A Very Simple Framework for 3D Human Poses Estimation Using a Single 2D Image: Comparison of Geometric Moments Descriptors

Dieudonne Fabrice ATREVI[a,*], Damien VIVET[b,**], Florent DUCULTY[a], Bruno EMILE[a]

[a] *Univ of Orleans, INSA Centre Val de Loire, PRISME EA 4229, F45072, Orleans, France*
[b] *University of Toulouse, ISAE-Supaero / DEOS, Toulouse, France*

Abstract

In this paper, we propose a framework in order to automatically extract the 3D pose of an individual from a single silhouette image obtained with a classical low-cost camera without any depth information. By pose, we mean the configuration of human bones in order to reconstruct a 3D skeleton representing the 3D posture of the detected human. Our approach combines prior learned correspondences between silhouettes and skeletons extracted from simulated 3D human models publicly available on the internet. The main advantages of such approach are that silhouettes can be very easily extracted from video, and 3D human models can be animated using motion capture data in order to quickly build any movement training data. In order to match detected silhouettes with simulated silhouettes, we compared geometrics invariants moments. According to our results, we show that the proposed method provides very promising results with a very low time processing.

Keywords: 3D Pose estimation, 3D modeling, Skeleton extraction, Shape descriptor, Geometric moment, Krawtchouk moment, Zernike moment, Hu moment, Hahn moment

[*] Principal corresponding author
[**] Corresponding author
Email addresses: fabrice.atrevi@univ-orleans.fr (Dieudonne Fabrice ATREVI), damien.vivet@isae.fr (Damien VIVET), florent.duculty@univ-orleans.fr (Florent DUCULTY), bruno.emile@univ-orleans.fr (Bruno EMILE)

1. Introduction

One of the main objectives of smart environments is to enhance the quality of life of the inhabitants. For this purpose, monitoring systems have to understand the needs and intention of a human in order to adapt the environment, for example in term of heating or lighting. Moreover, by monitoring the movement of a user, these systems could also be able to alert the user or ask for help in case of danger or if the movement could lead to an injury like a fall for example. Then, Human action recognition systems have a lot of possible applications in surveillance, pedestrian tracking and Human Machine Interaction. Human pose estimation is a key step to action recognition.

A human action is often represented as a succession of human poses [1]. 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 [2] 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 [3] or by rotation angles [4]. 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 even if tracking approaches can solve this ambiguty. 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 framework for the extraction of 3D skeleton pose assumptions from a single 2D image provided by a low cost webcam. Our approach focuses uniquely on the silhouette shape recognition. A silhouette database is constructed from 3D human pose and action simulator and is used in order to match the nearest silhouette and as a result possible 3D human pose. Section 2 presents the state of the art in the field of human pose estimation. Section 3 explains the methodology we applied in order to estimate the human pose from a single silhouette but also the 3D simulator used to build our training database. Section 4 provides the mathematical description of the geometrics moments (and their parameters) used and compared for this application. Finally, section 5 presents the results obtained by the approach on both our simulated and real database.

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 [5] and Hogg [6] in the eighties. Since last decades, scientists proposed many approaches. We can categorize these approaches into two main categories.

Most of the approaches use a 3D model or 3D detection for estimating the pose of a subject and for action classification. Bourdev and Malik [7] estimated the human pose from key points. They used an annotated dataset of human with 3D joins informations inferred using anthropometric constraints for human action classification. Wei and Chai [9] proposed an approach for solving the non-rigid structure from motion problem specifically for bodies. They claimed that with a minimum of five frames with 2D point correspondences, their approach is able to estimate bone lengths, camera scale and articulated pose. In [8], Valmadre and Lucey demonstrated that this assumption from Wei and Chai is false and this approach is only valid for rigid substructures of the human body (e.g. torso) rather than the entire bodys non-rigid structure. They introduced a deterministic solution to the problem of estimating camera scale and bone lengths for the bodys rigid torso. Recently, depth camera such as the Microsoft Kinect camera has been intensively used in tracking 3D human posture [10],

