3D Human Poses Estimation from a single 2D silhouette

Fabrice Dieudonné Atrevi, Damien Vivet, Florent Duculty, Bruno Emile

3D Human Poses Estimation from a single 2D silhouette

Fabrice Dieudonné Atrevi[1], Damien Vivet[1], Florent Duculty[1] and Bruno Emile[1]
[1]*Univ. Orléans, PRISME, EA 4229, F45072, Orléans, France*
{*damien.vivet, bruno.emile,florent.duculty*}*@univ-orleans.fr, atrevifabrice@gmail.com*

Keywords: Pose estimation, 3D pose, 3D modeling, skeleton extraction, shape descriptor, geometric moment, Krawtchouk moment.

Abstract: This work focuses on the problem of automatically extracting human 3D poses from a single 2D image. By pose we mean the configuration of human bones in order to reconstruct a 3D skeleton representing the 3D posture of the detected human. This problem is highly non-linear in nature and confounds standard regression techniques. Our approach combines prior learned correspondences between silhouettes and skeletons extracted from 3D human models. In order to match detected silhouettes with simulated silhouettes, we used Krawtchouk geometric moment as shape descriptor. We provide quantitative results for image retrieval across different action and subjects, captured from differing viewpoints. We show that our approach gives promising result for 3D pose extraction from a single silhouette.

1 INTRODUCTION

Recognizing human actions is really challenging for computer vision scientists and researchers since the last two decades (Wang et al., 2011). Nevertheless, human action recognition systems have a lot of possible applications in surveillance, pedestrian tracking and Human Machine Interaction (Aggarwal and Cai, 1999). Human pose estimation is a key step to action recognition.

A human action is often represented as a succession of human poses (Wang et al., 2013). As these poses could be 2D or 3D, so estimating them have attracted a lot of attention. A 2D pose is usually represented by a set of joint locations (Yang and Ramanan, 2011) whose estimation remains challenging because of the human body shape variability, viewpoint change, etc. Considering 3D pose, we usually represent it by a skeleton model parameterized by joint locations (Taylor, 2000) or by rotation angles (Lee and Nevatia, 2009). Such representation has the advantage to be Viewpoint-invariant however, estimating 3D poses from a single image still remains a difficult problem. The reasons are multiple. First, multiple 3D poses may have the same 2D pose reprojection. Second, 3D pose is inferred from detected 2D joint locations so 2D pose reliability is essential because it greatly affects skeleton estimation performance. In camera network used in a video-surveillance context, image quality is often poor making 2D joint detection a difficult task, moreover camera parameters are unknown making the correspondence 2D/3D difficult.

In this work we propose a new technique for the extraction of 3D skeleton pose assumptions from a single 2D image based on the silhouette shape recognition. This technique is based on the use of a 3D human pose and action simulator. A silhouette database is constructed from this simulator and is used in order to match nearest silhouette and as a results possible 3D human pose.

This article presents a silhouette shape description and comparison between different subjects and action steps and show that we can obtain 3D skeleton configuration by using only a single 2D silhouette detection. Section 2 presents related works in the human skeleton and action recognition. Section 3 presents the global framework of the method and the 3D simulation used. Section 4 deals with Krawtchouk shape descriptors applied to human silhouettes. Finally, section 3.1 and 5 present the databases and the obtained results.

2 RELATED WORKS

There are many methods in the state-of-the-art that deals with the human pose estimation and action recognition. Nevertheless, these tasks are still challenging for computer vision community. Human activity analyses started with O'Rourke and Badler

(O'Rourke et al., 1980) and Hogg (Hogg, 1983) in the eighties. Since last decades scientists proposed many approaches. We can categorize these approaches into two main categories: on one hand the methods using 3D information and on the other hand technics using only 2D data.

Most of the approaches use a 3D model or 3D detection for estimating the pose of a subject and for action classification. Rehg and Kanade (Rehg and Kanade, 1994) presented a 3D model-based hand tracking system that can cover the state of a 27 DOF skeleton. Gavrial and al.(Gavrila and Davis, 1996) used a 3D model-based tracking of unconstrained human movement. They used some sequence images acquired from multiple views for recovering 3D body pose of a human.

