# TRAYECTORIAS DE MOVIMIENTO DENSAS 3D+T COMO PRIMITIVAS CINEMÁTICAS PARA EL ANÁLISIS DE SECUENCIAS DE VIDEO DE PROFUNDIDAD

**Fabián Castillo, Lola Bautista, Fabio Martínez**
Biomedical Imaging, Vision and Learning Laboratory (BivL$^2$ ab), Universidad Industrial de Santander, Bucaramanga, Colombia.

Corresponding author: famarcar@uis.edu.co

## RESUMEN

Los sensores RGB-D han permitido atacar de forma novedosa muchos de los problemas clásicos en visión por computador, tales como la segmentación, la representación de escenas, la interacción humano-computador, entre otros. Con respecto a la caracterización de movimiento, las estrategias típicas en RGB-D están limitadas al análisis dinámico de formas globales y a la captura de flujos de escena. Estas estrategias, sin embargo, solo recuperan información dinámica entre cuadros consecutivos, limitando el análisis de largos desplazamientos. Este trabajo presenta una estrategia para el cálculo de trayectorias (3D+t), las cuales son fundamentales para la descripción cinemática local, permitiendo una descripción densa de movimiento. Cada trayectoria permite modelar palabras cinemáticas, las cuales en conjunto, describen gestos complejos en los videos. Estas palabras cinemáticas fueron procesadas dentro de un esquema de bolsa-de-palabras para obtener un descriptor basado ocurrencias. Este descriptor de trayectorias logró una exactitud del 80% en 5 gestos y 100 videos.

**Palabras clave:** RGB-D; Flujo de escena; Trayectorias densas de movimiento; Seguimiento; Características cinemáticas

Recibido: 21 de febrero de 2019. Aceptado: 17 de Mayo de 2019
*Received: February 21, 2019. Accepted: May 17, 2019*

## 3D+T DENSE MOTION TRAJECTORIES AS KINEMATICS PRIMITIVES TO RECOGNIZE GESTURES ON DEPTH VIDEO SEQUENCES

## ABSTRACT

*RGB-D sensors have allowed attacking many classical problems in computer vision such as segmentation, scene representations and human interaction, among many others. Regarding motion characterization, typical RGB-D strategies are limited to namely analyze global shape changes and capture scene flow fields to describe local motions in depth sequences. Nevertheless, such strategies only recover motion information among a couple of frames, limiting the analysis of coherent large displacements along time. This work presents a novel strategy to compute 3D+t dense and long motion trajectories as fundamental kinematic primitives to represent video sequences. Each motion trajectory models kinematic words primitives that together can describe complex gestures developed along videos. Such kinematic words were processed into a bag-of-kinematic-words framework to obtain an occurrence video descriptor. The novel video descriptor based on 3D+t motion trajectories achieved an average accuracy of 80% in a dataset of 5 gestures and 100 videos.*

**Keywords:** *RGB-D, scene flows, dense motion trajectories, tracking, kinematic features.*

*Cómo citar este artículo: F. Castillo, L. Bautista, F. Martinez. "Trayectorias de movimiento densas 3d+t como primitivas cinemáticas para el análisis de secuencias de video de profundidad", Revista Politécnica, vol. 15, no.29 pp.82-94, 2019. DOI: 10.33571/rpolitec.v15n29a7*

## 1. INTRODUCTION

Typically, computer vision applications are merely based on optical RGB representations, with dependency on appearance information, which implies several limitations, such as: segmentation of similar color objects, detection in dynamic scenarios, sensibility to illumination changes, or even strong variability recognition for different perspectives. The current RGB-D (Kinect) devices have allowed to introduce new information to better represent objects of interest from depth scenes. This new multimodal analysis has allowed addressing new perspectives for classical vision problems, such as: 3D reconstruction, human tracking, human interaction, among many others. Also this kind of analysis helps with typical problems of illumination changes and perspective. The computation of RGB-D primitives results useful to understand complex scenarios by capturing and characterizing more accurately the different objects of interest in a particular problem. Nevertheless, the depth sensors are limited in resolution and some little environment perturbations lead to noisy measurements. Additionally, the use of depth information is complex because of the nonlinear correlation between depth and optical information, for instance scale differences and strong dependency on intrinsic video device parameters. Even worst, the computation of motion primitives implies the association of temporal RGB frames along the sequences, which increase the complexity of computational approaches.

In the state-of-the-art has been proposed several strategies to recover low level features and to develop image descriptors from RGB-D information. For instance, Blum et. al. [1] introduces a RGB-D clustering algorithm that recovers regional patterns from SIFT points. Such approach is however dependent of features in cluster, and the SIFT points discard depth information. In [2] it was analyzed the representation of scenes by computing interest points in appearance and depth information, independently. Interestingly, in that work it is reported that a better characterization for recognition is achieved only using appearance features. Such evaluation can be justified because richness in RBG space while depth information has coarse levels of representation. Nevertheless, in the reported work authors suggest the use of appearance points but complemented with depth information in a low level scale by using average measures around points. In this sense, SIFT and HOG descriptors were extended in RGB-D sequences, achieving an improvement of 10% in a classical object recognition task [3].

