GARZÓN VILLAMIZAR [**]

PALABRAS CLAVE: DESCRIPTOR ESPACIO TEMPORAL, RECONOCIMIENTO DE ACCIONES, RECONOCIMIENTO DE PATRONES, DESCRIPTORES COMPACTOS.

DESCRIPCIÓN:

El reconocimiento de acciones es una tarea fundamental en diferentes áreas del conocimiento y aplicaciones, tales como: sistemas de realidad virtual, aplicaciones de tele-vigilancia, control de multitudes, videojuegos, entre otros. Esta tarea consiste en etiquetar, identificar o segmentar espacialmente objetos de interés que desarrollan una acción durante un intervalo temporal específico en una secuencia de video. A pesar de los grandes avances reportados en el estado del arte para el reconocimiento autómatico de acciones, existen múltiples limitaciones en cuanto a la caracterización y representación de las acciones debido a la complejidad de los escenarios de captura y la variabilidad del entorno. Además, la variabilidad de la apariencia, la geometría y la cinemática de las acciones humanas representan un desafío hoy en día en la comunidad de visión por computador.

El método propuesto presenta dos variaciones para calcular descriptores compactos de ocurrencias orientados a reconocer acciones en secuencias parciales de video. Para ello se obtienen trayectorias de movimiento que representan la actividad en desarrollo. Sobre ellas se aplica un proceso de conteo que se encuentra acotado por una región y se centra en cada trayectoria. En este sentido, se obtienen las ocurrencias de trayectorias vecinas respecto a cada centro, dentro de cada subregión acotada, o se codifican de manera binaria en caso de cumplir con un umbral determinado (Minimal number of trajectories MNT). Se obtiene una representación intermedia al construir un diccionario con 400 valores que luego se mapea a un histograma que codifica las características de la acción que se esté ejecutando. Los dos métodos obtienen resultados prometedores en la clasificación de

---

[*] Trabajo de investigación

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, PhD.

acciones utilizando un histograma de dimensionalidad reducida.

ABSTRACT

TITLE: SPATIO-TEMPORAL PATTERN REPRESENTATION FOR ACTION RECOGNITION [*]

AUTHOR: GUSTAVO ADOLFO GARZÓN VILLAMIZAR [**]

KEYWORDS: SPATIO-TEMPORAL DESCRIPTOR, ACTION RECOGNITION, PATTERN RECOGNITION, COMPACT DESCRIPTOR.

DESCRIPTION:

Action recognition is a fundamental task in different areas and applications such as virtual reality systems, surveillance, video games and others. This task consists in labeling, identifying or spatially segmentate objects of interest that perform an action during a specific time interval in video sequences. Despite significant advances in the state of the art for action recognition, there exist multiple limitations regarding characterization and representation of actions due to complex video adquisition scenarios and variability of the environment. Moreover, variability of appearance, geometry and kinematics of human actions represent a challenge for the computer vision community.

The proposed method exhibit two variations that are able to calculate compact occurrence descriptors that recognize actions over frame level, partial and full video sequences. Firstly, our methods obtain motion trajectories that represent the performed activity inside a given video file. Then, a counting process is executed over bounded region that is centered on each trajectory. This allows to identify neighboring trajectories inside each bounded subregion or, are codified as binary values if they comply with a given threshold (Minimal number of trajectories MNT). Therefore, a midlevel representation is obtained by calculating a dictionary with 400 values that will be mapped to a histogram that codify the features of the current human action. Both methods obtained promising results on classification and recognition tasks using a reduced size histogram.

---

[*] Research work

[**] Faculty of Physical-Mechanical Engineering. Department of Systems Engineering and Informatics. Advisor: Fabio Martínez Carrillo, PhD.

INTRODUCTION

Recognizing human actions is a fundamental task in many areas and applications, such as surveillance and crowd control Takahashi et al., 2011; [1] automatic annotation of human actions in videos [2] and video indexing [3], analysis of sports videos [4], HCI applications [5] and gesture-based video games interaction [6], among other examples [3] [7]. Nevertheless, such applications hardly offer ideal conditions with respect to environmental factors which difficult the action characterization. The typical challenges are reported because of scene variations such as different shapes and clothing, scale changes and movement on the background. Strong variations of the object of interest with respect to the geometry, appearance and motion patterns can also difficult the task of identification and recognition.

---

[1] Masaki Takahashi y col. "Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories". En: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE. 2011, págs. 9-16.

[2] Ivan Laptev y col. "Learning realistic human actions from movies". En: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, págs. 1-8.

[3] T Subetha y S Chitrakala. "A Survey on human activity recognition from videos". En: *Information Communication and Embedded Systems (ICICES), 2016 International Conference on*. IEEE. 2016, págs. 1-7.

[4] Mikel D Rodriguez, Javed Ahmed y Mubarak Shah. "Action mach a spatio-temporal maximum average correlation height filter for action recognition". En: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, págs. 1-8.

[5] Ying Wu y Thomas S Huang. "Vision-based gesture recognition: A review". En: *Gesture Workshop*. Vol. 1739. Springer. 1999, págs. 103-115.

[6] Guangming Zhu y col. "An online continuous human action recognition algorithm based on the kinect sensor". En: *Sensors* 16.2 (2016), pág. 161.

[7] Ronald Poppe. "A survey on vision-based human action recognition". En: *Image and vision computing* 28.6 (2010), págs. 976-990.

Action recognition have been approached with strategies that rely on action representation. As such, state of the art exhibit methods that use similar approaches, namely: representation of actions using human silhouettes, spatio-temporal patches representation (keypoints and spatio-temporal invariant interest points), motion primitives representation (optical flow based and long trajectories) and exhaustive learning strategies that aim to comprehend motion patterns using significant amounts of data.

The main contribution of this work is a compact action descriptor that codify distribution and occurrences of motion trajectories over a bounded region, in order to find spatio-temporal patterns that correspond to human actions, operating on frame-level and full video sequences. Such compact descriptor proved to be effective for action recognition specially over hardware with limited performance and, therefore, high-performance devices are not required to execute the proposed method.

RELATED WORK

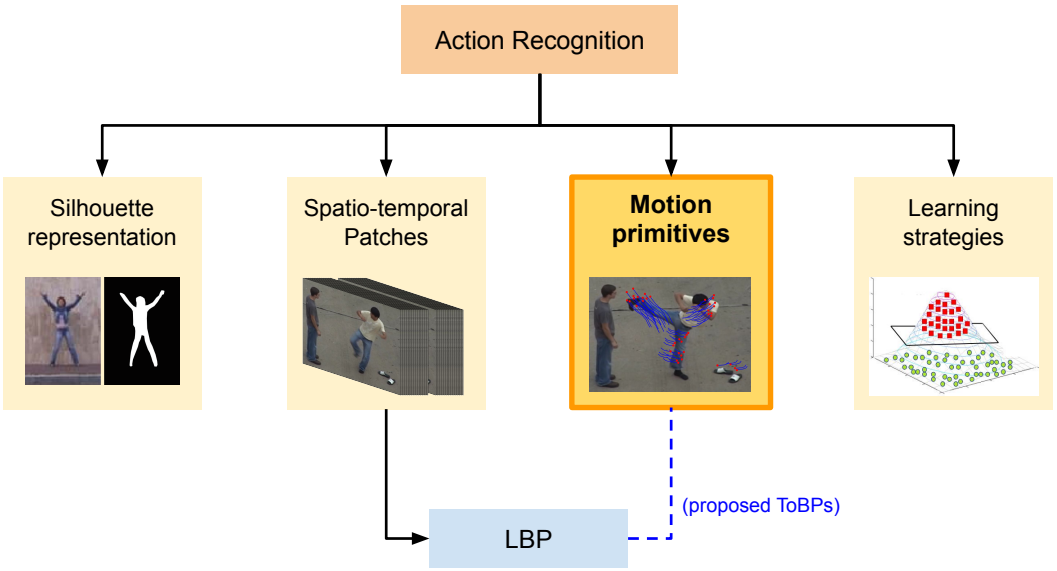
Representations of human actions have been studied in the last decades with strategies involving the use of shape-based strategies, local-interest point approaches, as well as motion kinematic primitives and dynamic texture descriptors such as Local Binary Patterns (LBPs). The following section shows an overall view on these methods, and a more detailed subsection regarding LBPs.

Global shape strategies involve the temporal segmentation of regions of interest and the association of the geometrical shape characterization with particular actions. For instance, Bobick et al. [8] studied the motion presence in video sequences by matching

[8] Aaron F. Bobick y James W. Davis. “The recognition of human movement using temporal templates”. En: *IEEE Transactions on pattern analysis and machine intelligence* 23.3 (2001), págs. 257-267.

temporal templates against stored shapes of fixed actions. This method uses a two-component setup in which a binary motion-energy image (MEI) represents where motion has occurred in an image sequence and a motion-history image (MHI) which is a scalar-valued image where intensity is a function of recency of motion. In order to classify any input sequence, a Mahalanobis distance is obtained from the description of input and each of the known movements. If multiple matches are obtained, the smallest distance is chosen. This method reported to be robust to linear changes in speed, and performs well in real-time applications but with limitations to perspective of capture and also, it assumes a static background.

Figure 1. Action recognition methods reported on the state of the art



Also, Gorelick et al. [9] proposed to analyze actions as volumetric shapes computed from silhouettes along video. This method uses the solution of the Poisson equation over a spatio-temporal shape S that exist inside a bounded surface. A set of particles close to each point $(x, y) \in S$ will perform a random walk until boundaries of S

[9] Lena Gorelick y col. "Actions as space-time shapes". En: *IEEE transactions on pattern analysis and machine intelligence* 29.12 (2007), págs. 2247-2253.

are touched. Average time of random walks is calculated using the solution of a Poisson equation with artificial boundary conditions for temporal limits (namely $t = t_0$ and $t = t_f$). Such solution allow to obtain features such as shape, saliency and spatio-temporal orientation. Default solution is widely used for the spatial domain but the existence of time add new objects that are richer in information compared to previous ones. Such proposed approach achieves relative invariance to scale-viewpoint changes but it is dependent of a proper silhouette computation to describe the volumes. Also, this method is dependent on resizing video sequences to fixed values.

Junejo et al. [10] introduced a silhouette representation that combines time series and Symbolic Aggregate approXimation (SAX). Firstly, human silhouettes are extracted using [9] method. Secondly, such shape is converted to a time series and a brute force approach is applied to deal with rotational misalignment. Said series are then transformed to SAX elements by defining a number of horizontal segments combined with fixed breakpoints to generate coefficients. Thereafter, a mapping process converts such information into a SAX representation of symbols. Similarly, Salim et al. [11] proposed a combination of Cartesian Coordinate features, Fourier Descriptors among others that complemented with appearance-based histograms achieved an action recognition. Such approaches are relatively fast but dependent of controlled scenarios of capture and are vulnerable when facing actions with similar kinematic structure.

On the other hand, local-interest point representations have focused on statistical frameworks that combine appearance and inter-frame-based features to recognize

---

[10] Imran N Junejo, Khurrum Nazir Junejo y Zaher Al Aghbari. “Silhouette-based human action recognition using SAX-Shapes”. En: *The Visual Computer* 30.3 (2014), págs. 259-269.

[11] Salim Al-Ali y col. “Human action recognition: contour-based and Silhouette-based approaches”. En: *Computer Vision in Control Systems-2*. Springer, 2015, págs. 11-47.

actions. For instance, Schuldt et al. [12] used local space-time features to represent motion schemes in image sequences. Such setup involved an internal scale-space representation that convoluted image frames with a gaussian kernel. Next, a second-moment matrix was calculated with spatio-temporal image gradients established inside a gaussian neighborhood w.r.t each point. This produced an expression that is processed using a Harris corner function, whose local maxima is associated with positions of said features. Thereafter, a support vector machine classification strategy is executed in order to find the hyperplane that maximizes the separation margin boundary.

Also, Laptev [13] proposed a spatio-temporal invariant interest points that maximizes a normalized spatio-temporal Laplacian operator over spatial and temporal scales. Later, Laptev and Lindeberg [14] proposed a space-time interest point representation that is projected to a lower-dimensional space to recover a main vector of representation. This strategy compared motion representations based on: spatio-temporal jets, position dependent histograms, position independent histograms, and principal component analysis computed over spatio-temporal gradients or optical flow. Evaluation shown that a competitive classification score is obtained by using local position dependent histograms. These approaches outperform occlusion problems but remain limited to motion direction and appearance variability.