Bourdev and Malik (Bourdev and Malik, 2009) estimated the human pose from key points. They used a data set of annotations of human with 3D joins informations inferred using anthropometric constraints for human action classification (Maji et al., 2011). Hiyadi and al. (Hiyadi et al., 2015) used the depht information obtained from Kinect sensor and a tracking algorithm for 3D human gestures recognition. Jian (Jiang, 2010) proposed an exemple-based method, based on the kd-tree achieves real-time performance, to prune the hypotheses. Ramakrishna and al. (Andriluka et al., 2010) proposed a three-stage process for 3D poses recovering in uncontrolled environment. Valmadre and Lucey (Valmadre and Lucey, 2010) used deterministic structure from multiple view of motion, based on the related work of Wei and Chai (Wei and Chai, 2009), for 3D pose estimation.

These approaches need multiple sensors or specific devices such as time of flight or active camera for acquiring 3D information. These models also, need good parametrization.

The second category of approaches, to which our proposed method belongs, used 2D models trained from various images. Baumberg and Hogg (Baumberg and Hogg, 1994) used active shape model to track pedestrians in real world scenes. They used the B-spline as a shape vector for training the model. Wren and al. (Wren et al., 1997) tracked people and interpreted their behaviour by using a multiclass statistical model of colour and shape to obtain 2D representation of head and hand. Gorelick and al. (Gorelick et al., 2005) used the solution of Poisson's equation to extract spatiotemporal features such as the saliancy, the orientation of the shape for action recognition and then human pose estimation. Guo and al. (Guo et al., 2009) used a geometrical normalized vector of dimension 13 for describing the shape of a human. Mori and Jitendra (Mori and Malik, 2002), or Agarwal and Triggs (Agarwal and Triggs, 2006) used the shape context in their research on human pose estimation. Gorce and al. (de La Gorce et al., 2011) estimated and tracked the human hand from monocular video through minimization of an objective function. This minimization is done using a quasi-Newton method, for which they provide a rigorous derivation of the objective function gradient. Yang and Ramanan (Yang and Ramanan, 2011) estimated the pose by capturing the orientation of each part with a mixture of templates modeled by linear SVMs. All of these methods focus on 2D image interpretation in order to detect human pose or action. For this purpose, learning is requiered and such algorithms need complex and expensive systems to get the training data set with the ground truth.

Our method is based on a very simple silhouette extraction and description. We use the robust Krawtchouk geometric moment to shape analysis in monocular image. For the database, we proposed to use software applications from the open source community. These softwares makes realistic simulation of various human poses and action possible. We have shown in this work that using 3D simulations for learning, without complex machine learning algorithm and with a simple real time shape descriptor we can achieve 3D pose estimation on real data with good accuracy from a unique 2D image.

3 METHODOLOGY

The proposed approach for pose estimation is based on shape analysis of human silhouette. The method can be decomposed into four parts: (1) simulated silhouette and skeleton database, (2) Human detection and 2D silhouette extraction, (3) silhouette shape matching, (4) skeleton scaling and validation. The workflow is presented Fig.1.

(1) First, silhouette and skeleton database is built thanks to opensource 3D software (see section 3.1). Such database is composed of human silhouettes and its corresponding 3D skeletons for different kind of actions we want to recognize. So, for a requested silhouette, it'll be possible to find the matching silhouette in the database and then the corresponding 3D skeleton.

(2) 2D silhouettte detection is a well-studied field in machine learning and computer vision. For this purpose we used classical real-time approach proposed by Dollar et al. (P. Dollar and Perona, 2010) based on multiscale HOG (Dalal and Triggs, 2005). Once the human silhouette is detected, we converted it in a 128 x 48 pixels image for solving the translation and scale