Regarding motion analysis, seminal works have analyzed temporal changes of silhouettes estimated from depth channels [4]. These approaches remove dependencies of appearance and result efficient in time computation. However, a main problem in such global representation is the dependency of perspective and the restrictions w.r.t occlusion. Also, a typical motion characterization is carried out from the computation of the appearance's velocity field between consecutive frames to recover the displacement of an object of interest characterized from its appearance. From a RGB-D perspective, several strategies has been proposed to recover scene fields among consecutive frames, achieving a 3D local displacement characterization [5,6,7]. Such scene flows give important kinematics primitives but they are however, prone to errors because of the low resolution depth maps that limit the correlation with optical information along the sequence [8]. In [9] was proposed a scene flow approach that firstly computes a classical dense optical flow and then add depth information to recover 3D velocity information. This approach is nevertheless limited to include only brightness restriction which interferes with the field correspondence in depth. In [10] was proposed a simultaneous localization and mapping (SLAM) strategy that generates a 3D point cloud at each time. Hence, an environment measurement model (EMM) computes a set of rigid transformation between consecutive clouds to obtain a graph of correspondence to estimate a scene flow. This approach was extended to compute trajectories to track robots in controlled environments. However, the graph representation is computationally complex and the cloud correspondence can fail when abrupt motions are present. Herbest et. al. [7] also proposed a scene flow strategy but using non-linear operators to model color and depth. This approach is able to recover dense flows even in uniform scenarios. A main drawback of scene flow characterization is the description only between consecutive pairs of frames, limiting the kinematics description of objects to first order primitives.

This work introduces a novel strategy that computes 3D+t long trajectories as local kinematics primitives of RGB-D sequences. These long trajectories allow recovering coherent local motion information along the sequences that can be used to analyze object in the scene. Firstly, a scene flow is computed along the video to obtain sequential velocity fields. Hence, a local point tracking is defined over a grid of pixels to follow the corresponding velocity vectors. Several trajectories are removed because of abrupt motions among consecutive frames or given the static performance. Each of the recovered trajectories was characterized from local kinematics metrics, such as average and variation of velocities, curvature and torsion. This kinematics is a word to represent a set of object motions recorded in videos. The motion words are coded into a bag of words scheme to validate the performance of obtained trajectories. Validation of proposed trajectories was carried out into a scheme of recognition of gestures recorded in RGB-D sequences.

## 2. MATERIALS AND METHODS

In this work is presented a computational strategy to recover 3D+t trajectories able to locally characterize RGB-D sequences. The proposed approach starts by recovering 3D velocity fields between consecutive frames as an initial scene flow characterization (subsection 2.1). These velocity fields help to track uniform points along the sequence that constitutes long term trajectories (subsection 2.2). A set of differential kinematics are computed along each trajectory to obtain a video representation (subsection 2.3). Finally, a bag-of-kinematics words is built to obtain a final descriptor of the proposed trajectories in the problem of gesture representation (subsection 2.4). Next subsections describe each of the steps considered in the proposed strategy.

### 2.1 RGB-D scene flow characterization

The computation of motion fields, between consecutive frames, from optical flow strategies is one of the most relevant primitives in computer vision to represent objects of interest. These strategies are relatively independent to appearance, robust to some changes of interest and can recover additional features to complement a recognition analysis. Classical optical flows approaches include the Lucas and Kanade [11] that solve a linear problem to find edge motion vectors in the scene. On the other hand, Horn and Schunck [12] introduces global restriction and variational approaches to recover dense motion field's estimations. These two main works have given rise to novel applications and new strategies based on motion quantification [9], [13], and [7].

Taking advantage of RGB-D sequences, scene flow characterization strategies allow recovering 3D local velocities as a set of three-dimensional displacement vectors in consecutive frames. In this work it was implemented a variational strategy that includes local and global restrictions to preserve motion discontinuities and to recover scene flow estimations following the model proposed by Quiroga et.al. in [9]. This 3D flow strategy models rotational and translational motions by using local and global restrictions of 3d scene points. Such restriction allows a better estimation of 3D rotation and non-rigid motions.

In general a 3D motion could be measured as the difference between two consecutive cloud points, represented in real world coordinates as: $(X,Y,Z)_{t_i} - (X,Y,Z)_{t_{i+1}}$. This motion can be approximated as a scene flow $\hat{v} = (\hat{v}_x, \hat{v}_y, \hat{v}_z)$ taking into account the $x = (x,y)^T$ projection over the brightness images, together with corresponding depth values $z(x)$. Then a first motion consistency is defined from both: brightness $\Phi_b$ and depth consistency $\Phi_d$, measured over a local region $W(x;v)$. For doing so, the brightness consistency is measured among consecutive frames as: $\Phi_b(v_x, v_y) := \left\| I_{t_{i+1}}(W(x;v)) - I_{t_i}(W(x;v)) \right\|$, while the depth consistency is expressed as: $\Phi_d(v_z) := \left\| Z_{t_{i+1}}(W(x;v)) - Z_{t_i}(W(x;v)) \right\|$. Hence a general 3D flow motion consistency is defined in whole space **x** as:

$$E_D(\mathrm{v}) = \sum_x \Psi\Phi_b(v_x, v_y) + \lambda\Psi\Phi_d(v_z) \qquad (1)$$

where $\Psi(s^2) = \sqrt{s^2 + \varepsilon}$ is a Charbonnier penalty and $\lambda$ is a predefined weight for both considered consistencies. Additionally, a second restriction was considered to capture large coherent 3D displacements. This displacement allows recovering coherent abrupt motion that represents the kinematics signature of some objects. In such cases, a set of sparse SIFT points $m(x_{t_i})$ are computed from optical images, from which are measured a 3D motion consistency w.r.t to couple points in the next frame $\delta_{3D}(x_{t_i}, x_{t_{i+1}})$. The minimization matching term is then defined, as:

$$E_M(\mathrm{v}) = \sum_x p(x)\,\Psi\left(\left\|\delta_{3D}(x_{t_i}, x_{t_{i+1}}) - \mathrm{v}(\mathrm{x}_{t_i})\right\|\right) \quad (2)$$

with $p(x) = 1$ in regions around of an SIFT interest point. A third restriction $E_R(\text{v})$ is defined over the captured field and acts as local regularization term to favor locally rigid motions and preserve motion discontinuities in depth. This restriction is minimized w.r.t the gradients of computed field in each of their axis, and weighted by a function $\omega(x) = e^{-\alpha|\nabla Z(x)|}$ that helps to regularize the field w.r.t the computed depth map. This flow rule minimization is expressed as:

$$E_R(\text{v}) = \sum_x \omega(x)\left|\nabla v_{t_i}(\text{x})\right| \qquad (3)$$

This approach uses a regularized variational function to obtain a total dense scene field, while preserving motion discontinuities. The scene flow is finally obtained as the sum of three restriction defined above. In Fig. 1 is illustrated a typical computation of scene flow for a couple of frames. It is shown a gray-map representation of the $\vartheta_x, \vartheta_y, \vartheta_z$ components separately.

Despite of 3D motions have demonstrated to be useful to describe dynamic scenes, these primitives are limited to measuring only each consecutive pair of frames. From RGB-D information, the implemented strategy achieves a dense 3D field representation of the scene.

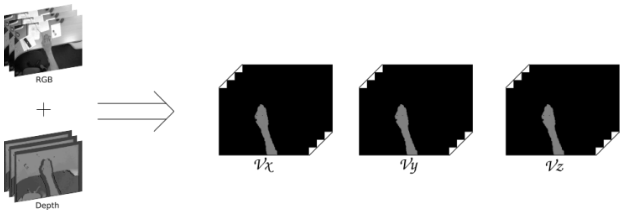

Fig. 1. Scene flow. The 3D motion between a pair of consecutive RGB-D images is obtained from the scene flow information. Please visit the site: (to be announced) to observe the image in color.
Source: The authors.

## 2.2 Computing long motion trajectories

Taking advantage of dense optical flow description, in the literature has been proposed new kinematics models that search to extend motion information in more than two frames. It consists of local motion trajectories that represent local primitives that follow interest points along the sequences, describing kinematics information in relatively large intervals of time [14,15]. For instance the KLT-Tracker uses an extension of pyramidal Lukas-Kanade [16] to follow relevant motion vectors, obtaining coherent motion trajectories of coherent edges along the sequence. Such representation is nevertheless dispersed and only produces few trajectories to represent the motion of an object. Also, Sun et. al. [14], proposed the capture of salient interest SIFT points by carrying out a coherent matching along the sequence. This approach extends the performance of SIFT points, being invariant to scale and rotation but only representing few points along the scene which can limit a statistical representation of the objects. On the other hand, Wang et. al. proposed a dense trajectories representation by following independent fields of motion along the videos. Such representation has been proofed to be useful in action representation tasks for classical RGB sequences.

Inspired in such dense trajectories, in this work is proposed an extension of dense trajectories computed in RGB-D spaces, which allows capturing $(3D + t)$ long trajectories. These proposed trajectories can enrich the kinematics description of the object by computing higher-order motion representations. Also, the trajectories can be used as bases to recover local RGB-D descriptors. For doing so, the proposed strategy starts by computing a point cloud from the RGB-D sequences at each time. This cloud point assumes a pre-processing and calibration of RGB and depth images. Then, a dense sampling is carried out from the cloud point $P_t(x) = (x_t, y_{t,} z_t)$ taking a grid of spatially distributed points along of each frame.

Thereafter, the scene flow, herein implemented, is computed among a couple of frames to obtain a basic 3D velocity representation of the sequence for the cloud of RGB-D points. In such way, each of the points $P_t(x) = (x_t, y_{t,} z_t)$ gotten from the dense grid is tracked to the following frame from the respective vector of displacement $\vartheta = (\vartheta_x, \vartheta_y, \vartheta_z)$, computed in the scene flow characterization. As in RGB strategies along consecutive frames, the displacement vector $\vartheta$ can represent incoherent abrupt motions. The displacement of the point is filtered by using a classical median operator, over a neighborhood $\Omega$, as, $P_{t+i}(x) = P_t(x) + med\{v(x_{t_i})\}_{\Omega}$, preserving the structure of points along the sequence. The set of points that are tracked according to the associated velocity vector form the $(3D + t)$ motion trajectory $(P_t(x), P_{t+1}(x), P_{t+2}(x), \ldots, P_{t+n}(x))$. A typical trajectory representation is illustrated in Fig. 2, where the displacement of point $(x, y)$ is drawn in blue, and for the z-component, each trajectory is drawn using a color-map, where green represents trajectories that are close to the camera, while blue represent low motion in z. This 3D tracking introduces important features about geometry of the object as well as the 3D motion displacement developed during a particular action.