[11]. Its advantage is that it can track 3D human posture without requiring the user to wear any special equipment. The use of captured depth image allows extracting depth-based edge and ridge data used to track human body parts [12]. However, unsupervised approaches using depth sensor require a complex algorithm to analyse the scene. Of course, run-time detection of a complex model is not always accurate and activity recognition is degraded. In the same way, the use of the skeleton extracted by Kinect for action recognition suffers for eroneous joint recognition in case of occlusions resulting in noisy skeletons [13]. In case the depth sensor data are considered reliable, motion analysis algorithms do not work well with Kinect [14]. Very recently, Ho *et al.* [15] propose new methods to take into account the sensor errors and to improve action recognition in a smart environment using depth sensor. All of these approaches need multiple sensors or specific devices such as time of flight or active camera for acquiring 3D information. The anatomical models used also, need a very good parametrization to be usefull. This category of methods is not suitable for our purpose. We want to estimate the 3D postures from monocular image without any prior depth informations about the person and in complete uncontrolled environment using one camera.

The second category of approaches, to which our proposed method belongs, used 2D models trained from various images. Indeed, identifying human posture with traditional 2D video cameras can be performed using computer vision techniques [16]. Nevertheless, recovering a 3D human pose from a single 2D image is an ill-posed problem because multple body configurations may have a similar silhouette aspect. Moreover, in realistic situations, body silhouette cannot be accurately detected because of occlusions or wrong background segmentation. Wren *et al.* [17] tracked people and interpreted their behaviour by using a multiclass statistical model of colour and shape to obtain a 2D representation of head and hand. Gorelick *et al.* [18] 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. Agarwal and Triggs [19] used the shape context in their research on human pose estimation. Gorce *et al.* [20] 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 [2] 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. So, learning is required and such algorithms need complex and expensive systems to get the training data set with the ground truth. For this purpose, motion capture data have to be collected for different motions and behaviors. Such technique has been widely used [21] and multiple mocap files are publically available on the internet. Our method is based on a very simple silhouette extraction and description. It could be compared to Shape-from-silhouette approach [22] but in our case, we use a single image to find the 3D pose. We also show that our method is robust in case of noised extracted silhouette.

Another difference from state of the art approach is that for generating the learning database, we proposed to use software applications from the open source community associated to available motion capture files. Those softwares makes realistic simulations of various human poses and action possible. Moreover, movement can be easily adapted in order to generate new pose and actions. This work is an extension of [23] and shows that (1) using only 3D simulations for learning, (2) without complex machine learning algorithm and (3) with a very 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 3D 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.

3.1. General workflow

As mentionned above, the proposed 3D pose estimation approach is composed of 4 parts, from the human detection and silhouette extraction to the pose estimation and validation. The entire workflow is presented Fig. 1. In this section, we summarized each of forth step.