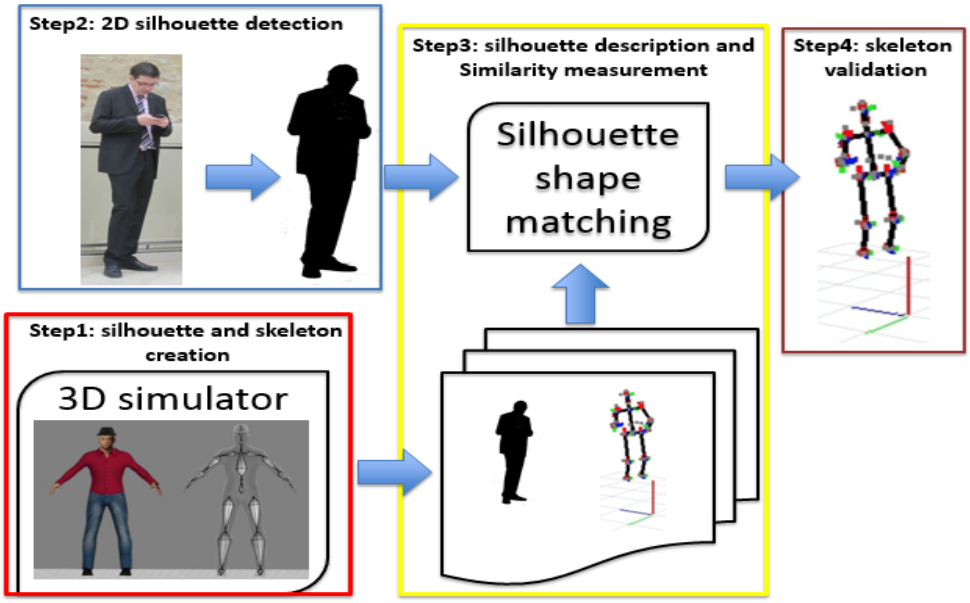


Figure 1: Human pose estimation methodology.

problem.

(3) Silhouette description and similarity measurement is the key point of our methodology. The main objective is to describe accurately the shape of the silhouette. For this task, we used the geometric moment of Krawtchouk because of its robustness compared to Hu, Zernike or Shapecontext descriptors. (See section 4) Based on this descriptor, a characteristic vector is computed for each silhouette in the database. The similarity between characteristic vector is measured with the Euclidean distance given by :

$$d(z^r, z^t) = \sum_{i=1}^{T} \left(z_i^r - z_i^t\right)^2 \quad (1)$$

where z^r et z^t is respectively the characteristic vector of request silhouette and the t th silhouette in the database.

(4) Skeleton scaling and validation. For each silhouette we retrieve a 3D skeleton. This skeleton is scaled to the current silhouette size. At this step we use ground truth simulated database to valide the approach. The confidence score is process by measuring the reprojection error of predicted joints on the silhouette.

3.1 Construction of the 2D/3D matching database

3.1.1 3D human avatar and action simulation

In order to build our simulated humans, we choose to use a professional free and open-source 3D computer graphics software called *Blender*[1] associated with a free software to create realistic 3d human *makehuman*[2] (see Fig. 2). These avatars can be animated thanks to motion capture data in order to simulate very realistic actions.

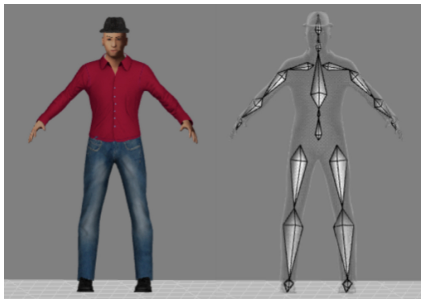


Figure 2: 3D simulated avatar and its associated skeleton

In these softwares, we simulate different human avatars with different morphologies and clothes and animate them with different realistic motions taken from the CMU motion capture database[3]

3.1.2 Database construction

In the 3D computer graphics software, we positioned on an emisphere a virtual camera looking at the subject. For each movement of the avatar, we record

[1] https://www.blender.org/

[2] http://www.makehuman.org/

[3] *The data used in this project was obtained from mocap.cs.cmu.edu.*

both: 2D image and silhouette (see fig 3), 3D camera poses and 3D joints and bones poses. As a result for each subject's pose we can collect the detected silhouette related to its 3D skeleton which contains 19 bones. We recorded in 4 subjects with different phenotypes and for 4 differents animations: walk cycle, basket action, jumb and climb. As a result, we obtained 2925 couples silhouette / 3D skeleton.
For each silhouette, we calculated the feature vector of the shape descriptors presented in section 4 and the 2D poses of reprojected joints for quantitative evaluation of the method.