Motion primitives, such as optical flow and long trajectories, have been incorporated in local statistical frameworks to develop strategies robust to appearance changes.

---

[12] Christian Schuldt, Ivan Laptev y Barbara Caputo. “Recognizing human actions: a local SVM approach”. En: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. IEEE. 2004, págs. 32-36.

[13] Ivan Laptev. “On space-time interest points”. En: *International journal of computer vision* 64.2-3 (2005), págs. 107-123.

[14] Ivan Laptev y Tony Lindeberg. “Local descriptors for spatio-temporal recognition”. En: *Lecture notes in computer science* 3667 (2006), págs. 91-103.

For instance, Wang *et. al.* [15] proposed motion trajectories as sample feature points tracked according to a local optical flow directions. Around the computed motion trajectories were computed local volumes described by Histogram of gradients (HoG), histograms of optical flow (HoF) and histograms of motion boundaries (MBH). Such volumes were used to built a Bag-of-words (BoW) representation of actions in video sequences. This strategy achieved action recognition in uncontrolled videos but with some restrictions for abrupt camera motions. To outperform such problems, in [16] was proposed a set of improved motion trajectories that filter background and camera motions by using additional matching restrictions from SURF descriptors and taking into account the homography parameters from RANSAC algorithm. Results showed a significant accuracy but additional restrictions increased the computational cost with respect to the dense field strategy.

Additionally, recent action recognition methods are based on deep learning and recurrent architectures that codify features according to the correlation of thousands of video samples. Such strategies have demonstrated high capability to recognize human actions at different scenarios. For example, Baccouche et al. [17] proposed a double-step deep model for action recognition that firstly extend two-dimensional convolutional networks to the 3D scenario in order to learn spatio-temporal features. These features are then used to train a LSTM that will execute action classification for full sequences. Similarly, Rahmani et al. [18] introduced a method that used artifi-

---

[15] Heng Wang y col. “Dense trajectories and motion boundary descriptors for action recognition”. En: *International journal of computer vision* 103.1 (2013), págs. 60-79.

[16] Heng Wang y Cordelia Schmid. “Action recognition with improved trajectories”. En: *Proceedings of the IEEE international conference on computer vision*. 2013, págs. 3551-3558.

[17] Moez Baccouche y col. “Sequential deep learning for human action recognition”. En: *International Workshop on Human Behavior Understanding*. Springer. 2011, págs. 29-39.

[18] Hossein Rahmani, Ajmal Mian y Mubarak Shah. “Learning a deep model for human action recognition from novel viewpoints”. En: *IEEE transactions on pattern analysis and machine intelligence*

cial 3D human models from different viewpoints to extract dense trajectories, which are represented with a bag of features. Afterwards, a deep fully connected network learned trajectory descriptors from different viewpoints but correspondant to the same human action. Lastly, real video sequences become the input for that learned network, obtaining cross-view action descriptors that is mapped to a Support Vector Machine to obtain action labels. Nevertheless, these approaches demand large training sets and the adjustment of hyper-parametric functions that require special computational demands to achieve proper results.

LOCAL BINARY PATTERNS (LBP) ON ACTION RECOGNITION

One of the most interesting local descriptions for computer vision problems correspond to Local Binary Patterns (LBP) [19][20]. The main idea behind LBP descriptors is to compute local differences around of a central point, and then generate a bit-vector quantification using a binary representation. If a neighborhood value is larger than central point then the bit is marked as one, otherwise the bit is fixed as zero. This compact representation allows a more stable description of local features to solve several problems, such as, texture modelling, pattern recognition, and even to model local dynamic patterns by extending the concept at consecutive frames [21].

Regarding action recognition, several strategies have exploited LBP coding to locally

---

40.3 (2018), págs. 667-681.

[19] Timo Ojala, Matti Pietikäinen y David Harwood. "A comparative study of texture measures with classification based on featured distributions". En: *Pattern recognition* 29.1 (1996), págs. 51-59.

[20] Timo Ojala, Matti Pietikainen y Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". En: *IEEE Transactions on pattern analysis and machine intelligence* 24.7 (2002), págs. 971-987.

[21] Thierry Bouwmans y col. "On the role and the importance of features for background modeling and foreground detection". En: *Computer Science Review* 28 (2018), págs. 26-91.

represent actions. For instance, Nanni et al. [22] proposed a strategy that used a masked image that is divided in 7x7 subregions, where each one produced a histogram that is obtained by adding characteristics from Local Ternary Patterns (LTP) and Local Binary Patterns at Three Ortogonal Planes (LBP-TOP). Particularly, LTP stage produced three posible values $\{0, 1, -1\}$ as the outcome of center-neighbor comparison and introduces a threshold τ that improves discriminative power of simple LBPs. Thereafter, the authors apply LTP at a current frame i to planes xt and yt, and a given number of previous and later frames are considered in order to create a volume. Afterwards, a random subspace ensemble of support vector machines obtains a label using the sum rule. Nevertheless, this approach is highly dependent on masked human silhouettes, is sensible to occlusion and require a proper background computation.

Yeffet et al. [23] proposed to recognize actions using local trinary patterns by coding differences among consecutive frames during time. Then, in partial intervals of the video are computed occurrence histograms of the obtained local trinary patterns with values $\{0, 1, -1\}$. This process is calculated over a grid of patches in order to estimate motion coherence. Therefore, a sum of square differences (SSD) is employed to calculate similarities between patches at frame t with frames $t - 3$ and $t + 3$. The obtained information is mapped to a linear kernel support vector machine, and further experiments are run on several segments of sequences in parallel. This approach achieve interesting results but its computation is dense and requires rigid splitting of videos.

---

[22] Loris Nanni, Sheryl Brahnam y Alessandra Lumini. “Local ternary patterns from three orthogonal planes for human action classification”. En: *Expert Systems with Applications* 38.5 (2011), págs. 5125-5128.

[23] Lahav Yeffet y Lior Wolf. “Local trinary patterns for human action recognition”. En: *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE. 2009, págs. 492-497.

Nguyen et al. [24] also proposed to codify textures and motion using uniform LBP and Local Trinary Pattern (LTP) respectively, from neighboring image patches and a set of tracked points. Such technique, namely Spatial Motion Patterns (SMP), allowed to codify distinct velocities in local positions, but sequences with a significant amount of background motion might not perform well since only one spatial scale was considered for the analysis.

The LBP scheme has also been extended on depth sequences [25][26] for action classification and partial recognition. In such cases, the LBP descriptors are computed from Depth Motion Maps (DMMs) to complement the dynamic texture representation.

---

[24] Thanh Phuong Nguyen y col. “Revisiting lbp-based texture models for human action recognition”. En: *Iberoamerican Congress on Pattern Recognition*. Springer. 2013, págs. 286-293.

[25] Yun Yang y col. “Action recognition using completed local binary patterns and multiple-class boosting classifier”. En: *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*. IEEE. 2015, págs. 336-340.

[26] Rawya Al-Akam, Salah Al-Darraji y Dietrich Paulus. “Human Action Recognition from RGBD Videos based on Retina Model and Local Binary Pattern Features. In 26”. En: *Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*. 2018, págs. 1-7.

1. RESEARCH PROBLEM

Nowadays, many advances have been made to recognize activities in increasingly complex scenarios, but it is almost always assumed that the activity is present from the beginning to the end of the input sequence, *i.e.,* its temporal support interval is *a priori* delimited. Despite significant efforts, the proposed methods are therefore very sensitive to phase shift of actions, which induces considerable limitations in real-time applications, where sequences are untrimmed and which require any-time prediction capability, for instance in surveillance or online video indexing systems.

Nevertheless, such complex scenarios hardly offer ideal conditions with respect to environmental factors which difficult action characterization. The typical challenges are reported because of scene variations such as different shapes and clothing, scale changes and movement on the background. Strong variations of the object of interest with respect to the geometry, appearance and motion patterns can also difficult the task of identification and recognition.

Research question

Taking this context into account, the following question is posed: How can we codify spatio-temporal primitives at frame-level for action recognition?

2. OBJECTIVES

General objective

- To propose a computational strategy based on representations of space-time primitives for action recognition.

Specific objectives

- To obtain motion trajectories for video datasets.
- To characterize motion trajectories using kinematic or appearance primitives.
- To propose a spatio-temporal descriptor for action recognition.
- To evaluate the performance of the proposed descriptor for action recognition.

3. ACTION RECOGNITION FROM FRAME-LEVEL MOTION OCCURRENCE DESCRIPTORS

This chapter is part from two research articles: “A fast action recognition strategy based on motion trajectory occurrences” [27] and “Online action recognition from trajectory occurrence binary patterns (ToBPs)” published on the proceedings of the International Conference on Advances in Emerging Trends and Technologies (ICAETT 2019) [28].

Human visual system has developed robust mechanisms to perceive and recognize complex actions by coding coherent information. Such fact has been widely demonstrated in Johansson experiments, where actions were properly identified by analyzing few bright spots spatially distributed that changed over time [29]. In such experiments it was demonstrated that visual systems coded the spatial distribution of interest points that have been moving coherently through time. Inspired in such natural mechanisms, in this work is introduced a novel action recognition strategy based on the local spatial characterization of motion of points of interest. Such motion of interest are herein computed from *motion trajectories* (section **??**) that remain coherent in a temporal interval of time $\triangle t$ (see in Figure 2-a and Figure 3).

A two-layered *trajectory-occurrence* scheme (section **??**) is proposed to locally compute the existence of trajectories into subregions of a circular grid that is fixed at each

---

[27] Gustavo Garzon y Fabio Martinez. “A fast action recognition strategy based on motion trajectory occurrences”. En: *Pattern Recognition and Image Analysis* 29 (2019).

[28] Gustavo Garzon y Fabio Martinez. “Online action recognition from trajectory occurrence binary patterns (ToBPs)”. En: *Proceedings of the International Conference on Advances in Emerging Trends and Technologies*. 2019.

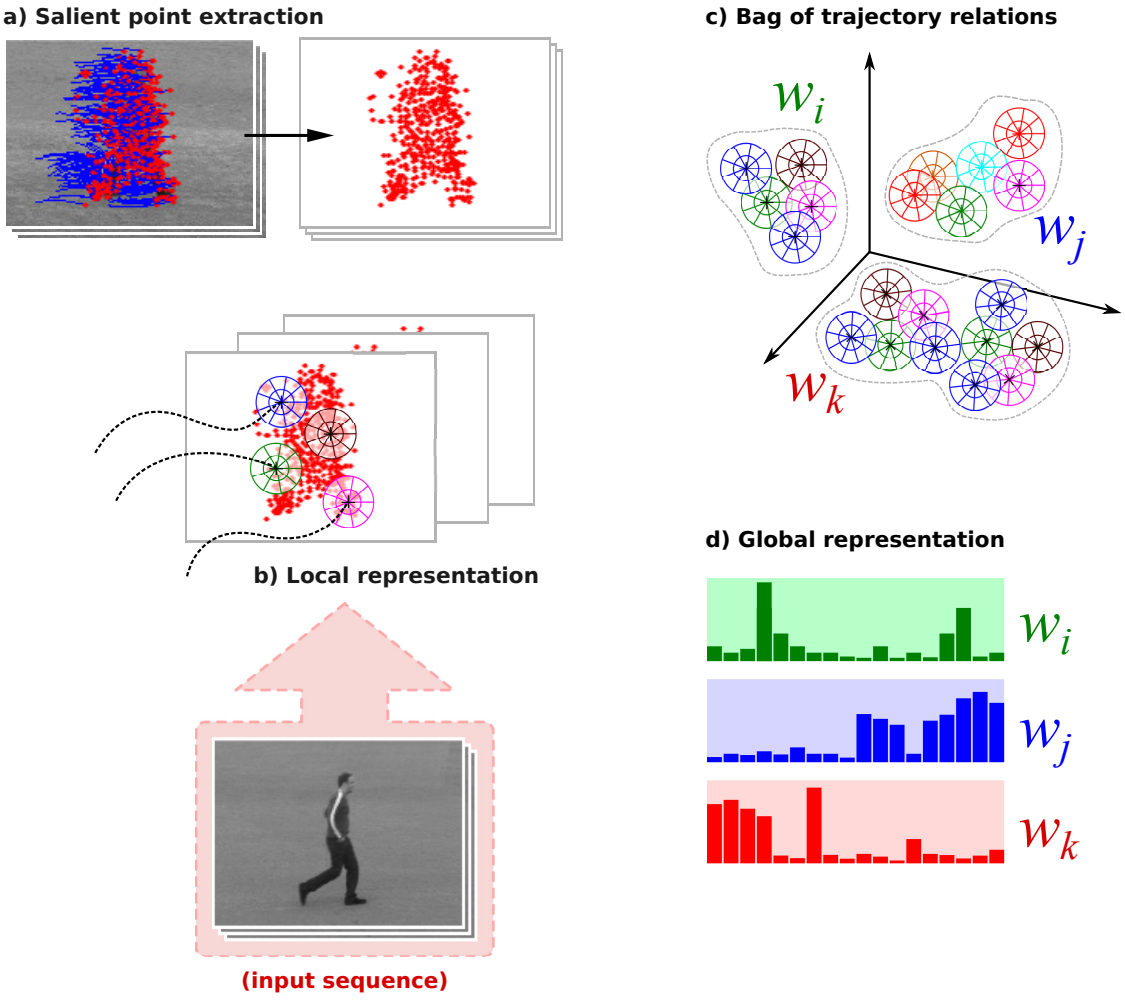
[29] Gunnar Johansson. “Visual perception of biological motion and a model for its analysis”. En: *Perception & psychophysics* 14.2 (1973), págs. 201-211.