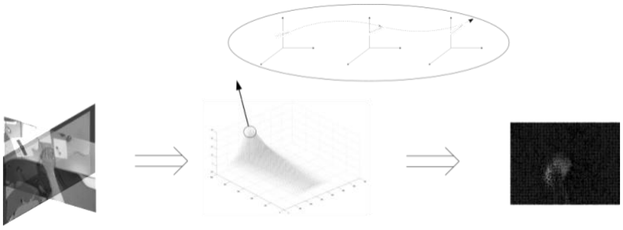

Fig. 2. (3D+t) motion trajectories. In left subplot is illustrated the RGB-D information of captured hand. Then, between consecutive frames is computed a scene flow (middle subplot) that recovers the velocity of hand and hence its main geometrical structure. From such scene flow is tracked 3D points along video to obtain (3D+t) long trajectories as illustrated in right subplot. In this plot, color of trajectories represent the depth displacement, being green a color that represent that trajectories are close to the camera, while red color indicates that the trajectories are far. Please visit the site: (to be announced) to observe the image in color.
Source: The authors.

Some spatial filters are implemented to remove trajectories that have strong motions among consecutive frames or trajectories with static performance. These spatial filters are implemented by using the temporal thresholding of the trajectory's variance. An illustration of the process to compute 3D motion trajectories is presented in Fig. 2, where left plot represent raw cloud point information, and, how scene flow is computed among consecutive frames (middle plot). 3D+t long trajectories (right panel) are then obtained by tracking several points from scene flow.

### 2.3 Kinematics trajectory features

The set of $(3D+t)$ long trajectories captured along the video could be used as kinematic independent words to represent the gestures in a video. These trajectories are rich in dynamic spatio-temporal information and result fundamental to characterize motion. Then, a set of kinematic differential measures were computed from each trajectory to represent the words that represent a specific activity. In this work, it was computed the average $\mu$ and standard deviation $\sigma$ of the local speed along each trajectory, as: $\{\mu(\|\mathrm{v(x)}\|), \sigma\left(\big\|\|\mathrm{v(x)}\|\big\|\right)\}$.
Since trajectories track motion points along the sequence and represent long local information, additional kinematics and analysis can be carried out. For instance, a complementary measure in this work was the curvature, that allows describing how rapidly the trajectory is bending along the video. Also the torsion was considered in this work as an additional 3D measure of motion trajectory. The computation of such metrics were implemented according to [27], following finite Euclidean differences. For instance the curvature is approximated by the circle that is circumscribed around three consecutive points of the trajectory $P_{i-1}(x), P_i(x), P_{i+1}(x)$. Then, the curvature $\kappa$ can then be expressed as:

$$\kappa\big(P_i(x)\big) = 4\frac{\sqrt{\hat{s}(\hat{s}-a)(\hat{s}-b)(\hat{s}-c)}}{abc} \qquad (4)$$

where $\hat{s} = (a+b+c)/2$ and $a = \|P_{i-1} - P_i\|$, $b = \|P_i - P_{i+1}\|$ and $c = \|P_{i-1} - P_{i+1}\|$ are three consecutive segments of the trajectory. Also, the torsion $\tau$ could be approximated using six consecutive segments of the trajectory, as:

$$\tau\big(P_i(x)\big) = 6\frac{H}{\Big(\kappa\big(P_i(x)\big)\Big)} \qquad (5)$$

with $H$ as the height of the tetrahedron formed from each of the six consecutive segments of the trajectory. An important feature of the proposed $(3D+t)$ trajectories is the posibility to expand kinematics features that involve depth information, as the torsion.

### 2.4 Mid-level trajectory representation for recognition

The set of computed motion trajectories constitutes a set of kinematics primitives to represent the particular performance of objects of interest. Each of the trajectories was codified with a set of kinematic measures as words of representation into Bag-of-kinematic words (BoKW). This mid-level representation is widely used in different areas of interest such as natural language processing,

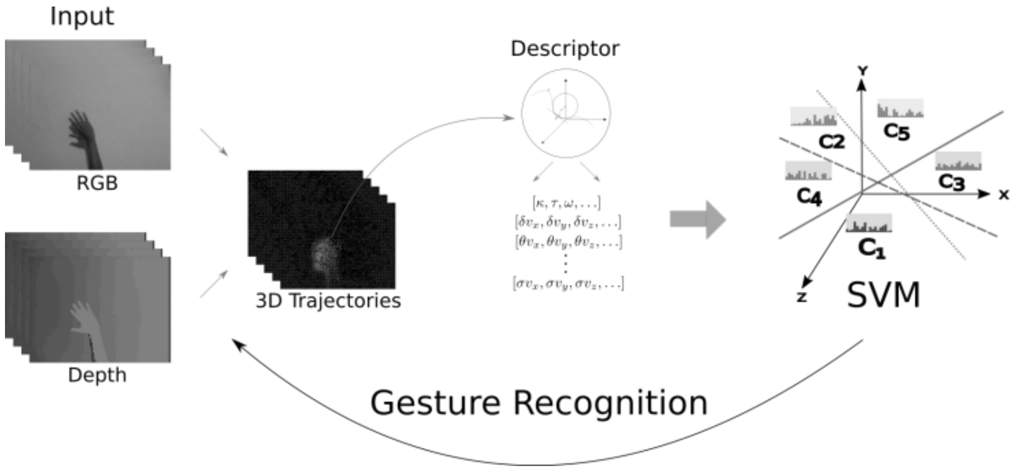
Input
RGB
Depth
3D Trajectories
Descriptor
$[\kappa, \tau, \omega, \ldots]$
$[\delta v_x, \delta v_y, \delta v_z, \ldots]$
$[\theta v_x, \theta v_y, \theta v_z, \ldots]$
$\vdots$
$[\sigma v_x, \sigma v_y, \sigma v_z, \ldots]$
Y
X
Z
C2
C5
C3
C4
C1
SVM
Gesture Recognition

Fig. 3. Recognition from proposed (3D+t) trajectories. (a) RGB-D acquisitions. (b) The dense scene flow is obtained and the (3D+t)trajectories are calculated. (c) From long (3D+t)trajectories, a set of kinematics primitives are computed to code motion information that can be associated with gestures. In this work was computed mean and standard deviation velocities, as well as the curvature and torsion of each trajectory (d) This local kinematic primitives are coded as motion words into a scheme of bag-of-kinematics-words and then occurrence histograms are computed. Such histograms are mapped to SVM to predict the gesture. Please visit the site: (to be announced) to observe the image in color. Source: The authors.

computer vision for detection and recognition of objects in images and videos, among others.