Figure 3: Human silhouette extracted

4 KRAWTCHOUK POLYNOMIAL AND MOMENTS

4.1 Krawtchouk Polynomial

The n-th order of Krawtchouk polynomial is based on the hypergeometric function and is defined as:

$$K_n(x;p,N) = \sum_{k=0}^{N} \left(a_{k,n,p} x^k\right) = {}_2F_1\left(-n,-x;-N;\frac{1}{p}\right) \tag{2}$$

where $x,n = 0,1,2,...,N\,et\,N > 0, p \in (0,1)$ and the hypergeometric function defined as:

$${}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \left(\frac{(a)_k (b)_k z^k}{(c)_k} \frac{z^k}{k!}\right) \tag{3}$$

$$(a)_k = a(a+1)...(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)} \tag{4}$$

Equation (4) is the Pochhammer symbol.
The set of (N+1) Krawtchouk polynomial forms the complete set of discrete basis functions with weight function

$$w(x;p,N) = \left(\begin{array}{c} N \\ x \end{array}\right) p^x (1-p)^{N-x} \tag{5}$$

and satisfies the orthogonality condition :

$$\sum_{k=0}^{N} w(x;p,N) K_n(x;p,N) K_m(x;p,N) = \rho(n;p,N)\delta_{nm} \tag{6}$$

where $\rho(n;p,N) = (-1)^n \left(\frac{1-p}{p}\right)^n \frac{n!}{(-N)_n}$ and δ_{nm} is the Kronecher function.
In order to eliminate the large variability in the dynamic range, a normalization process is applied. Then, the set of normalized (weighted) Krawtchouk polynomials is defined by (Yap et al., 2003) as:

$$\bar{K}_n(x;p,N) = K_n(x;p,N) \sqrt{\frac{w(x;p,N)}{\rho(n;p,N)}} \tag{7}$$

4.2 Krawtchouk Moment

Krawtchouk moment is firstly used in image analysis by P.T Yap and al.(Yap et al., 2003). Based on the weighted Krawtchouk polynomials, the (n + m) order of Krawtchouk moment for an N x M image with intensity function $f(x,y)$ is defined as:

$$Q_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \bar{K}_n(x;p1,N-1)\, \bar{K}_m(y;p2,M-1)\, f(x,y) \tag{8}$$

The parameter p1 and p2 can be viewed as a translation factor. Indeed, if p = 0.5 + Δp, the weighted Krawtchouk polynomials are shifted by about NΔp. The direction of shifting relies on the sign of Δp, with the polynomials shifting along + x direction when Δp is positive and vice versa. This property allows to extract the local properties of an images. For software like Matlab, there is a matrix form of the Krawtchouk moment. In matrix form, it is defined as:

$$Q = K_2 A K_1^T \tag{9}$$

where $Q = \{Q_{ji}\}_{i,j=0}^{i,j=N-1}$,
$K_v = \{\bar{K}_i(j;pv,N-1)\}_{i,j=0}^{i,j=N-1}$ *and*
$A = \{f(j,i)\}_{i,j=0}^{i,j=N-1}$

4.3 Feature extraction

For a given image of human silhouette, we used Krawtchouk moment to describe the shape of the human belong to the image. That means to calculate the characteristic vector of the image with different values of the moment. Thanks to the ability of Krawtchouk moment to extract feature of specific regions of the image, we divided each silhouette in two parts (up and bottom) (fig. 4) with the parameter $p1 = 0.5,\ p2 = 0.1$ (for the up) and $p1 = 0.5,\ p2 = 0.95$

(for the bottom). Then, we calculated two characteristic vectors and combined them to get one vector descriptor. Each human silhouette extracted is converted to a common space 128 x 48 to get the invariance to translation and scale. For rotation invariance, we supposed that the vertical is preserved.

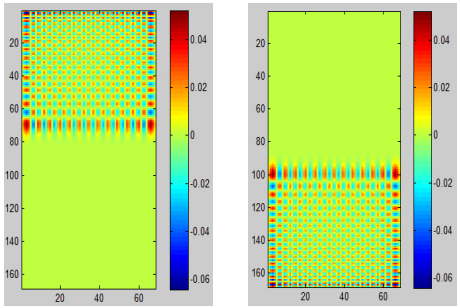


Figure 4: Krawtchouk polynomial for up and bottom

According to some related works, we chose to calculate Krawtchouk moment with parameter (m = n). In order to find the best value of n, we used a database with 600 simulated silhouettes and done cross validation over all. The fig 5 show that from order (n = m = 24), we got a stable and best accuracy for pose recognition. So, the final feature vector has 48 dimensions.