frame and around of each active trajectory. Regarding first occurrence layer, two alternative approaches to locally characterize occurrences were proposed: a spatial counting process that regarded the spatial distribution of motion by using a counting process, or, a thresholded scheme that codify subregions from said counting process that featured a minimal required amount of motion as binary values. It is worth to clarify that, in this paragraph, motion is associated with a number of trajectories appearing in the current bounded region. A second occurrence layer is defined as a regional counting process of local atom descriptors, that is, a midlevel representation (section 3.3). For doing so, from a set of training videos is firstly computed a dictionary of local occurrence atoms. Then any action can be globally represented in the video sequences by projecting the spatial distribution atoms to the learned dictionary and obtaining a global histogram representation. This representation can be also computed at frame level, allowing an online action recognition.
Finally, such global/frame histograms are mapped to a previously trained Support vector machine to obtain an action label (section 3.4). The pipeline of the proposed method is illustrated in Figure 2.

3.1. MOTION TRAJECTORIES REPRESENTATION

The herein proposed work starts by computing a set of dense trajectories, allowing a primary motion representation of the action present in the video. As visual perception, such motion trajectories are a set of salient spots that are tracked in a set of frames and recover temporal coherence of local motion. In this work was implemented the improved motion trajectories proposed by Wang et al [15]. These dense motion trajectories have demonstrated a proper representation of actions, by following points of interest from a dense optical flow field. The spatio-temporal points of interest $P_t = (x_t, y_t)$ in frame I_t are tracked w.r.t. the velocity field direction $\omega = (u_t, v_t)$, and are smoothed by using a median filter M at different spatial scales (spaced with

Figure 2. Proposed approach: (a) salient points are tracked for the input video sequence in order to obtain a (b) local representation based on occurrence of neighboring trajectories. This is then represented as a (c) Bag of Spatial Representation Words (BoSRW) that will codify a (d) global descriptor for the entire video sequence.



a factor of $1/\sqrt{2}$), resulting in a tracked position in frame I_{t+1}:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\overline{x}_t, \overline{y}_t)}$$

with $(\overline{x}_t, \overline{y}_t)$ being the rounded position of (x_t, y_t). Then, each trajectory is expressed as the concatenation of points $(P_t, P_{t+1}, P_{t+2}, ..., P_{t+n})$ along time.

Formally, the set of trajectories $\mathbf{T} = \{\tau_\mathbf{1}, \tau_\mathbf{2}, ..., \tau_\mathbf{N}\}$ can be considered as paths that appear during the course of a video, and describe the motion of objects. Each path trajectory $\tau_\mathbf{i} = \{\tau_{t_1}, \tau_{t_2}, \ldots, \tau_{t_k}\}$ is a set of k spatial points in certain consecutive frames within the video.

Once the raw trajectories are computed, some statistical motion filters are implemented to remove trajectories with insignificant or incoherent motion information, according to the local variation of norm velocity. The estimation of these trajectories was also improved by taking into account a camera motion correction assuming a homography relationship between consecutive frames.

A per-frame action representation is defined as the set of spatial points that correspond to *active* trajectories, described as the spatial position τ_{t_k} of each trajectory τ_i on the current frame.

An example of computed motion trajectories are illustrated in figure 3 and figure 4 for different activities recorded in open scenarios. As illustrated in the second row, the red points indicate the frame representation from motion trajectories, while the blue lines code the motion history of each red point. In the third row is shown the point representation that globally can describe the actions by the spatial coding configuration. The local distribution of salient points is defined as a spatial point process around motion trajectories.

Figure 3. *First row:* a set of frames of six different actions from KTH dataset. *Second row:* motion trajectories for six different actions. *Third row:* red pixels indicating active trajectories for each frame.

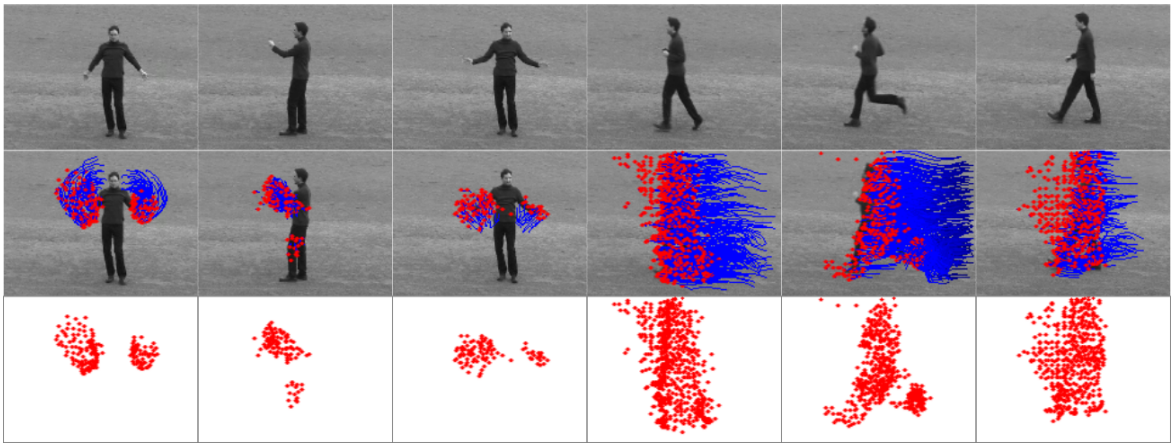
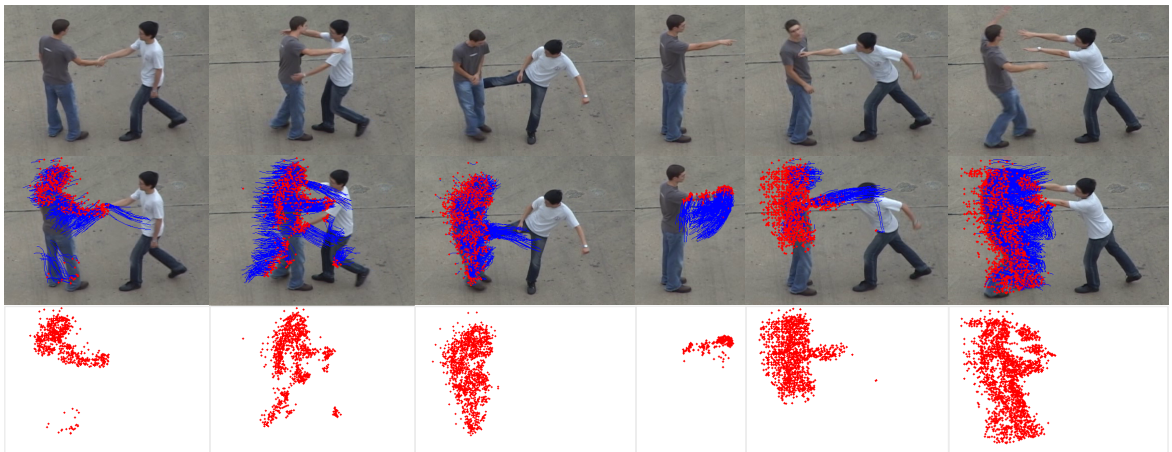


Figure 4. *First row:* a set of frames of six different actions from UT-Interaction dataset. *Second row:* motion trajectories for six different actions. *Third row:* red pixels indicating active trajectories for each frame.



3.2. LOCAL OCCURRENCE CHARACTERIZATION

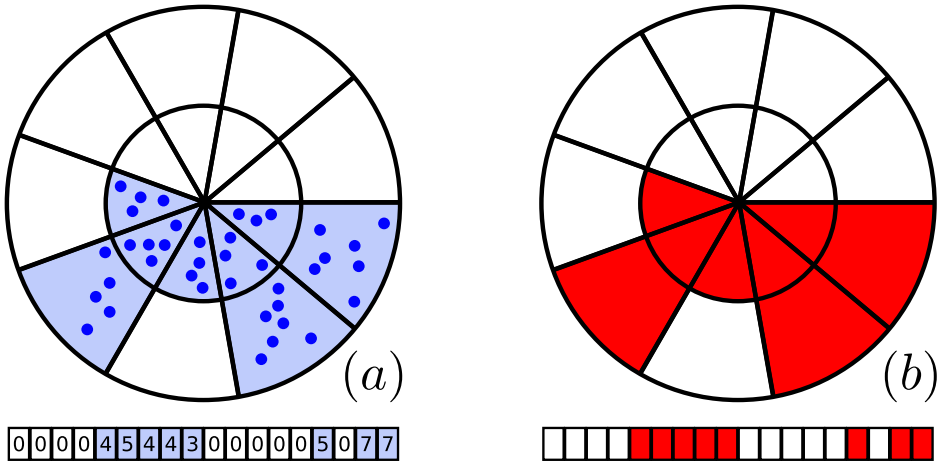
Motion trajectories tend to form spatial clusters in regions with a significant number of active points. This fact implies the existence of motion patterns that can be measured in local regions centered around active points. A local point representation is herein achieved by representing each active point trajectory in frame t as the spatial distribution of neighborhood trajectories. In such case the signature of each point is defined by the spatial density of points in its neighborhood. Hence neighborhood positions describe events at random locations, formulated as a finite-set-valued random variable $\mathbf{u} = \{u_1(\tau), u_2(\tau), ..., u_n(\tau)\}$. Those atomic point distributions represent a local spatial point process of motion trajectories contained in a bounded local region.

Aforesaid region is composed by γ concentric circles and α angular divisions. This setup formed subregions $r_i \in \mathbb{R}^2$ that will contain occurrences of neighboring trajectories with respect to the active trajectory that is positioned in the center of said circular region.

The counting process that take place over this circular region is performed using either of two approaches, namely: a spatial counting process that relies on spatial distribution of neighboring trajectories as a signature for actions, or, an occurrence binary pattern scheme that threshold the number of trajectories existing in each subregion r_i with respect to a minimal number of trajectories (MNT), as shown in figure 5.

These local representations of actions constitute the input for a mid-level representation, that will describe the spatial point distribution in a higher level. In such way, the proposed strategy comprises a hierarchical model that allows to characterize spatial point distributions from a local to a global level to represent actions like observed in natural visual perception systems.

Figure 5. Approaches for the proposed local counting process: (a) local spatial counting process and (b) local occurrence binary pattern scheme.



3.2.1. Local spatial counting process approach As for the first approach, a counting scheme is obtained by locally counting the occurrences of neighboring motion trajectories that fall inside the proposed circular bounded region. For doing so, a counting process is defined around of active point trajectories. A circular distribution is then achieved by counting motion neighboring trajectory points that fall into each subregion r_i of the circular grid. A general scheme of the circular counting process is depicted in Figure 6.