Particularly in this work, a set of kinematic words, computed from a set of training videos, were used as input in a non-supervised k-means algorithm to recover k representative kinematics words that represent in general whole actions in video. Then, on testing step, a new video is coded with $(3D + t)$ long trajectories and then a set of kinematic words are computed for the whole video. Thereafter, each kinematic word is projected to the trained dictionary to compute the most similar centroid. From such comparison is built an occurrence histogram that represents the descriptor of each video.

Finally, the occurrence histograms are computed for all the dataset, i.e., training and testing videos. The set of training histograms are used to compute a machine learning model that allows a posterior classification. In this work it was implemented a support vector machine strategy (SVM) because its well-known performance in different areas of recognition and the efficiency to obtain results [28, 29]. Then, for testing it was mapped the histogram occurrences to the previously trained SVM model and it was obtained automatically a label of the gesture in video. Fig. 3 shows the general scheme used to recognize particular RGB-D gestures from the proposed $(3D + t)$ motion trajectories.

### 2.5 Data

In literature it has been reported RGB-D datasets for different purposes [30, 31, 32]. Nevertheless, such datasets are namely captured for static recognition tasks and only independent images are captured to describe the object observations. Also, some proposed datasets are captured to describe postural gestures but with strong limitations along time. In such cases, few disperse frames are taken from videos to represent main poses of a particular action. Also some datasets are restricted to two continuous frames to test scene flow algorithms.

A novel motion RGB-D dataset is herein reported that result useful to evaluate strategies that compute kinematic information along sequences. From this dataset is possible to capture smooth trajectories along time and also to compute scene flows, allowing to measure the coherence along the video sequence. The proposed dataset was captured from one kinect V1 camera and the raw sequence of frames was recovered with the open frameworks OpenNI (https://structure.io/openni) and libfreenect (https://github.com/OpenKinect/libfreenect). Each video sequence was captured in a spatial resolution of 640x480 with a temporal resolution of 30 fps in

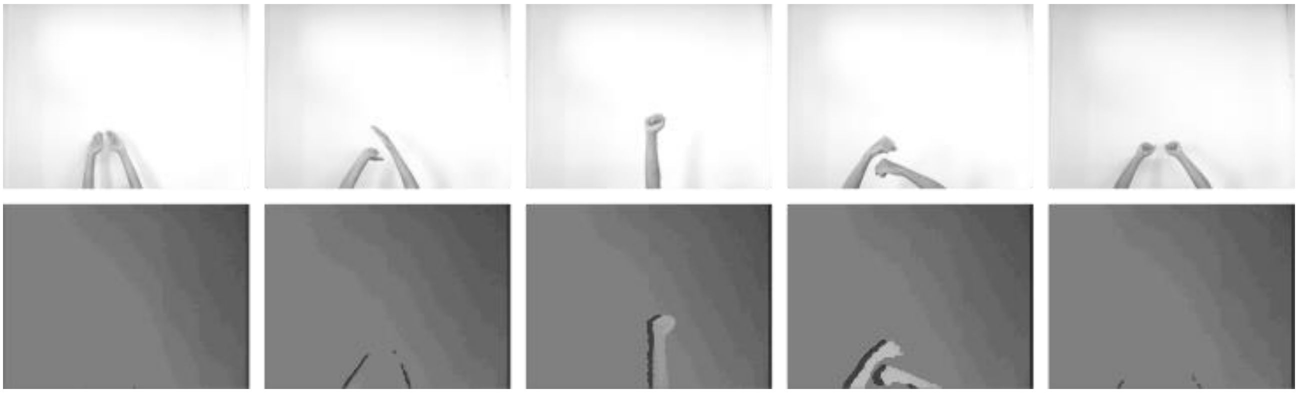

Fig. 4. Dataset to evaluate the proposed approach. This dataset is available at: (to be announced) repository (to be announced). In the first row it is shown the first frame of each one of the actions in the intensity channel, and their corresponding depth maps in the second row. A total of five different gestures were computed into a semi-controlled scenario. The 3D motions as representative keys of the recorded gestures. Please visit the site: (to be announced) to observe the image in color. Source: The authors.

both channels RGB and depth. The proposed dataset was captured in a controlled scenario, with uniform background and relative small camera jitters. A total of five persons were recorded where each one develops five different actions with one or both hands. The selected actions involve several gestures with different motion representative signatures and with displacements in the depth axis. Each selected gesture was recorded five times to obtain a range of statistical variation of each motion. A total of 125 videos were recorded. In Fig. 4 is illustrated the different gestures captured in the proposed dataset. The dataset is available at: (to be announced) repository (to be announced).