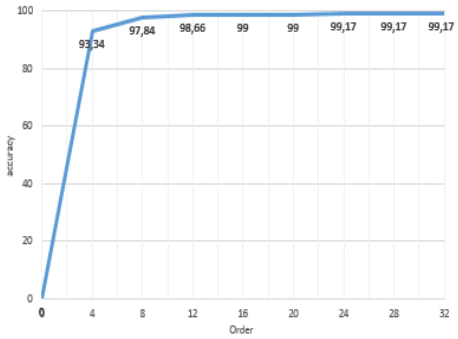


Figure 5: Accuracy of cross validation with differents value of n

5 EXPERIMENTS

In section 3.1 we have shown that for each 2D image of silhouette of the database, we store both the silhouette vector descriptors and the associated 3D skeleton composed of 19 joints. Then, for a test image with extracted silhouette, similarity is computed between the processed vector of descriptors and database descriptors using the Euclidian distance. As a result we extract the corresponding silhouette in the database and its joints 3D poses. Note that the approach does not only give the more suitable silhouette but gives in a classified way the N^{th} most probable silhouettes. In order to evaluate the given result, we used the simulation. By knowing the real skeleton of the test image, we can process the reprojection error of the estimated 3D joints. According to experimental result, when the mean error is less than 5 pixels, the pose of the result is considered similar to the pose of the request silhouette. For this empiric threshold, the difference between two silhouettes is hardly visible for a human.

5.1 Representativity and descriptor robustness to noise

Silouette extraction is still an active reseach field. It is well known that extraction is subjected to noise. First point was to check our descriptors robustness to noise. For this, we conducted experiments with two databases of simulated data for a human avatar with different morphology and different actions. The first database contains 2925 training data with Gaussian noise around the contour of the shape and the second database contains 608 unlearning data. The aims of this experience is to evaluate the capacity of shape descriptors to encode various shapes with different value of the standard deviation of Gaussian noise. Considering $x_0 = [0,0]$ the center of the silhouette, let $x_i = [\rho_i, \theta_i]$ the polar coordinates of a contour point. The noise $\Delta\sigma$ is applied on ρ_i. $\Delta\sigma \hookrightarrow \mathcal{N}(0, std)$ with $std = \{0,1,2,3\}$. Example of noised silhouettes are presented on figure 6.

Figure 6: Noised silhouettes with $\Delta\sigma \hookrightarrow \mathcal{N}(0, std)$ and $std = \{1,2,3\}$

The aim of this experience is to see if the shape descriptor can perfectly encode a silhouette and make the difference between closed postures. The silhouette in the database can be very similar because we extracted it from a video of the motion, so two near frames provide a very similar silhouette. For $std = 0$, we have the original silhouette and for $std > 0$, the Gaussian white noise is added on the silhouette. Figure 7 shows that the more the std increases, the more

the recognition accuracy decreases. For this test we used a training data set composed of 2925 and a testing test of 608 silouhettes. For a single neighbour ($N = 1$), with $std = \{0,1,2,3\}$, the recognition rate is respectively $RR = \{98.81, 96.43, 74.6, 44.84\}$. But, if we augment the number of N assumption returned by the program, the recognition rate grows up quickly. For $N = 7$ and $std = \{0,1,2,3\}$, the RR are $\{100, 100, 96.43, 73.41\}$. Considering that the silhouettes are very similar and the noise very strong, the method gives very good results. For the rest of the article we will consider $N = 7$ first silhouettes given by the matcher.

In order to estimate the 3D extracted skeleton, we use the same request silhouette as for previous experiment. For each extracted silhouette, we process the reprojection error and evaluate the accuracy for different value of N. The Figure 8 shows skeletons estimations from a single monocular image. For this result, the reprojection error of the first image (human walking) is **2.4739 px** and that of the second image (human in cross position) is **1.2614 px**. This means error show that the retrieval pose is near to the original pose. Note that, in the database, there a no avatar with the similar appaerance, so this error is reasonable.

The images that we used as request in fig 8 are simulate image. So, we got a perfect result with low reprojection error.

The result of this 3D skeleton extraction of fig 9 is perfect because this pose is unique and easy to find. The silhouette extraction is too easy because we have a static and uniform background.

In fig 10, we used a real world image extracted from a walking action video. The pose that we choose is similar but not exact with pose in the learning database. So, we don't expect to get a very simular 3D pose as result, but some pose similar. The result show a good result in term of the shape of the pose. But, confusion was made between right and left foot and arm.