As illustrated in Figure 7, because the density of important motion regions, some of the point signatures can share counting information, producing redundant data in the representation. This fact result relevant in our description because important motion patterns can be represented in several atomic signatures. Then, any particular spatio-temporal boundary $\mathbb{C}$ of action sequence is represented by a set of l-dimensional occurrence points coded from a circular grid. The boundary $\mathbb{C}$ could be considered as total video sequence in classification tasks or at frame level for online action recognition.

Figure 6. Bounded region detail: position of neighboring trajectories is characterized using a circular grid layout corresponding to bounded local region $\mathbb{C}$. Trajectories that end (green), go through (blue) and start (orange) in the tracking interval L are taken into account. In this illustration, the spatial distribution is computed as a signature of the red point (bottom-right).

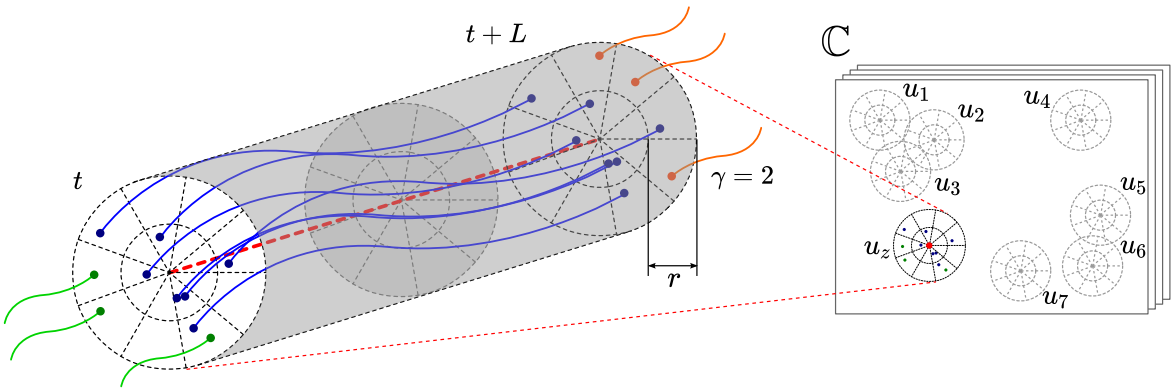


Figure 7. Grid overlapping on a per-frame representation. Relations between trajectories are efficiently characterized and represent a robust signature for some actions in a local perspective.

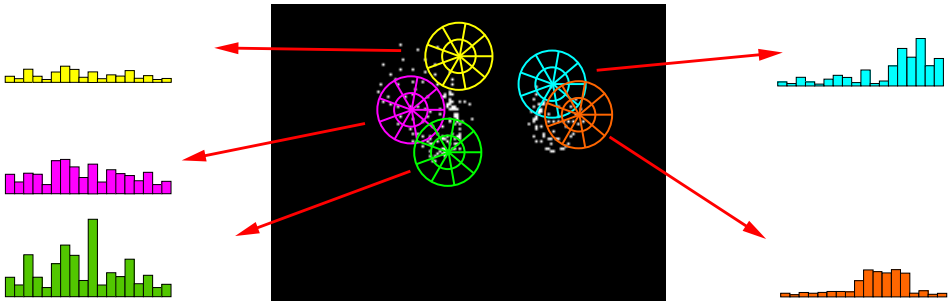
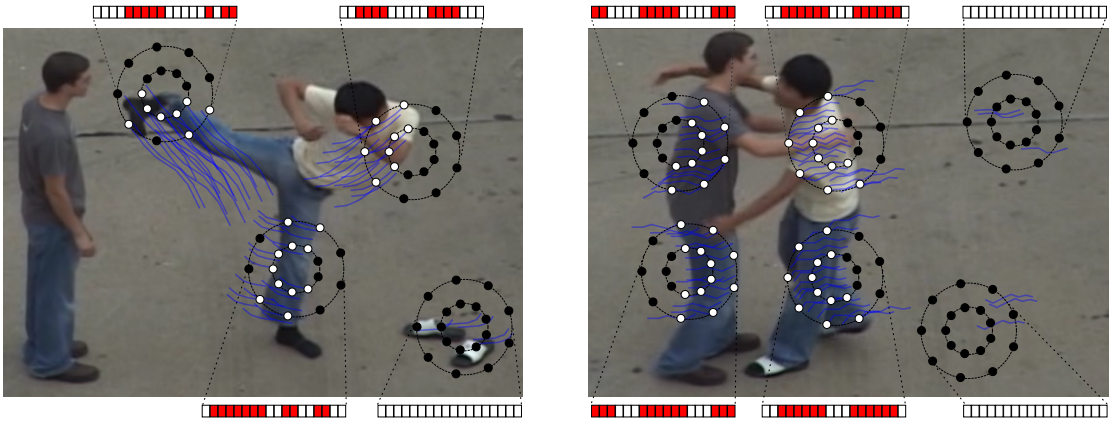


Figure 8. Detail of the proposed LBP-based scheme: bounded subregions r_i with a given motion density n_i are codified with a value of 1 (if $n_i \geq \tau$) or 0 otherwise. Subregions with not enough trajectories are excluded from the model.



3.2.2. Local occurrence binary patterns (ToBPs) A compact and robust representation is carried out by computing occurrence motion patterns as a LBP-like operator (ToBPs). For doing so, a bit string vector is formed by considering binary values to represent local occurrences of neighboring trajectories if they comply with a given threshold. Our strategy considers the number of trajectories n_i passing through region r_i as a component of a dynamic texture distribution T such that: $T \approx t(m_\tau(n_1), m_\tau(n_2), ..., m_\tau(n_k))$ where, a function m_τ compares each value n_i with a Minimal Number of Trajectories (MNT) τ which is detailed in figure 8.

The feature vector is then defined as $u := \sum_{1 \leq i \leq \gamma \cdot \alpha} 2^{i-1} m_\tau(n_i)$

The pipeline of the proposed local occurrence binary patterns approach is illustrated in Figure 9.

3.3. Mid-level representation

As is well known, visual mechanisms operate from fine to coarse different scale of analysis to understand and define the world around. The proposed work completes the local spatial representation by coding the circular distribution points and binary

pattern occurrences in a mid-level representation.

Firstly, a dictionary of representative spatial words is computed from a non supervised strategy over a set of training videos. For doing so, a k-means was herein implemented to compute k representative centroids, defined as: $D = [d_1, d_2, ...d_k] \in {}^{\mathbf{1}\times\mathbf{K}}$. Each of these centroids are recovered from an objective function: $D(k) = \min_{d_k} = \sum_{m=1}^{M}\sum_{k=1}^{K} \|\mathbf{u}(\tau)_m - d_k\|_2^2$, as the K points closer to each of the members of resulting groups. In this sense, the actions are assumed that can be globally described by the distribution of these centroids d_i in each particular video (see figure 9-c).

Once the dictionary is computed, a mid level action representation is obtained by coding the set of spatial words in global histogram occurrences w.r.t the learned centroids. This representation could be defined into a particular time interval Δ_t such as the complete video or even at level of a single frame. For each temporal interval, each computed word $\mathbf{u_n}$ contributes with the count occurrence only for the centroid i with minimal Euclidean distance s, process defined as: $s(i) = 1$, if $i = \arg\min_i \|\mathbf{u_n} - d_j\|^2$ **s.t.** $\|s\|_0 = 1$. A normalized version of the set of K centroid occurrences constitutes the histogram representation for the action in such particular interval of time.

A clear advantage of the Bag of Words scheme is that actions can be described from partial sequences and is also robust to occlusion problems during capture. Compared with circular distribution points at mid-level, this analysis is not directly related to the neighboring trajectories over regions, allowing scale invariance in the description of motion patterns. Redundant information enrich the description of actions and add favorable complexity to the model. This representation of motion clusters result in a compact scheme for action recognition and constitutes the input for classifying human action sequences.

Figure 9. Pipeline of the proposed LBP-based method: (a) Trajectories are calculated for frame t_i. (b) A circular grid is defined to perform a LBP-based measurement resulting in a vector of motion density. (c) Such vector is used to create a regional dictionary representation (d) An activity classification process assign a label $\hat{y}_{t_i}$ for each frame t_i of the video sequence.

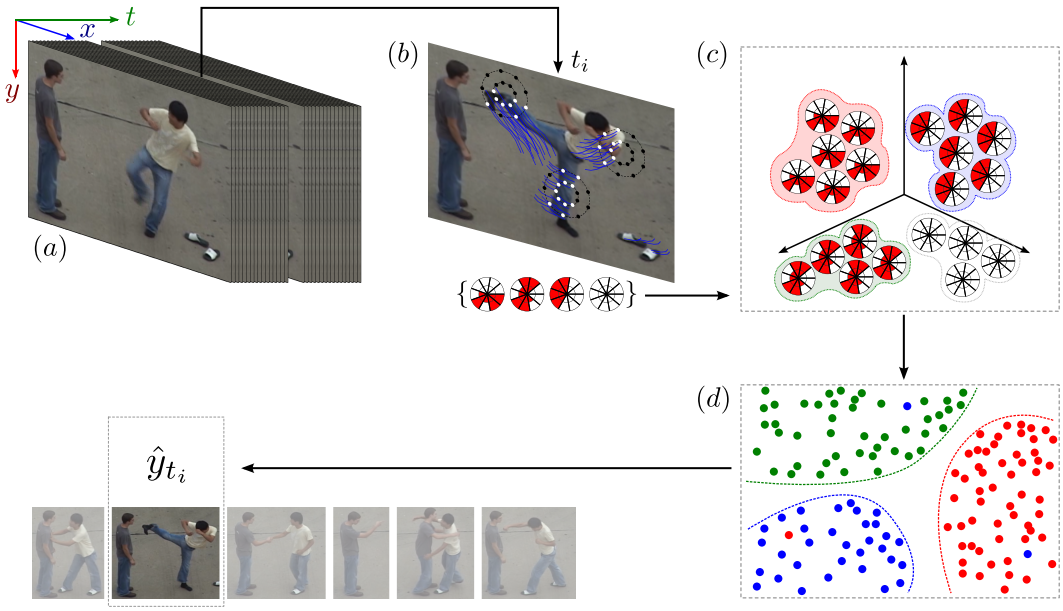
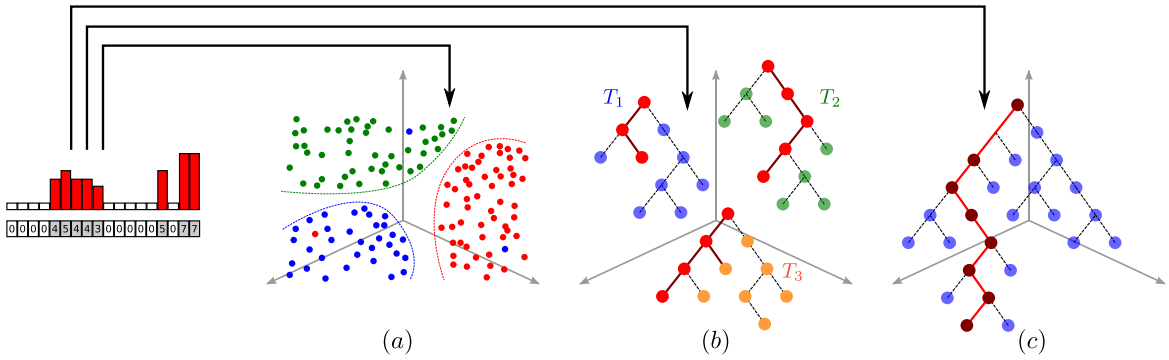


Figure 10. Different classification methods: (a) Support Vector Machines, (b) Random Forest and (c) Decision Trees.



3.4. Action classification and recognition

The computed spatio-temporal descriptor at the larger region ocurrence scale, coded from fine to coarse active point representation, is used as the input of machine learning strategies to obtain an automatic action recognition. To achieve a proper trade-off between prediction time and accuracy, in this work were evaluated different classification methods based on kernel functions and random forest strategies. On the one hand, a multi-class Support Vector Machine (SVM) was herein implemented by using a *One against one SVM multiclass classification* and evaluated using the linear and the Radial Basis Function (RBF) kernels [30]. The SVM with linear kernels allows to deal with fast action recognition, one of the main goal of the proposed work, while the RBF lead with non-linear class separations (figure 10-(a)).