## 3. RESULTS

The proposed strategy computed a set of $(3D + t)$ long trajectories as kinematics primitives to represent gestures captured in RGB-D scenarios. A first qualitative evaluation of the computed trajectories is illustrated in Fig. 5, in which the trajectories are computed for a particular gesture with two hands. For this gesture, a hand rotation in a vertical axis is carried out. As observed in the figure, the most important information is coherent with hand motion. The color of each trajectory is represented in a color map, where gray scale color represents motion in depth that is closer to the acquisition camera.

Secondly, a quantitative evaluation was performed by measuring the capability of representation of (3D+t) motion trajectories. For doing so, each trajectory captured in the sequence was characterized by using a set of kinematics, namely: velocity means, velocity deviations, curvatures and torsion. Then, a bag of kinematic words was codified to represent each of the gestures described in the evaluated dataset. A k-fold cross validation was performed to obtain a statistical significance of the results, by fixing k = 25. For the whole considered experiments each of the computed trajectories was characterized with a total of 8 scalar kinematics. Also the BoKW was trained with a dictionary of a total of 250 centroids. Hence, the final gesture descriptor recover very compact histograms 250 occurrence bins w.r.t the learned kinematic dictionary.

In a first experiment it was evaluated the capability of representation of the proposed strategy for the three most different gestures, namely: “RING”, “TRIANGLE” and “FLEXION”. In Table 1 is reported the obtained results for each of the subject included in the experiments. In general, the proposed approach achieves an average accuracy of 76% for a total of 100 videos. In subject 2 there exist some limitations because some noise in the sequence limit the computation of an appropriate scene flow and then some few trajectories are recovered to represent the gestures.

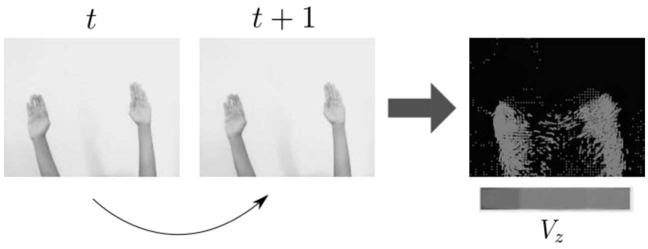

Fig. 5. 3D trajectories obtained for the gesture “ring”. Color represents the displacement in depth coded in the color-map shown in the color-bar. Please visit the site: to observe the image in color and appreciate the difference in depth with respect to the camera, achieved by the proposed approach. Please visit the site: (to be announced) to observe the image in color. Source: The authors.

Table 1. Gesture classification of 4 subjects, 3 classes and 8 bins in the final descriptor. Mean accuracy: 76%.

| **Subjects** | **Accuracy (%)** |
|---|---|
| **$P_1$** | 80 |
| **$P_2$** | 64 |
| **$P_3$** | 88 |
| **$P_4$** | 72 |

Source: The authors.

In the same direction, in a second experiment it was included an additional gesture into the framework of evaluation, the gesture “ROTATION”. In Table 2 are reported the results obtained for this experiment. Interestingly, the proposed approach achieves better rates of recognition with an additional gesture, obtaining and average score of 80% for a total of videos of 100. This result also show the stable performance of the proposed approach to analyze and characterize different gestures captured in RGB-D sequences.

In a third experiment, as reported in Table 3, was evaluated the entire dataset by including all of the gestures. In such case, the average accuracy of the proposed approach is 77% which is favorable in terms of the capability of the approach to represent gestures. It can be observed that the subject $P_3$ has a lower score because the actor develops the action at a different velocity w.r.t to the others, which combined with some artifacts in scene, results in few trajectories of representation.

Table 2. Gesture classification of 4 subjects, 4 classes and 8 bins in the final descriptor. Mean accuracy: 80%.

| **Subjects** | **Accuracy (%)** |
|---|---|
| **$P_1$** | 92 |
| **$P_2$** | 72 |
| **$P_3$** | 68 |
| **$P_4$** | 88 |

Source: The authors.

Table 3. Gesture classification of 4 subjects, 5 classes and 8 bins in the final descriptor. Mean accuracy: 77%.

| **Subjects** | **Accuracy (%)** |
|---|---|
| **$P_1$** | 92 |
| **$P_2$** | 72 |
| **$P_3$** | 68 |
| **$P_4$** | 88 |

Source: The authors.

In Table 4 is reported an additional experiment with all gestures but using a different combination of subjects, where all the possible combinations $\binom{4}{3}$of 3 persons were used to carried out the analysis. From this experiment it can be confirmed that subject $P_3$ reports some variations w.r.t to the rest of the population, which impact in lower results. The proposed approach is however robust to capture this variations and try to predict correctly the developed gesture.

Table 4. Gesture classification of 3 subjects, 5 classes and 8 bins in the final descriptor. Mean accuracy: 75.33%.

| **Subjects** | **Accuracy (%)** |
|---|---|
| **$P_1$, $P_2$, $P_3$** | 73.33 |
| **$P_1$,$P_2$, $P_4$** | 84 |
| **$P_1$,$P_3$, $P_4$** | 76 |
| **$P_2$,$P_3$, $P_4$** | 68 |

Source: The authors.

Finally, a confusion matrix was computed using a total set of 5 gestures (Fig. 7). As it can be observed, almost all the gestures have a rate recognition above of 80%. Particularly, the gesture rotation reports some misclassifications because the kinematics similarities with the gestures 3 and 5. In general, the $(3D + t)$ motion trajectory demonstrates to be robust to represent kinematics gestures.