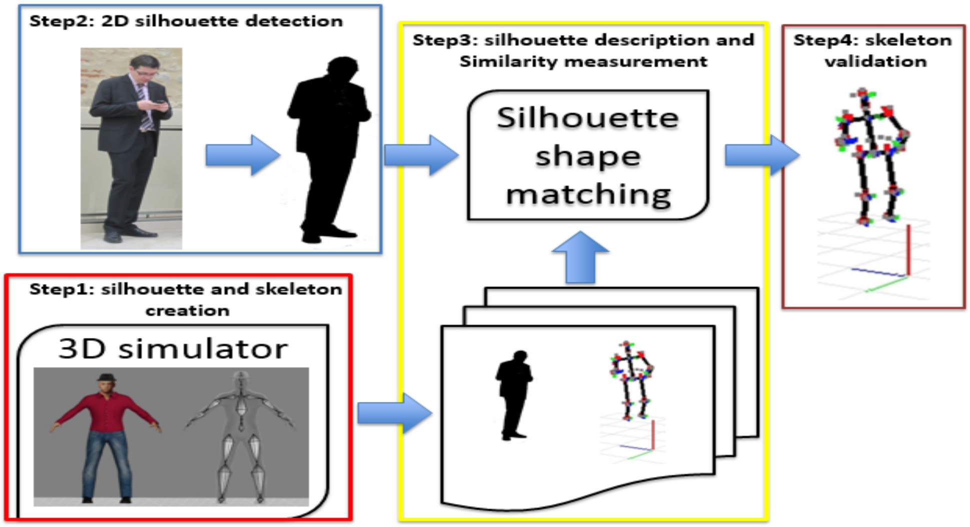

Step2: 2D silhouette detection
Step3: silhouette description and Similarity measurement
Silhouette shape matching
Step4: skeleton validation
Step1: silhouette and skeleton creation
3D simulator

Figure 1: Human pose estimation methodology.

(1) 2D silhouette and 3D skeleton database is built thanks to open source 3D software Blender (see section 3.2 for more details on the database construction). Such database is composed of human silhouettes and its corresponding 3D skeletons for different kind of postures, extracted from multiples actions dataset, like walk, run, climb. So, for a requested silhouette, it'll be possible to find an approximate silhouette in the database and then the corresponding 3D skeleton.

(2) 2D silhouette detection is a well-studied field in machine learning and computer vision. For this purpose, we used classical real-time approach human detector proposed by Dollar et al. [24] based on multiscale HOG to focus the region of interest associated to a statistical background substraction. Once the human silhouette is detected, we converted it to a 48 x 128 pixels image for solving the translation and scale 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. Since silhouettes can be consider as shape, different shapes descriptors can be use to describe them. Numerous shape descriptors have been proposed in the literature and can be categorized as contour-based and region-based descriptors. The first category, describe the distribution of the boundary information of the shape and by the way, ignoring the interior content which can be important for some shapes. In opposite, the region-based descriptors exploit both boundary and internal content for the shape description. In this last category, one group of descriptor is the geometric moments which have been very popular since their introduction in the 60s. For our application, we use four moments in the silhouette description task (See section 4 for details). Based on those descriptors, a feature vector is computed for each silhouette in the database and the similarity between characteristic vector is measured with the Euclidean distance given by :

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

where z^r et z^t is respectively the characteristic vector of request silhouette and the t^{th} silhouette in the database. We choose to present this simple distance as other metric distance tested (cosinus distance, correlation distance, Bhattacharyya distance, L2...) did not improve significantly the results. The search of corresponding silhouette in the database is linear and then have O(n) time complexity. Others search strategy can be investigate to improve the complexity.

(4) Skeleton scaling and validation. For each silhouette we retrieve the n nearest 3D skeletons in the dabase obtained at the end of the last step. In order to get the final posture, different technics can be use. One can take the nearest silhouette in the database and then its corresponding skeleton. One can also, consider n nearest silhouettes and compute a mean skeleton by using the n corresponding skeleton. The final skeleton is scaled to the current silhouette size by geometric transformation. For validation purpose, we use ground truth simulated database to validate the approach. The confidence score is processed by measuring the 3D/2D reprojection error of predicted joints on the silhouette and an empirical fixed threshold is used to decide which result is good pose estimation.

3.2. Construction of the 2D/3D matching database

One of the novelty of our approach is to generate easly big database of human 3D skeleton and its corresponding silhouette thanks the advance in computer graphic and virtual modelisation. By the way, it'll be possible to generate different silhouettes (different size, corpulence, camera view point) from real world motion information. This approach contribut to reduce the time and financial relate to the generation of such database. Indeed, for one motion information, took from motion capture dataset, we can generate many silhouettes.