On the other hand, the random forest (RaF) classification (figure 10-(b)) consists on a set of decision tree classifiers grouped in an Ensemble Learning strategy, allowing to overcome sensibility problems of decision by taking mode decisions over the set of computed trees. In our specific approach, a set of independent DT algorithms (figure

[30] Chih-Chung Chang y Chih-Jen Lin. "LIBSVM: a library for support vector machines". En: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pág. 27.

10-(c)) are trained over different parts of global occurrence histograms to reduce the variability in the prediction. The final prediction may be carried out by averaging the predictions of individual trees f_b, or taking the majority vote as expressed $\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b \ \wedge \ \hat{y} = \arg\max\{f_1, \dots f_B\}$. Once the trained models are adjusted from spatio-temporal descriptor, the proposed strategy result efficient in time to perform online prediction over sequences of actions.

4. EVALUATION AND RESULTS

4.1. DATASETS

KTH [12] contains six human action classes: *walking, jogging, running, boxing, hand-waving* and *handclapping*. 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. Each video has a spatial resolution of 160×120 pixels and a frame rate of 25 fps. The proposed approach was evaluated following the original experimental setup which specifies training, validation and test groups as well as five-fold cross validation suggested in [12].

Weizmann [9]**:** this dataset is composed by 90 video sequences that comprehend ten actions such as *walk, run, jump-forward, gallop-sideways, bend, one-hand wave, two-hands wave, jump-in-place, jumping-jack* and *skip*. Validation was carried out using a leave-one-out strategy, taking into account the standard validation reported in the state of the art.

UT-Interaction [31] contains six different human interactions between different people: *shake-hands, point, hug, push, kick* and *punch*. 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 [31].

The proposed method generated a descriptor that was assessed two-fold: employing full video sequences (action classification) in order to obtain a prediction, and using consecutive fragments of a sequence (frame-level action recognition) to obtain a cumulative prediction for online applications.

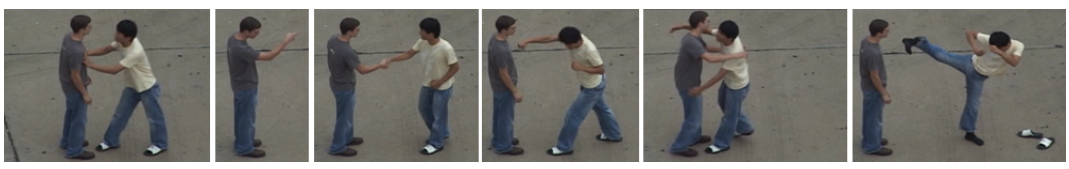
---

[31] M. S. Ryoo y J. K. Aggarwal. *UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)*. http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html. 2010.

Figure 11. KTH dataset: Six subjects performing different actions, namely (from left to right): boxing, handclapping, handwaving, jogging, running and walking.

Figure 12. Weizmann dataset: nine subjects performing actions such as (from left to right, top to bottom): bend, jumping-jack, jump-forward, jump-in-place, run, gallop-sideways, skip, walk, one-hand wave and two-hands wave.

Figure 13. UT-Interaction dataset: shows subjects performing actions such as (from left to right): push, point, shake-hands, punch, hug and kick.



4.2. ACTION CLASSIFICATION

In order to predict human actions from video sequences, the proposed descriptor becomes the input for a classification process. Such procedure imply a previous knowledge regarding the class of each action, which is represented as a label for each element of the descriptor. Our strategy implemented supervised learning algorithms for the classification stage, namely, Decision Tree, Random Forest and Support Vector Machines (linear and RBF kernel). Best overall scores were obtained for both SVM and Random Forest classifiers, as shown in figures 15 and 20. Further experimental results were obtained in order to estimate the best parameters for each circular grid (table 1 and 2), Minimal Number of Trajectories (figures 16 and 17) and optimal number of visual words (figure 14).
Aditional testing is reported in order to evaluate the influence of length of tracking over general classification performance (figures 18 and 19). Similarly, online action recognition was assessed by means of simulating a cumulative video sequence as shown in figures 21,22,23 and 24.

4.2.1. Local spatial counting process approach Occurrence-based descriptor was evaluated in two different tasks: activity classification using complete sequences and partial sequence recognition from frame-level action representations. In both experiments a very compact descriptor achieved a trade-off between accuracy and fast activity prediction, being ideal for real time applications.
A first evaluation of the proposed descriptor was carried out to obtain the best configuration in terms of number of concentric circles (γ) and their respective radius (r), on each evaluated dataset. The number of circles over our circular grid was set as $\gamma = \{4, 5, 6, 7, 8\}$, increasing the size to reach of the neighborhood around the averaged center. Also, the radius of each circle was set as $r = \{6, 8, 10\}$ to admit more

$\gamma \setminus r$	**6**	**8**	**10**
4	88.52	90.26	90.26
5	90.73	89.68	89.33
6	90.61	89.91	88.41
7	88.87	87.83	87.60
8	**91.20**	88.64	88.18

$\gamma \setminus r$	**6**	**8**	**10**
4	76.66	76.66	73.33
5	73.33	72.22	**78.88**
6	74.44	72.22	72.22
7	75.55	73.33	71.11
8	78.88	71.11	71.11

Table 1. Preliminary experiment for KTH (top) and Weizmann (bottom) datasets: accuracy (%) for combinations of γ circles and radius r (px).

trajectories inside each region. These results were obtained w.r.t a global descriptor of 400 words and using the SVM with a RBF kernel, as strategy for classification.

In table 1-top and 1-bottom is reported the performance of the proposed approach using different (γ, r) configurations on KTH and Weizmann datasets, respectively. For KTH the maximum score of $91.2\,\%$ was achieved with the tuple ($\gamma = 8$, r=6), standing out small radius steps because of the spatial frame resolution in this dataset. Hence a larger number of γ circles was necessary to codify actions in terms of trajectory densities around each point, mainly because the dynamic closeness of some activities, such as: jogging and running. As reported in table 1, the proposed approach achieves more than $87\,\%$ of accuracy for almost all configurations, demonstrating its robustness to represent actions given the different (γ, r) configurations.
Regarding Weizmann dataset, the best score ($78.88\,\%$) was achieved with the tuple ($\gamma = 5$, r=10). This dataset groups most action classes and therefore the classification is more challenging, w.r.t to KTH dataset. Larger radius sizes have better per-

formance, since the spatial frame resolution is much more larger than for KTH. Also, this frame resolution allows a better density trajectory description around each point, and therefore fewer γ circles are sufficient to describe the actions. As expected, for Weizmann dataset are also reported stable results for different (γ, r) configurations, obtaining more than $70\,\%$ of accuracy to describe the 10 different actions.

The proposed approach was also evaluated on UT-Interaction dataset that represent activities in almost real conditions. This dataset is split on: set 1, which was captured with a relatively static camera, and set 2, where videos were recorded with some camera jitters. In tables 2-top and 2-bottom are reported the performance of the proposed approach for set 1 and set 2, respectively. As expected, the best accuracy ($71.66\,\%$) is obtained on set 1 with a tuple configuration of ($\gamma = 8$, r= 8), while for set 2 with tuple ($\gamma = 4$, r=6) was achieved a $65\,\%$. The set 2 of UT-Interaction has better performance on more reduced occurrence space, since most of the trajectories could be related with camera motions, introducing artifacts to the analysis of activities. In contrast, the set 1 achieves a better performance doing a wide analysis from several circles with a radius of 8 px. Interestingly enough and despite of the challenge of the UT-Interaction dataset, almost whole (γ, r) configuration achieves similar scores showing a stable performance of the proposed descriptor.

In a second experiment, the dictionary size, that code main words for all actions on training video sequences, was adjusted for different k centroids, which allowed to find an adequate balance between dimensionality of descriptor and accuracy. Using a SVM with a RBF kernel, the size of the dictionary was increased from histograms with sizes of 100 to 1600. For all evaluated datasets the best overall accuracy was obtained for $k = 400$ visual words, as shown in figure 14.

It is worth noting that much of the state-of-the art strategies achieve stable results for descriptors with thousands of values, while the proposed approach achieve the best

$\gamma \setminus r$	**6**	**8**	**10**
4	65	65	66.66
5	65	68.33	70
6	65	70	68.33
7	66.66	70	70
8	68.33	**71.66**	66.66

$\gamma \setminus r$	**6**	**8**	**10**
4	**65**	56.66	61.66
5	60	53.33	60
6	56.66	65	56.66
7	61.66	58.33	53.33
8	61.66	60	53.33

Table 2. Preliminary experiment for UT-Interaction dataset (set1 - top, set2 - bottom): accuracy (%) for combinations of γ circles and radius r (px).

Figure 14. Accuracy for different k-values given best configurations over KTH, Weizmann and UT-Interaction datasets.

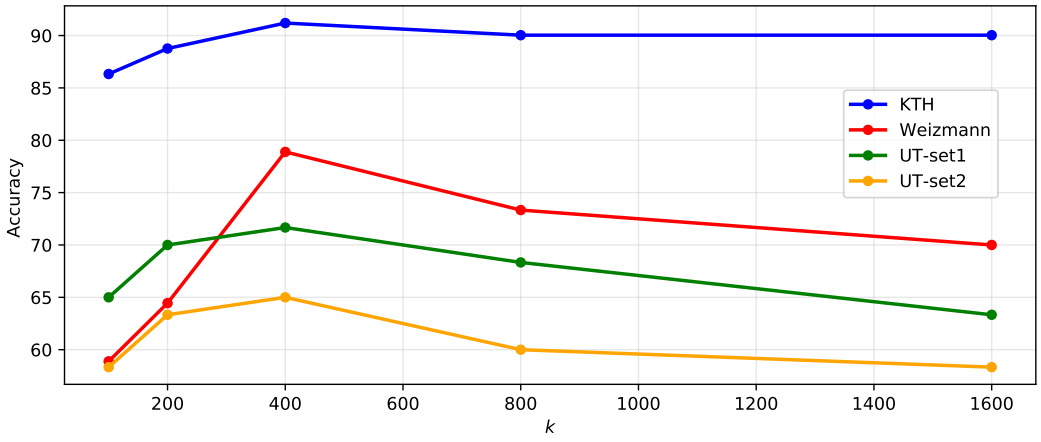
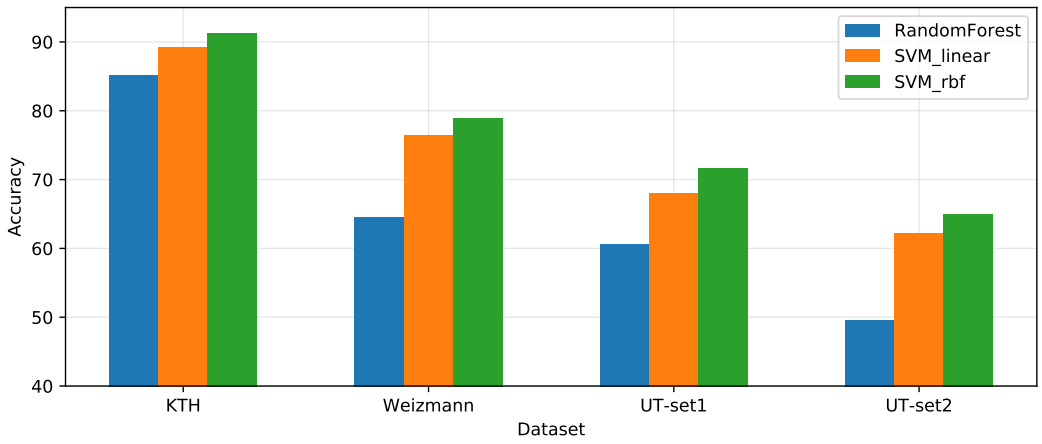


Figure 15. Classification strategies for KTH, Weizmann and UT-Interaction datasets. Best overall accuracies are shown.



performance by only using 400 scalar values to represent a complete video sequence. Such compact descriptor result ideal as a first stage of identification, allowing also to highlight the importance of active spatial points to represent actions developed by the subject. Additionally, the descriptor shows a good trade-off between dimension size and accuracy.