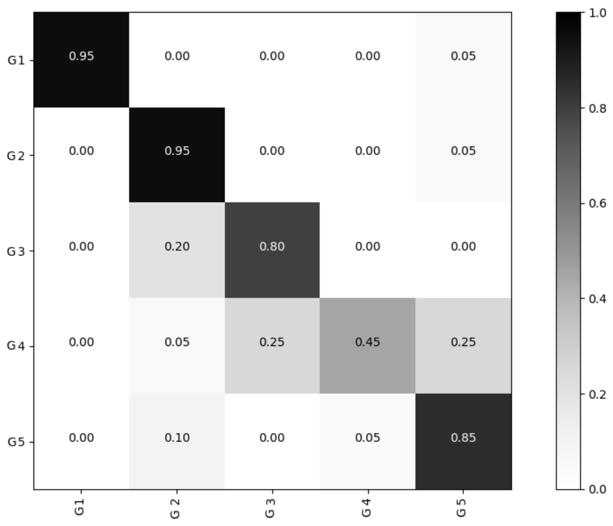

Fig. 7. Confusion matrix for the best results obtained in the experiment. In each row/column is represented the correlation of each gesture G with the other.

## 4. DISCUSSION

This work introduced a novel approach to recognize 3D gestures, captured from a kinetic sensor, and using long 3D+t motion trajectories. The complete video sequence is characterized by a set of long motion trajectories with the main advantage of describing motion in space as well as in depth. Then, each of the trajectories was characterized using kinematic primitives, like speed, velocity angle, curvature and torsion. Such primitives allow to build a mid-level representation and then a 3D gesture descriptor is coded as occurrence histograms projected over such representation. These descriptors are mapped to a support vector machine strategy to obtain predictions.

In literature has been proposed some approaches that are based on detection of interest points from appearance information [1,2]. These approaches are however dependent on image properties, and require a proper definition of the object of interest. In contrast, the proposed approach holds a relative independence of appearance, and gesture description is only based on motion information. Over a dataset with more than 100 video sequences the proposed approach achieves accuracy scores in average of 80%. These gestures were performed by different actors, which result fundamental to evaluate color dependency of descriptor.

Other approaches, like proposed in [4] used temporal information to code gesture descriptors in RGB-D scenarios. Such approaches are sensible to occlusions and the well definition of shape. In such sense the mid-level representation of our approach allows to recover and characterize incomplete gestures. Recently, other approaches have tried the motion characterization of RGB-D sequences by computing scene flow methodologies [5,6,7]. These approaches result very important to build kinematic representations in 3D+t scenes. However, a main limitation of such approaches is that the description of motion fields is applied only among consecutive frames, limiting the kinematic history description. The herein proposed approach takes advantage of such 3D motion field and selects a set of points, which are followed during time. The resulting long trajectories allow a better characterization of kinematics, with the possibility of incorporating higher representations such as the curvature and torsion.

In general the kinematic achieves more than 90% for some specific gestures, because their well-defined motion signature. Other gestures are however composed of several common motions, and the description result difficult when using only kinematic information.

## 5. CONCLUSIONS

This work presented a novel strategy to compute long motion primitives in RGB-D space. The primitives are represented as (3D+t) trajectories that are a set of points tracked along the sequence, according to specific scene flow information. These motion trajectories were characterized with differential kinematics and included in a bag-of-words representation. A new dataset was also herein computed to evaluate motion information from particular objects of interest in RGB-D sequences. In the task of classification, the motion trajectories proved to be robust to represent different RGB-D gestures, from very compact descriptors.

For a set of five different gestures, developed by 4 different actors, the proposed strategy achieved an 80% of average accuracy, by only using histograms of 8 bins. Future work includes the development of trajectories that only follow interest points in depth and the evaluation on more rich datasets and scenarios.

## 6. ACKNOWLEDGEMENTS

This work was partially funded by the Universidad Industrial de Santander. The authors acknowledge the Vicerrectoría de Investigación y Extensión (VIE) of the Universidad Industrial de Santander for supporting this research registered by the project: *Reconocimiento continuo de expresiones cortas del lenguaje de señas*, with SIVIE code 2430.

## 7. REFERENCES

[1] Blum M., Springenberg, J.T., Wülfing, J. and Riedmiller, M. A learned feature descriptor for object recognition in RGB-D data, 2012 IEEE International Conference on Robotics and Automation (ICRA), 2012. pp. 1298–1303. DOI: 10.1109/ICRA.2012.6225188.

[2] Zhao, Y., Liu, Z., Yang, L. and Cheng, H. Combing RGB and depth map features for human activity recognition, Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Asia-Pacific, 2012. pp. 1–4.

[3] Bo, L., Ren, X. and Fox, D. Depth kernel descriptors for object recognition., 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2011. pp. 821–826. DOI: 10.1109/IROS.2011.6095119.

[4] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. and Blake, A. Real-time human pose recognition in parts from single depth images, 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. pp. 1297–1304. DOI: 10.1109/CVPR.2011.5995316.

[5] Vedula, S., Baker, S., Rander, P., Collins, R. and Kanade, T. Three-dimensional scene flow, Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999. Volume 2, pp. 722–729. DOI: 10.1109/ICCV.1999.790293.

[6] Huguet, F. and Devernay, F. A variational method for scene flow estimation from stereo sequences, 11th IEEE International Conference on Computer Vision ICCV, 2007. pp. 1–7. DOI: 10.1109/ICCV.2007.4409000.