In order to evaluate the stability and therobustness of our approach, we considered the successive detections during a complete movie of the movement. Note that there is no use of the time line and each frame is processed independently. Figure 11 (a) shows the tracking results of four human's joints during the execution of the climbing motion. The red curve show the real position over time and the green curve show the estimate position over time. We can note that the shape of different curves is the same. That means that the successives detections are stable in time and that our shape descriptor is reliable. We can note that there is an offset due to shape scaling. The means error over the motion execution is 1.9765 *px*. Figure 11 shows that the shape of the curve changes as a fonction of the motion. The means error obtained form the jump motion is 1.9892 *px*. This discrimination factor confirmed that the 3D poses can be used for actions classification in a video.

5.2 Application to action recognition on real data

We used the same shape descriptor for human action classification in video, with the public Weizmann database (see Figure 12). As we do not use temporal information, our method consists in matching each frame to an action class and took the class with the highest associated rate as the class action.

The database is a collection of 90 low-resolution (180 x 144, deinterlaced 50 fps) video sequences showing 9 different people, each performing 10 natural actions: run, walk, skip, jumping-jack, jump, gallop-sideways, wave-two-hands, waveone- hand, or bend. On Weizmann data base, we made a cross validation with the different movements and with the different phenotypes. In each case and for each frame, we apply our shape matching method to each frame. As the resulting silhouette from the database belongs to a specific movement class we simply count the number of occurencies. The more represented class is then considered as the detected movement.

Based on this very simple workflow, we got 71.66% of good action classification. The confusion matrix is shown on the fig.13. Of course, this accuracy rate is lower than the recent accuracy obtained on the same database (Blank 99.64% (Blank et al., 2005) and Gorelick 97.83% (Gorelick et al., 2007)). But both of these approach used space-times cubes to analyse the motion while we do not consider yet the temporal correlation between successives frame.

According to Gorelick et al.: *many successive frames from the first action (about run) may exhibit high spatial similarity to the successive frames from the second one. Ignoring the dynamics within the frames might lead to confusion between the two actions*. As the approach does not take into account time dimension, frame to frame comparison leads to misclassification for these very similar frame to frame actions: run, skip and jump.

In future work, we will use our proposed approach combined with the multi-hypothesis tracking techniques (with N neighboors) to improve the accuracy of action classification. By this way, we will take into account the temporal information and the dynamic of the action.

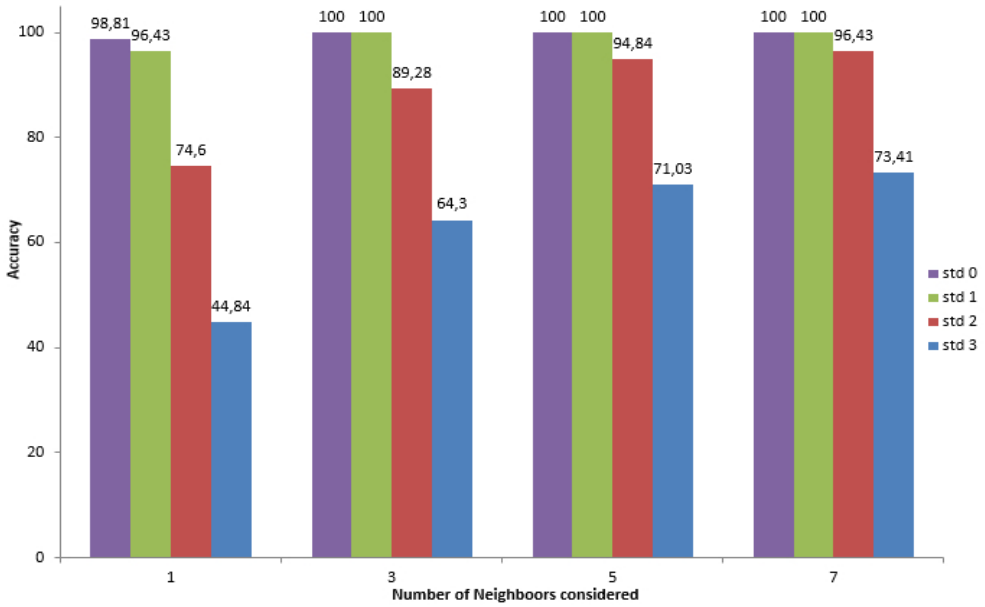


Figure 7: Histogramm of accuracy: colors represent the noise amplitude resp. $\{0,1,2,3\}$ pixels. The abscisses represent the number N of neighboors considered $\{1,3,5,7\}$.