In a third experiment, several classification strategies were evaluated over the best (γ, r) configurations obtained in previous experiments (figure 15). The classification strategies herein selected allow to deal with non linear spaces, with appropriate balance regarding computational time. The classification strategies were: Random Forest (RaF), SVM (with linear kernel) and SVM (with RBF kernel). The whole strategies were evaluated under a grid search parameter framework to obtain best configurations w.r.t to the proposed action descriptor and the respective dataset evaluated. In three evaluated datasets the SVM with RBF kernel achieved the best performance in terms of accuracy. Such fact could be associated to the properties of kernel to separate actions through radial basis functions w.r.t support vectors of reference rather than linear boundaries. It should be also noted that linear kernel respond to a proper performance of accuracy but with the main advantage of low computational comple-

xity, which could be a perfect tool to obtain fast prediction on real-time applications. Despite the lower performance of random forest classifier, the obtained results on KTH and Weizmann dataset are competitive and useful on low-level architectures. Also this classification strategy is interesting as a pre-processing step to discover most relevant features on the descriptor. Finally, prediction of human actions constitute a task that often requires reduced execution times. Our descriptor takes only 0.036 seconds in average to correctly recognize human actions on standard video sequences.

4.2.2. Local occurrence binary pattern approach Firstly, we evaluated the proposed approach to obtain an optimal value for the Minimal Number of Trajectories (MNT) threshold τ. Our circular grid was preconfigured with $\alpha = 9$ angles and MNT was set as $\tau = \{2, 3, 4, 5, 6, 7, 8\}$, increasing the amount of trajectories n_i admitted inside each region r_i around an averaged center, over all datasets.

In figure 16 is reported the performance of the proposed approach using different MNT (τ) values over UT-Interaction dataset. For the UT sequence 1 the maximum score of $75\,\%$ was achieved with $(\gamma = 8, r = 8), \tau = 6$, as for sequence 2, the maximum score of $70\,\%$ was obtained with $(\gamma = 4, r = 6), \tau = 2$. Sequence 2, which shows camera jitters and human interactions, require a smaller density of trajectories and a smaller reach ($\gamma = 4$) w.r.t. sequence 1 ($\gamma = 8$). Also, these features are explained because of the dynamic closeness of activities such as punching and pushing in sequence 1, and on the opposite, due to the uncontrolled scenarios of sequence 2.

On the other hand, results for KTH are shown in figure 17 featuring values of MNT from 2 to 8, achieving an accuracy of 90.03 % for $(\gamma = 8, r = 6)$ and $\tau = 6$. This is explained by the periodic featured actions and the constant spatial cropping of the sequences. Regarding Weizmann dataset, the best score ($78.88\,\%$) was achieved with $\tau = 6$ and $(\gamma = 5, r = 10)$. This dataset groups most action classes and therefore

Figure 16. Reported accuracy for different values of MNT occurring on each cell of the proposed circular grid over UT-Interaction dataset.

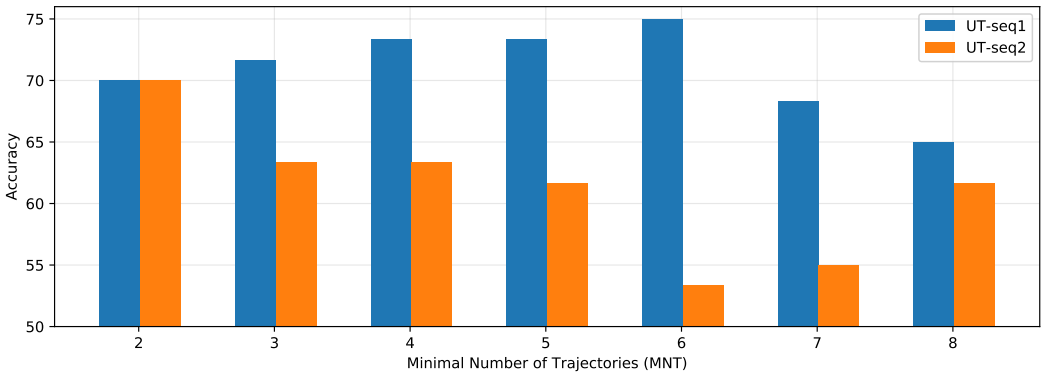
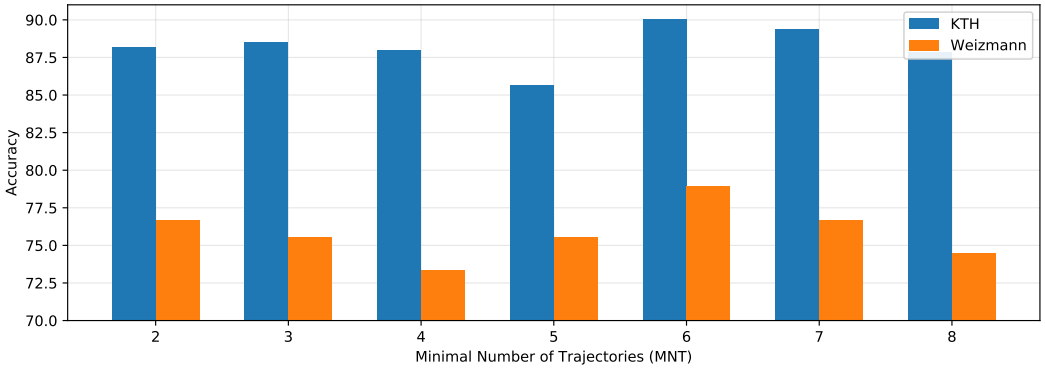


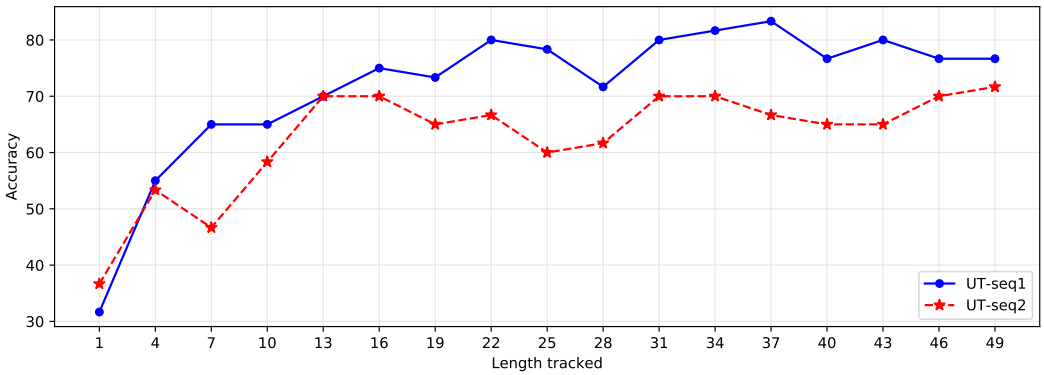
Figure 17. Reported accuracy for different values of MNT occurring on each cell of the proposed circular grid over KTH and Weizmann datasets.



the classification is more challenging. Larger radius sizes have better performance, since the spatial frame resolution is much more larger than for KTH. In general, the ToBPs coding shows stable results for different configurations, with a mode above of $70\,\%$ for very different and challenging datasets.

Secondly, we analyze the impact of trajectories length to describe actions. The experimental setup involved to increase the number of frames for trajectory tracking from 1 to 49 frames and a recognition using a SVM with a RBF kernel. Figure 18 shows a crescent shape with a local maximum on $= 37$ for UT-Interaction dataset, but resulted unstable for neighboring configurations (figure 23 green line). Best per-

Figure 18. Reported accuracy for different lengths of tracking over UT-Interaction dataset.



formance is then found at $= 22$ because a favorable trade-off between descriptor size ($\alpha \times \gamma \times 22$) and online action recognition behaviour (figure 23 blue line). Sequence 2 obtained a local maxima on $= 49$ but it implied a significant increase in descriptor size ($\alpha \times \gamma \times 49$) for obtaining only an extra 1 %. Thus, the default length of $= 15$ yields a sufficient performance, as shown in figure (figure 23 red line).

As for Weizmann dataset, we used the same experimental configuration and obtained local maxima on $= 15$ and $= 37$ as shown in figure 19. Therefore $= 15$ yields a favorable trade-off with a descriptor of only ($\alpha \times \gamma \times 15$). Regarding KTH dataset, best performance was achieved with only $l = 15$ which yields a compact descriptor size of ($\alpha \times \gamma \times 15$).

In a third experiment, three classification strategies were evaluated over the best MNT configurations of ToBPs. These classification strategies adequately dealt with non linear spaces, and shown a reasonable trade-off in terms of computational time. Namely, we used: Random Forest (RaF), SVM (with linear kernel) and SVM (with RBF kernel). These methods were evaluated under a grid search parameter setup to obtain optimal configurations w.r.t to the proposed action descriptor.

As shown in figure 20, the SVM + RBF kernel obtained the best performance in terms of accuracy and low computational complexity which is adequate for real-time

Figure 19. Reported accuracy for different lengths of tracking over Weizmann dataset.

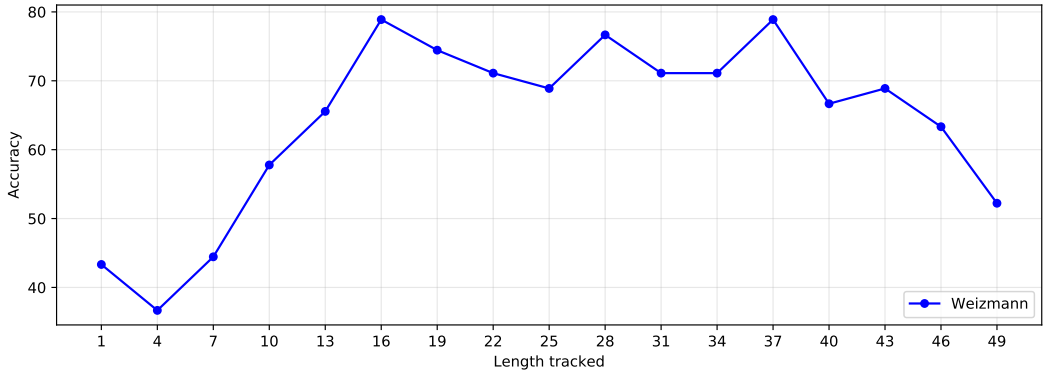
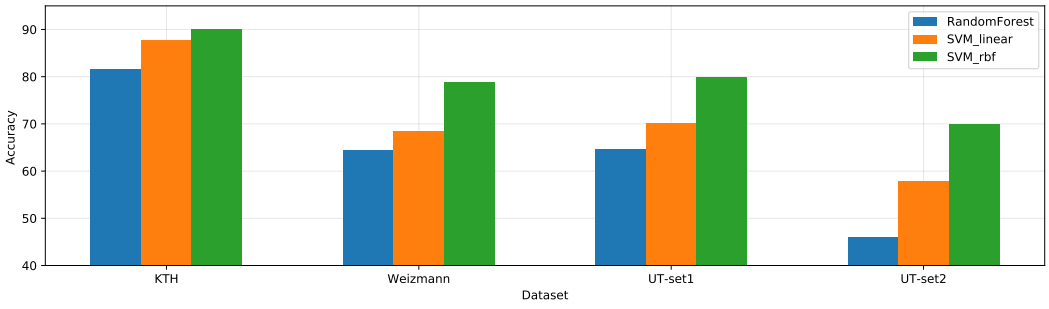


Figure 20. Classification strategies for KTH, Weizmann and UT-Interaction datasets. Best overall scores are shown.



classification tasks. We emphasize that the proposed descriptor achieves a favorable performance with only 400 scalar values for full video sequences, compared with state-of-the-art strategies that in general require thousands of scalar values. Moreover, our descriptor obtains a balanced ratio between accuracy and dimensionality, which makes it adequate for online classification tasks.