[7] Herbst, E., Ren, X. and Fox, D. RGB-D flow: Dense 3-d motion estimation using color and depth, 2013 IEEE International Conference on Robotics and Automation (ICRA), 2013. pp. 2276–2282. DOI: 10.1109/ICRA.2013.6630885.

[8] Khoshelham, K. Accuracy analysis of kinect depth data, ISPRS workshop laser scanning, 2011. Volume 38, pp. W12.

[9] Quiroga, J., Devernay, F. and Crowley, J. Local/global scene flow estimation, 20th IEEE International Conference on Image Processing (ICIP), 2013. pp. 3850–3854. DOI: 10.1109/ICIP.2013.6738793.

[10] Endres, F., Hess, J., Sturm, J., Cremers, D. and Burgard, W. 3-D mapping with an RGB-D camera. IEEE Transactions on Robotics,

30(1):177–187, 2014. DOI: 10.1109/TRO.2013.2279412.

[11] Lucas, B., Kanade, T., et al. An iterative image registration technique with an application to stereo vision, Proceedings of the 7th international joint conference on Artificial intelligence, 1981. pp. 674-679.

[12] Horn, B. and Schunck, B. Determining optical flow. Artificial intelligence, 17(1-3):185–203, 1981.

[13] Quiroga, J., Brox, T., Devernay, F. and Crowley, J. Dense semi-rigid scene flow estimation from RGB-D images. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8695. Springer, Cham. pp. 567–582. DOI: 10.1007/978-3-319-10584-0_37.

[14] Sun, S.-W., FrankWang, Y.-C., Huang, F. and Liao, H.Y. Moving foreground object detection via robust SIFT trajectories. Journal of Visual Communication and Image Representation, 24(3):232–243, 2013. DOI: 10.1016/j.jvcir.2012.12.003.

[15] Wang, H., Kläser, A., Schmid, C. and Liu, C.-H. Action recognition by dense trajectories, 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. pp. 3169–3176. DOI: 10.1109/CVPR.2011.5995407.

[16] Shi, J. et al. Good features to track. 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1994. pp. 593–600.

[17] Dollár, P., Rabaud, V., Cottrell, G. and Belongie, S. Behavior recognition via sparse spatio-temporal features. 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. pp. 65–72. DOI: 10.1109/RME.2005.1543049.

[18] Cai, Z., Han, J., Liu, L. and Shao, L. RGB-D datasets using Microsoft kinect or similar sensors: a survey. Multimedia Tools and Applications. 76(3): 4313–4355, 2017. DOI: 10.1007/s11042-016-3374-6.

[19] Xia L. and Aggarwal J.K.. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013. pp. 2834–2841. DOI: 10.1109/CVPR.2013.365.

[20] Li, W., Zhang, Z. and Liu, Z. Action recognition based on a bag of 3D points. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010. pp. 9–14. DOI: 10.1109/CVPRW.2010.5543273.

[21] Yang X. and Tian, J.L.. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012. pp. 14–19. DOI: 10.1109/CVPRW.2012.6239232.

[22] Lowe, D.G.. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 100 2004. DOI: 10.1023/B:VISI.0000029664.99615.94.

[23] Xiao, Y., Zhao, G., Yuan, J. and Thalmann, D. Activity recognition in unconstrained RGB-D video using 3D trajectories. In SIGGRAPH Asia 2014 Autonomous Virtual Humans and Social Robot for Telepresence, 2014. pp. 4. DOI: 10.1145/2668956.2668961.

[24] Willems, G., Tuytelaars, T. and Van Gool, J. An efficient dense and scale-invariant spatio-temporal interest point detector. Computer Vision–ECCV 2008, 2008. pp. 650–663. DOI: 10.1007/978-3-540-88688-4_48.

[25] Laptev, I. and Lindeberg, T. Space-time interest points. Proceedings 9th International Conference on Computer Vision, Nice, France, 2003. pp. 432–439. DOI: 10.1109/ICCV.2003.1238378.

[26] Ren, X., Bo, L. and Fox, D. RGB-(D) scene labeling: Features and algorithms. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012. pp. 2759–2766. DOI: 10.1109/CVPR.2012.6247999.

[27] Boutin, M. Numerically invariant signature curves. International Journal of Computer Vision

40(3): 235-248, 2000. DOI: 10.1023/A:1008139427340.

[28] Schuldt, C., Laptev, I. and Caputo, B. Recognizing human actions: a local SVM approach. Proceedings of the 17th International Conference on Pattern Recognition ICPR. 2004. pp. 32-36. DOI: 10.1109/ICPR.2004.1334462.

[29] Basura, F. et al. Rank pooling for action recognition. IEEE Transactions On Pattern Analysis And Machine Intelligence, 39(4): 773-787, 2017. DOI: 10.1109/TPAMI.2016.2558148.

[30] Janoch, A. et al. A category-level 3D object dataset: Putting the kinect to work. Consumer depth cameras for computer vision. Springer, London, 2013. pp. 141-165. DOI: 10.1007/978-1-4471-4640-7_8.

[31] Lai, K. et al. RGB-D object recognition: Features, algorithms, and a large scale benchmark. Consumer Depth Cameras for Computer Vision. Springer, London, 2013. pp. 167-192. DOI: 10.1007/978-1-4471-4640-7_9.

[32] Ni, B., Wang, G., Moulin, P. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. Proceedings of IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011. pp. 6-13. DOI: 10.1109/ICCVW.2011.6130379.