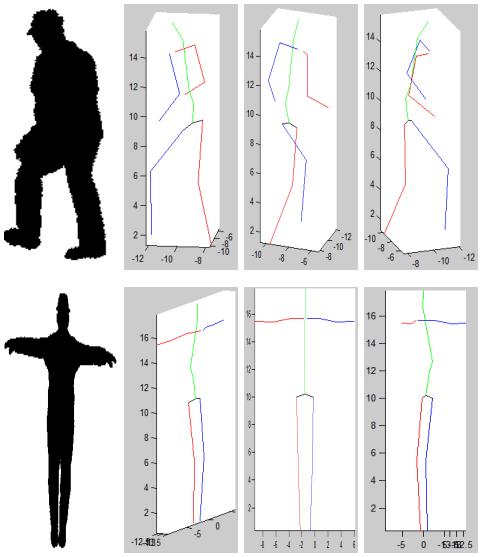


Figure 8: 3D pose estimation result: Left, the resquest silhouette and from left to right, the 3D estimated skeleton from various viewpoints

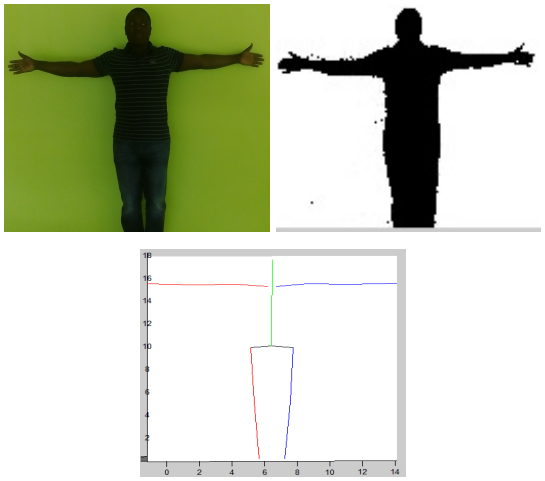


Figure 9: Real world data 1

6 CONCLUSIONS

In this paper, we presented a new approach for 3D human pose estimation and action classification in video. The learning database is easily generated thanks to open source softwares which allow any human pose simulaion. The proposed posture recognition method is based on the geometric Krawtchouck moment and gives promising results. Both application to 3D pose estimation and action classification have been presented. In our work, we tested different moment order and selected the best suitable for our approach. We compared our approach with some related work in action classification and we concluded that this approach can be improved by using multi-hypothesis tracking during action identification and classification. In future work, we will use a combination of local and global shape descriptor for improving the pose estimation, and use the estimated poses to construct an action model for activity classification.

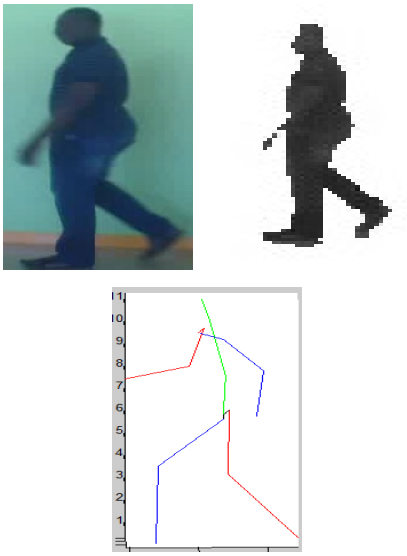
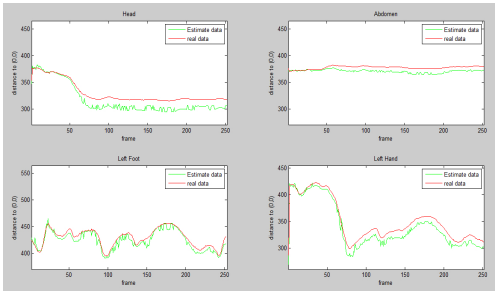
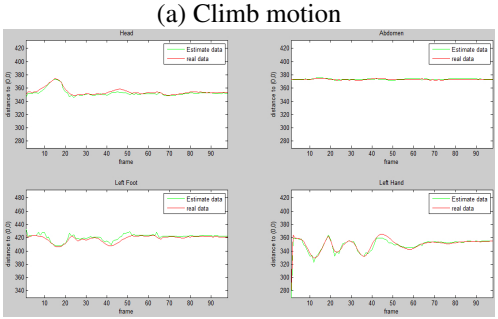


Figure 10: Real world data 2



(a) Climb motion



(b) Jump motion

Figure 11: Tracking result

7 ACKNOWLEDGE

This work is part of LUMINEUX project, supported by the Regional Centre-Val de Loire (France). The authors would like to acknowledge the Conseil Regional of Centre-Val de Loire for its support.