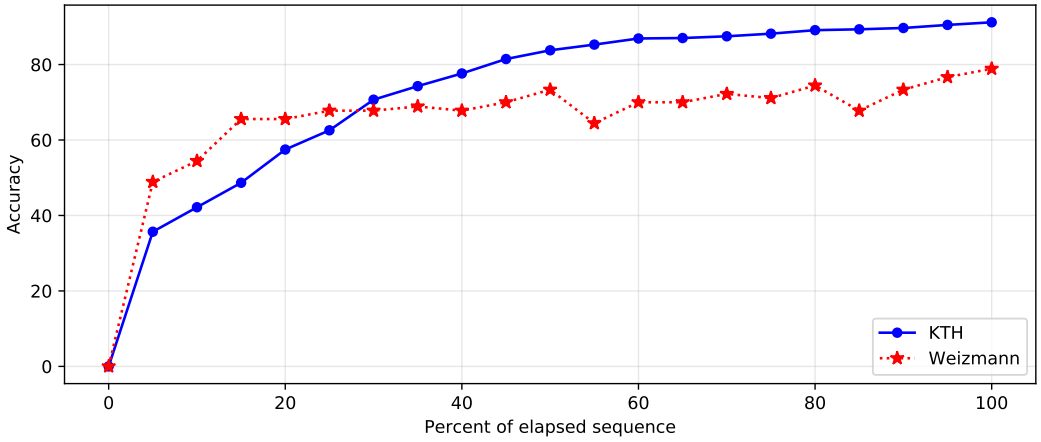
4.3. FRAME-LEVEL ACTION RECOGNITION

4.3.1. Local spatial counting process approach A second evaluation of the proposed approach was carried out on partial sequences computed at frame level representations. Such experiments deal with the capability of the strategy regarding online and partial recognition of action, setting face to incomplete information of the dynamic. To obtain a statistical quantification of partial recognition performance, the proposed approach was tested using different temporal percentages of the video sequences, and an averaged result was registered for all datasets.

In figure 21, blue line illustrates the recognition performance of the proposed approach for the KTH dataset. A proper action prediction is achieved by using almost half of the video sequences, reporting more than 70 % of accuracy. Also, a natural increasing of prediction is obtained while more frames are used into the prediction. Such fact is associated to the periodic nature of recorded actions.

Additionally, the red line in figure 21 represents the recognition achieved by the proposed approach for Weizmann dataset. A fast recognition is herein achieved by using only the 15 % of video sequences. After that, the proposed approach remains relatively stable, showing some mild increasing at the end of the sequences. Some little accuracy variations could be associated to action duration of different activities, and the relative non-regular cropping of video sequences.

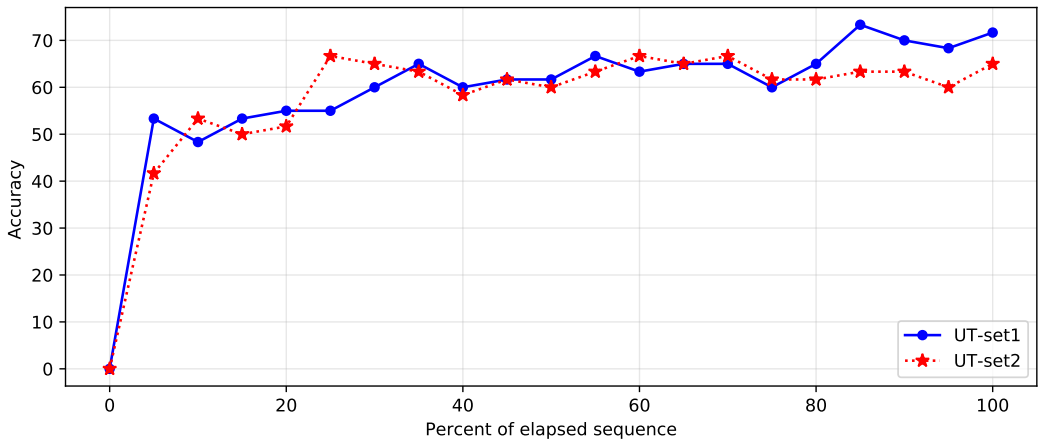
Figure 21. (Blue line) Performance simulation for an online application over KTH dataset. Our method achieves promising results with just $25\,\%$ of the total number of frames. (Red line) Performance simulation for an online application over Weizmann dataset. Our method achieves promising results with just $15\,\%$ of the total number of frames.



Finally, in figure 22 is illustrated the online recognition performance for UT-interaction in both set of videos. Despite of difference of both sets, w.r.t, to the camera motion and other actions on the background of the scenes, the proposed approach achieves a very similar performance on the task of recognition. Interestingly enough, the proposed approach achieves the best score (with $\sim 67\,\%$) on set 2 on an intermediate representation, by using only the 25 % of the video sequences. Also for set 1, the best score ($\sim 73\,\%$) is achieved by using the 85 % of the video sequence. These partial results are even better that the obtained results on classification evaluation, previously described. For both cases, the prediction, as in other datasets, remain relatively stable for almost all entire sequences.

4.3.2. Local occurrence binary pattern approach Online action recognition requires an updated prediction for each frame t_i of the video sequence and then a evaluation was executed over partial sequences on a frame level representation. This strategy aims to verify the performance of the proposed descriptor when the

Figure 22. Performance simulation for an online application over UT-Interaction dataset.

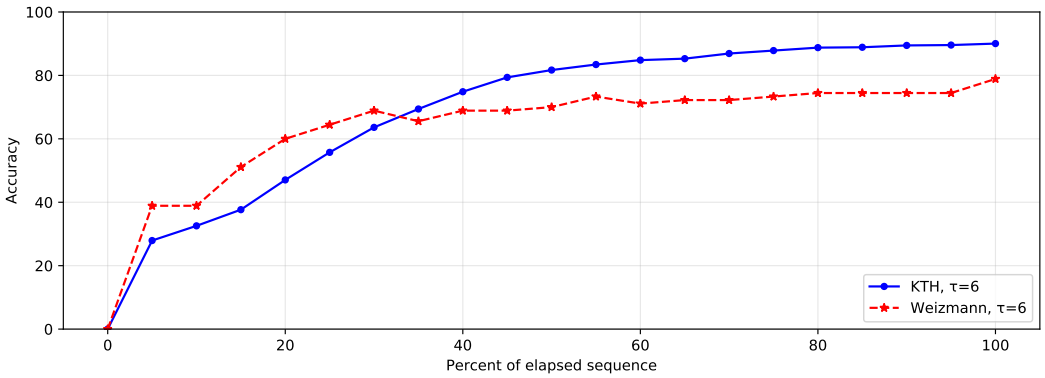


dynamics of motion are incomplete. In order to calculate statistics for partial recognition, our method was proved using distinct temporal percentages of the sequences, and for all the datasets an averaged score was reported.

Online recognition performance for KTH and Weizmann datasets is shown in figure 23. A competitive prediction for KTH (blue) is obtained with approximately 40 % of the sequence for an accuracy of more than 74 %. The periodic behaviour of the actions explain the sustained increase in the prediction score when more frames are employed for the prediction task. On the other hand, a red dashed line illustrates online recognition accuracy for Weizmann dataset, where a promising score is acquired with only 20 % of the total number of frames. Also, after 40 % of the sequence the score shows stability and a moderate increase at the end of the sequence. Existent accuracy fluctuations are associated with a distinct duration between actions of the same class and the different scale cropping of the sequences. It is worth noting that a quick prediction is obtained with just 10 % of the sequences.

Lastly, predictive behaviour for UT-Interaction is illustrated in figure 24 including set 1 (blue) and set 2 (red). Set 1 obtains promising results with just 15 % of the elapsed sequences and sustained an increasing score with small variations. As for set 2,

Figure 23. Performance simulation for an online application over KTH dataset (blue): our method achieves promising results with just $30\,\%$ of the total number of frames. (Red) Weizmann dataset: promising results are obtained with just $20\,\%$ of the total number of frames.



a reduced performance is explained by the noise introduced by additional featured subjects and background motion, in addition to changes in the spatial cropping of the sequences. Particularly, reported results correspond to a length of tracking $l = 22$ for set 1, $l = 15$ for set 2 and an extra test with $l = 37$ (green dotted) for set 1, which shown a final increase in prediction score for the entire sequences, but an inadequate performance for online recognition.

4.4. EXECUTION TIME AND PERFORMANCE

An extra set of experiments aimed to measure execution times of: dense and improved motion trajectories calculation, local processing and characterization of trajectories, both spatial counting process and occurrence binary patterns descriptors, dictionary calculation and action classification using a SVM with a RBF kernel.
In order to obtain an estimation of the temporal performance of the proposed method, timing on each block of the process was measured as shown in figure 25. Besides, online and full sequence action recognition steps are highlighted for a better understanding of the proposed approach.

Figure 24. Performance simulation for an online application over UT-Interaction dataset. Promising results are obtained with just $15\,\%$ of the total number of frames for sequence 1 (blue).

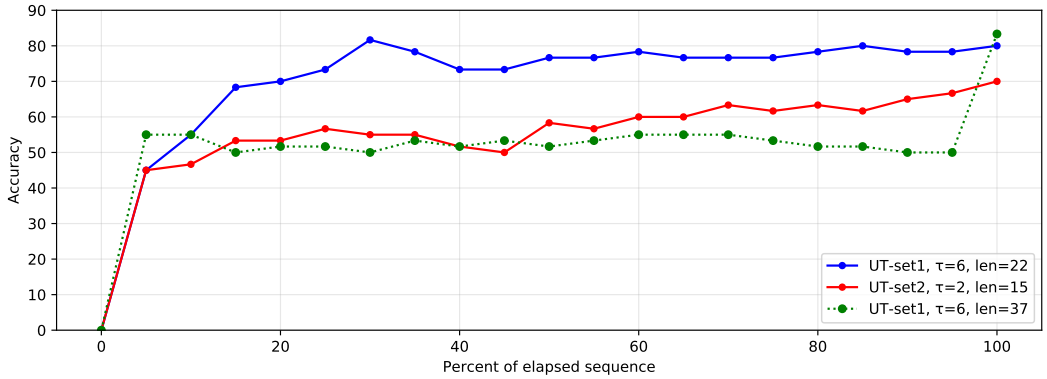
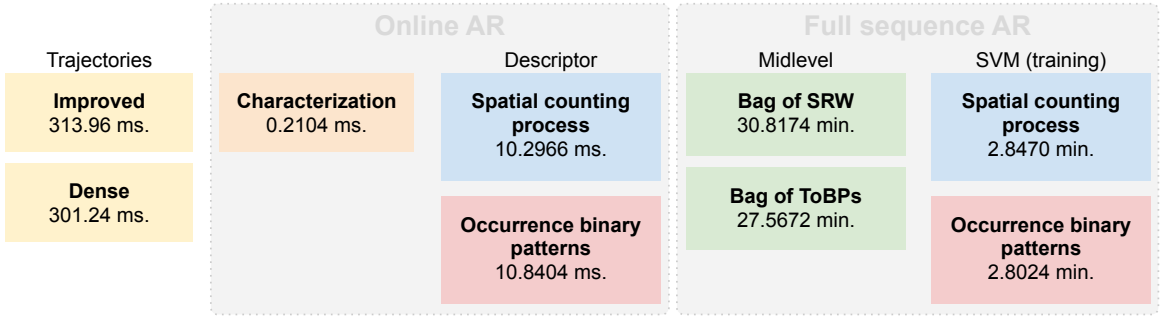


Figure 25. Averaged per-frame execution times for each stage of the proposed method over KTH dataset. Gray bounded regions highlight online action recognition (characterization and descriptor) and full sequence action recognition (midlevel and classification).



A competitive result is obtained on execution time for midlevel stage (≈ 30 min.) considering that KTH dataset is composed of 2391 video sequences. Same reasoning applies for classification process during training and testing steps. Overall execution time during online AR steps also featured a fast performance, considering our descriptor size of 400 scalar values at midlevel representation.

Furthermore, it is worth noting that improved trajectories are slower to compute compared with dense trajectories. This is due to the extra load from background motion filtering and other upgrades that reduce the number of reported trajectories. Also, a slight increase is evident for the occurrence binary patterns strategy with respect to the spatial counting process approach. Extra adjustments, filtering and thresholding of occurrences inside subregions explain this behaviour. As for the Bag of ToBPs, such advantage w.r.t. BoSRW is explained because the sparse structure of its descriptor when occurrence values were not compliant with the fixed MNT threshold. Action classification stage showed a similar elapsed time for both strategies, which makes sense, based on the equal dimension size of both descriptors.

5. DISCUSSION

The proposed approach introduced a reduced size descriptor that recognize actions through a local and global occurrence analysis of active trajectories. Such recognition exploits spatio-temporal information that lies within neighboring trajectories. A proximity criteria defines which trajectories will belong inside a neighborhood that is bounded by concentric circles and angles.