3.2.1. 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. In these Softwares, we simulate different human avatars with different morphologies and clothes and animate them with different realistic

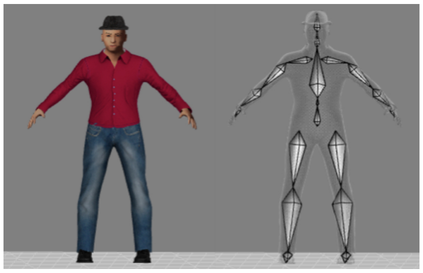


Figure 2: 3D simulated avatar and its associated skeleton

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

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

motions taken from the CMU motion capture database[3]. These motion capture files have been generated using numerous wearable markers and provide a very good precision of the reconstructed 3D motion. For our work, we choose to reduce the number of joints to be estimated to 19. Let's note that illumination conditions and point of view of the camera can be easily modified with this software, in order to generate required database. For example, a configuration using four virtual cameras positioned on a virtual sphere centered on the subject is presented in Fig. 3. The position of the camera on the sphere, camera intrinsic parameters, and sphere radius can be adjusted in order to match the type and the pose of the camera used in the real application (video monitoring, smart home, behaviour analysis, etc.). The number of camera can be more, it depend on the intrinsic parameters of the camera and the scene configuration.

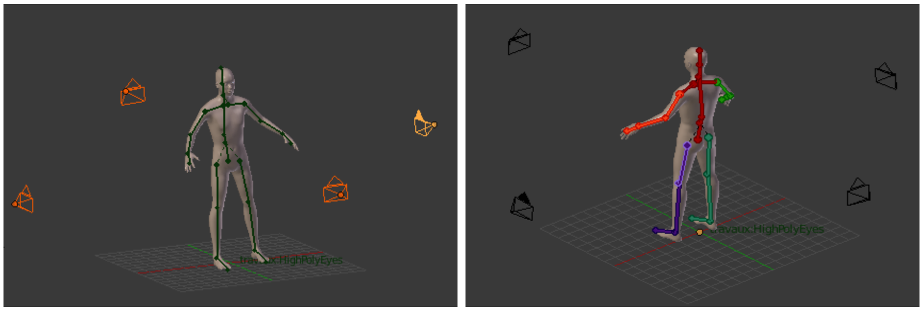


Figure 3: Blender software view for 4 cameras based database generation. Camera are positionned on a virtual 3D sphere centered on the subject.

3.2.2. Database construction

Once the avatars are generate with the software makehuman, we import them into the 3D graphics software "blender". We positioned on a hemisphere some virtual cameras looking at the subject. Thanks to the motion capture files, we animated these avatars. Then, for each movement of the avatar, we can record both: 2D image and silhouette (give by cameras), 3D camera poses and 3D

[3] http://mocap.cs.cmu.edu/

joints and bones poses (in world coordinate). As a result for each subject's pose, we can collect the detected silhouette related to its 3D skeleton (containing 19 bones). For the purpose of this paper, we recorded in four subjects with different phenotypes and playing four different animations: walk cycle, basket action, jump, and climb. Geometric transformation can be done on 3D skeleton in order to convert them from 3D world coordinate to 2D image coordinate(3D/2D projection). This transformation is used to compute the reprojection error for quantitave evaluation purpose. For each silhouette in the database, we then extract the feature vector with the shape descriptors presented in section 4.

4. Shape descriptors

In order to describe the silhouettes, we have processed and compared four well-known shape descriptors based on invariant and orthogonal moments. Such moments have been proved to be a good region-based descriptor in a multitude of machine learning application and for content-based image retrieval [25][26]. As we assume that the 3D pose is directly linked to the 2D shape of the silouette, the main objective is to reprensent the shape accurately. An ideal descriptor for our pose recovery problem would be able to distinguish between different body poses while being able to generalize over body dimensions, variations in viewpoint and local boundary noise.