Figure 12: Some images of Weizmann database

	walk	run	skip	jack	jump	pjump	side	wave	wave2	bend
walk	100	0	0	0	0	0	0	0	0	0
run	0	50	50	0	0	0	0	0	0	0
skip	16,7	16,7	50	0	16,7	0	0	0	0	0
jack	0	0	0	83,3	0	0	0	0	16,7	0
jump	16,7	16,7	33,3	0	16,7	0	0	167	0	16,7
pjump	0	0	0	0	0	83,3	16,7	0	0	0
side	0	0	0	0	16,7	16,7	66,7	0	0	0
wave1	16,7	0	0	0	0	16,7	0	83,3	0	0
wave2	16,7	0	0	0	0	0	0	0	83,3	0
bend	0	0	0	0	0	0	0	0	0	100

Figure 13: Confusion matrix

REFERENCES

Agarwal, A. and Triggs, B. (2006). Recovering 3d human pose from monocular images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(1):44–58.

Aggarwal, J. and Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440.

Andriluka, M., Roth, S., and Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 623–630. IEEE.

Baumberg, A. and Hogg, D. (1994). *Learning flexible models from image sequences*. Springer.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402.

Bourdev, L. and Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1365–1372. IEEE.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *In: IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893.

de La Gorce, M., Fleet, D., and Paragios, N. (2011). Model-based 3d hand pose estimation from monocular video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(9):1793–1805.

Gavrila, D. M. and Davis, L. S. (1996). 3-d model-based tracking of humans in action: a multi-view approach. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pages 73–80. IEEE.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *In ICCV*, pages 1395–1402.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

Guo, K., Ishwar, P., and Konrad, J. (2009). Action recognition in video by covariance matching of silhouette tunnels. In *In: XXII Brazilian Symposium on Computer Graphics and Image Processing*, pages 299–306.

Hiyadi, H., Ababsa, F., Bouyakhf, E. H., Regragui, F., and Montagne, C. (2015). Reconnaissance 3d des gestes pour l'interaction naturelle homme robot. In *Journées francophones des jeunes chercheurs en vision par ordinateur*.

Hogg, D. (1983). Model-based vision: a program to see a walking person. *Image and Vision computing*, 1(1):5–20.

Jiang, H. (2010). 3d human pose reconstruction using millions of exemplars. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1674–1677.

Lee, M. W. and Nevatia, R. (2009). Human pose tracking in monocular sequence using multilevel structured models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):27–38.

Maji, S., Bourdev, L., and Malik, J. (2011). Action recognition from a distributed representation of pose and appearance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3177–3184. IEEE.

Mori, G. and Malik, J. (2002). Estimating human body configurations using shape context matching. In *Computer VisionECCV 2002*, pages 666–680. Springer.

O'Rourke, J., Badler, N., et al. (1980). Model-based image analysis of human motion using constraint propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):522–536.

P. Dollar, S. B. and Perona, P. (2010). The fastest pedestrian detector in the west. In *In: Proceedings of the British Machine Vision Conference*, pages 1–11.

Rehg, J. M. and Kanade, T. (1994). Visual tracking of high dof articulated structures: an application to human hand tracking. In *Computer VisionECCV'94*, pages 35–46. Springer.

Taylor, C. (2000). Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 677–684 vol.1.

Valmadre, J. and Lucey, S. (2010). Deterministic 3d human pose estimation using rigid structure. In *Computer Vision–ECCV 2010*, pages 467–480. Springer.

Wang, C., Wang, Y., and Yuille, A. (2013). An approach to pose-based action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 915–922.

Wang, L., Wang, Y., and Gao, W. (2011). Mining layered grammar rules for action recognition. *International Journal of Computer Vision*, 93(2):162–182.

Wei, X. K. and Chai, J. (2009). Modeling 3d human poses from uncalibrated monocular images. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1873–1880. IEEE.

Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):780–785.

Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE.

Yap, P.-T., Paramesran, R., and Ong, S.-H. (2003). Image analysis by krawtchouk moments. *Image Processing, IEEE Transactions on*, 12(11):1367–1377.