In order to characterize motion, trajectories existing around a fixed center were assigned to certain bounded subregions over a concentric circular grid. A spatial point process is then implemented and occurrences of neighboring trajectories on each subregion are stored. An additional alternative were to perform a threshold of a minimal number of trajectories expected to fall on each subregion, taking a value of 1 if that condition is true or 0 otherwise, leading to a sparse descriptor. The first alternative generated a more detailed descriptor and thus, produced a slightly better accuracy. Even so, it is an interesting fact that the second alternative obtained only 1.2 % less, taking into account the simplicity of its content.

Experimental results evidence a competitive performance on classification and recognition task under academic and well known action datasets. A main advantage of the proposed approach is the compact frame-level representation of the action, which result useful on low level computational architectures and for applications that require online predictions on video streamings.

In literature, much of the proposed action recognition strategies are dedicated to classify action on well cropped sequences, which also namely require high dimensional descriptors to represent actions. A seminal word coding action representation was proposed by Laptev et al. [2], that codes actions from occurrences of appearance patches computed along the video. This approach requires video descriptors with size of 4000 scalar values, and fails in much of the cases because of the dependency

of appearance of the actions. Currently, as an alternative to appearance based descriptors, the work proposed by Wang et al. [16] compute cuboids descriptors around motion trajectories along video sequences. Such cuboids codes gradient appearance and motion information of the actions. Significant results are achieved using this strategy, over challenging dataset but requiring high dimensional descriptors to represent the activities. In such case, the best configuration is achieved with cuboids of size $32 \times 32 \times 16$, and global occurrence histograms of size 4000. In contrast, the proposed approach operates at two level of occurrences, obtaining very compact descriptors to represent actions. In terms of occurrences around trajectories, the size of descriptor is $9 \times 8 \times 16$, while the global occurrence is achieved with compact histograms of 400 scalar values. Additionally, the strategy is able to recover a global representation from partial representations.

Currently, sophisticated deep learning strategies are proposed in the state-of-the-art to deal with action challenges, such as camera motion, scene representation and even w.r.t to the dynamic representation of the actions [17][18]. Such approaches nevertheless require a huge number of samples to achieve coherent results, and also require complex setups of training to achieve proper accuracy results. Because much of these approaches are based on convolutional representations, a fixed interval of the action is required. Other approaches based on LSTM architectures have been proposed to deal with temporal and dynamic correlation of actions. For instance, Veeriah et al. [32] introduced a LSTM-based proposal that uses derivatives of gait states as a threshold criteria for a temporal segmentation of actions. This approach achieves a 92.12 % accuracy over KTH dataset, using 450 scalar values obtained with PCA from a 56.000 sized feature vector. Regarding the proposed approach (ac-

---

[32] Vivek Veeriah, Naifan Zhuang y Guo-Jun Qi. “Differential recurrent neural networks for action recognition”. En: *Proceedings of the IEEE international conference on computer vision*. 2015, págs. 4041-4049.

curacy of 91.2 %), in comparison, there is not statistical significance on the obtained results over the KTH dataset. Also, the LSTM-based approach achieve a compact description but requiring an additional processing of dimensionality reduction from the PCA.

The proposed method can be adapted to different classifying strategies and is not dependent on high performance hardware specifications, which facilitates online action recognition and a low-cost implementation. Taking into account the current reported results, our strategy can also be adapted to recurrent neural networks architectures in order to exploit the richness of sequential information regarding human action. Moreover, our method can be used as an input for finding correlation over different motion trajectories and the joint variability of motion kinematics in order to recognize actions.

6. CONCLUSIONS

The present work proposed a computational strategy that uses representations of spatio-temporal primitives in order to classify and recognize human actions. Firstly, motion trajectories were calculated from video sequences of different durations, featuring a fixed length of tracking for each path. Secondly, these paths were characterized using spatio-temporal primitives that rely exclusively on kinematic features. Thirdly, said information was codified inside a spatio-temporal descriptor that features two local approaches: spatial counting processes and occurrence binary patterns.

Moreover, a midlevel representation managed to create a dictionary of salient words by implementing a k-means to compute a fixed number of centroids. This codification is build for entire video sequences and partial sequences as the global space were the dictionary is used. The resulting data became the input for an action classification and recognition stage.

Experimental results comprised action classification on full sequences and frame-level action recognition emulating a partial sequence. Accuracy scores reported a competitive performance with a compact descriptor that describe actions with at most 72 local features and only 400 visual words. Such efficient features make this computational strategy adequate for environments with restricted hardware resources and might be adapted for performing action recognition over embedded devices.

On the other hand, the proposed strategy result inadequate when a video sequence contains few or no trajectories. Also, frames with size greater than 300px add a noticeable delay in the execution time. Additionally, if using the default $l = 16$ length of tracking, sequences with an inferior number of frames will lack trajectories and thus, accuracy is negatively affected.

Further research is needed in order to improve the adequacy of current motion pri-

mitives. Additional testing will be carried out over different uncontrolled datasets, with an increased amount of sequences. Also, an implementation that exploits the advantages of deep learning architectures is planned, particularly recurrent neural networks. Higher order statistics are also being considered in order to strengthen our motion representation.

BIBLIOGRAPHY

Al-Akam, Rawya, Salah Al-Darraji y Dietrich Paulus. “Human Action Recognition from RGBD Videos based on Retina Model and Local Binary Pattern Features. In 26”. En: *Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*. 2018, págs. 1-7 (vid. pág. 24).

Al-Ali, Salim y col. “Human action recognition: contour-based and Silhouette-based approaches”. En: *Computer Vision in Control Systems-2*. Springer, 2015, págs. 11-47 (vid. pág. 19).

Baccouche, Moez y col. “Sequential deep learning for human action recognition”. En: *International Workshop on Human Behavior Understanding*. Springer. 2011, págs. 29-39 (vid. págs. 21, 58).

Bobick, Aaron F. y James W. Davis. “The recognition of human movement using temporal templates”. En: *IEEE Transactions on pattern analysis and machine intelligence* 23.3 (2001), págs. 257-267 (vid. pág. 17).

Bouwmans, Thierry y col. “On the role and the importance of features for background modeling and foreground detection”. En: *Computer Science Review* 28 (2018), págs. 26-91 (vid. pág. 22).

Chang, Chih-Chung y Chih-Jen Lin. “LIBSVM: a library for support vector machines”. En: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pág. 27 (vid. pág. 38).

Garzon, Gustavo y Fabio Martinez. “A fast action recognition strategy based on motion trajectory occurrences”. En: *Pattern Recognition and Image Analysis* 29 (2019) (vid. pág. 27).

— “Online action recognition from trajectory occurrence binary patterns (ToBPs)”. En: *Proceedings of the International Conference on Advances in Emerging Trends and Technologies*. 2019 (vid. pág. 27).

Gorelick, Lena y col. “Actions as space-time shapes”. En: *IEEE transactions on pattern analysis and machine intelligence* 29.12 (2007), págs. 2247-2253 (vid. págs. 18, 19, 40).

Johansson, Gunnar. “Visual perception of biological motion and a model for its analysis”. En: *Perception & psychophysics* 14.2 (1973), págs. 201-211 (vid. pág. 27).

Junejo, Imran N, Khurrum Nazir Junejo y Zaher Al Aghbari. “Silhouette-based human action recognition using SAX-Shapes”. En: *The Visual Computer* 30.3 (2014), págs. 259-269 (vid. pág. 19).

Laptev, Ivan. “On space-time interest points”. En: *International journal of computer vision* 64.2-3 (2005), págs. 107-123 (vid. pág. 20).

Laptev, Ivan y Tony Lindeberg. “Local descriptors for spatio-temporal recognition”. En: *Lecture notes in computer science* 3667 (2006), págs. 91-103 (vid. pág. 20).

Laptev, Ivan y col. “Learning realistic human actions from movies”. En: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, págs. 1-8 (vid. págs. 16, 57).

Nanni, Loris, Sheryl Brahnam y Alessandra Lumini. “Local ternary patterns from three orthogonal planes for human action classification”. En: *Expert Systems with Applications* 38.5 (2011), págs. 5125-5128 (vid. pág. 23).

Nguyen, Thanh Phuong y col. “Revisiting lbp-based texture models for human action recognition”. En: *Iberoamerican Congress on Pattern Recognition*. Springer. 2013, págs. 286-293 (vid. pág. 24).

Ojala, Timo, Matti Pietikäinen y David Harwood. “A comparative study of texture measures with classification based on featured distributions”. En: *Pattern recognition* 29.1 (1996), págs. 51-59 (vid. pág. 22).

Ojala, Timo, Matti Pietikainen y Topi Maenpaa. “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”. En: *IEEE Transactions on pattern analysis and machine intelligence* 24.7 (2002), págs. 971-987 (vid. pág. 22).

Poppe, Ronald. “A survey on vision-based human action recognition”. En: *Image and vision computing* 28.6 (2010), págs. 976-990 (vid. pág. 16).

Rahmani, Hossein, Ajmal Mian y Mubarak Shah. “Learning a deep model for human action recognition from novel viewpoints”. En: *IEEE transactions on pattern analysis and machine intelligence* 40.3 (2018), págs. 667-681 (vid. págs. 21, 58).

Rodriguez, Mikel D, Javed Ahmed y Mubarak Shah. “Action mach a spatio-temporal maximum average correlation height filter for action recognition”. En: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, págs. 1-8 (vid. pág. 16).

Ryoo, M. S. y J. K. Aggarwal. *UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)*. http://cvrc.ece.utexas.edu/SDHA2010/Human_Inter 2010 (vid. pág. 40).

Schuldt, Christian, Ivan Laptev y Barbara Caputo. “Recognizing human actions: a local SVM approach”. En: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. IEEE. 2004, págs. 32-36 (vid. págs. 20, 40).

Subetha, T y S Chitrakala. “A Survey on human activity recognition from videos”. En: *Information Communication and Embedded Systems (ICICES), 2016 International Conference on*. IEEE. 2016, págs. 1-7 (vid. pág. 16).

Takahashi, Masaki y col. “Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories”. En: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE. 2011, págs. 9-16 (vid. pág. 16).

Veeriah, Vivek, Naifan Zhuang y Guo-Jun Qi. “Differential recurrent neural networks for action recognition”. En: *Proceedings of the IEEE international conference on computer vision*. 2015, págs. 4041-4049 (vid. pág. 58).

Wang, Heng y Cordelia Schmid. “Action recognition with improved trajectories”. En: *Proceedings of the IEEE international conference on computer vision*. 2013, págs. 3551-3558 (vid. págs. 21, 58).

Wang, Heng y col. “Dense trajectories and motion boundary descriptors for action recognition”. En: *International journal of computer vision* 103.1 (2013), págs. 60-79 (vid. págs. 21, 28).

Wu, Ying y Thomas S Huang. “Vision-based gesture recognition: A review”. En: *Gesture Workshop*. Vol. 1739. Springer. 1999, págs. 103-115 (vid. pág. 16).

Yang, Yun y col. “Action recognition using completed local binary patterns and multiple-class boosting classifier”. En: *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*. IEEE. 2015, págs. 336-340 (vid. pág. 24).

Yeffet, Lahav y Lior Wolf. “Local trinary patterns for human action recognition”. En: *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE. 2009, págs. 492-497 (vid. pág. 23).

Zhu, Guangming y col. “An online continuous human action recognition algorithm based on the kinect sensor”. En: *Sensors* 16.2 (2016), pág. 161 (vid. pág. 16).

APPENDICES

Anexo A. Academic Products

Journals

- G. Garzón, F. Martínez. “A fast action recognition strategy based on motion trajectory occurrences”. Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. Springer.
  Status: accepted, planned for publishing: No. 3, 2019.

Conference papers

- G. Garzón, F. Martínez. “Online action recognition from trajectory occurrence binary patterns (ToBPs)”. International Conference on Advances in Emerging Trends and Technologies ICAETT 2019. Springer.
  Status: accepted, planned for publishing: November, 2019.

- W. Moreno, G. Garzón, F. Martínez. “Frame-Level Covariance Descriptor for Action Recognition”. Colombian Conference on Computing. 2018. DOI: 10.1007/978-3-319-98998-3_22.

Other collaborations

- Y. Gutiérrez, G. Garzón, F. Martínez. “Towards clinical significance prediction using K-trans evidences in prostate cancer”. XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA 2019). To be published on: IEEE Xplore Digital Library. DOI: 10.1109/STSIVA.2019.8730282