In case of low orders of geomentric moments, it is possible to interpret their meaning. For instance:

- m_{00} represent the mass of image (for binary image, it's an area of the object);
- In case that the image is considered a probability density function ($m_{00} = 1$), m_{01} and m_{10} are the mean value;
- In case of zero means, m_{20} and m_{02} are variances of horizontal and vertical projections and m_{11} is a covariance between them;
- m_{01}/m_{00} and m_{10}/m_{00} define the gravity or centroid of the image.

In this section, we'll describe the four different shape descriptors that we compared during our experiments : Hu, Zernike, Krawtchouk and Hahn geometrics moments.

4.1. Hu geometric moments

The first geometric moments used in computer vision was introduced by Hu[27]. The general formulation of two-dimensional $(p+q)^{th}$ order moment for an image is defined as:

$$M_{pq} = \sum_x \sum_y x^p y^q f(x,y) \tag{2}$$

where p and q are integers: $p, q \in \{0, 1, 2, ..., N\}$ with N defining the maximal order.

To normalize for translation into the image plane, Hu shown that the central moment based on the image centroids of coordinate $(\bar{x}, \bar{y})$ should be used and can be express as:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x,y) \tag{3}$$

where

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}} \tag{4}$$

In shape recognition application field, Hu introduced seven (07) invariants moments based on the normalized central moments. The first 6 descriptors encode a shape with invariance to translation, scale and rotation. The 7th descriptor ensures skew invariance, which, we hope, will enable us to distinguish between mirrored images.

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned}$$

The computation of Hu invariants moments is very simple but present several drawbacks even invariant to rotation, scaling and translation[28]:

- Information redundancy: the invariants moment have high degree of information redundancy due to the non orthogonality of the basis.
- Noise sensitvity: The higher order moments are too sensitive to noise.
- Large variation in the dynamic range of values: Large variation in the dynamic range of values is observed for different orders, since the basis involves power of p and q. The consequence of that is numerical instability when the image size is large.

For our application, for each silhouette of the database, we computed the feature vector composed of the 7 invariants moments:

$$\mathcal{F}_{Hu} = [\phi_1 \ldots \phi_7]^T$$

$\mathcal{F}_{Hu}$ will then be used as a descriptor in the classification step in order to get the nearest matching silhouette and by the way, the nearest pose.

To overcome the limitations associated with invariant and geometric moments, Teague [29] suggested the use of continuous orthogonal moments. He introduced two different continuous-orthogonal moments, Zernike and Legendre moments, based on the orthogonal Zernike and Legendre polynomials, respectively. Recent work introduced news orthogonal moements for shape analysis and image reconstruction. After used Hu's moments as reference, we will show details about the set of 3 others orthogonal moments used in our work: Krawtchouk, Hahn and Zernike moments.

4.2. Krawtchouk shape descriptor

4.2.1. Krawtchouk Polynomial and moments

Krawtchouk moments are firstly introduced in image analysis by P.T Yap *et al.* [30]. These moments are computed using the discrete classical Krawtchouk polynimials. The n^{th} order of Krawtchouk polynomials are based on the hypergeometric function and is defined as:

$$K_n(x;p,N) = \sum_{k=0}^{N} \left(a_{k,n,p}x^k\right) = \, _2F_1\left(-n,-x;-N;\frac{1}{p}\right) \tag{5}$$

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

$$_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \left(\frac{(a)_k (b)_k z^k}{(c)_k} \frac{z^k}{k!}\right) \tag{6}$$

$$(a)_k = a(a+1)...(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)} \tag{7}$$

Equation (7) is the Pochhammer symbol.

The set of (N+1) Krawtchouk polynomial forms the complete set of discrete basis functions with the weight functions:

$$w(x;p,N) = \left(\begin{array}{c} N \\ x \end{array}\right) p^x (1-p)^{N-x} \tag{8}$$

and satisfies the orthogonality condition:

$$\sum_{x=0}^{N} w(x;p,N) K_n(x;p,N) K_m(x;p,N) = \rho(n;p,N)\delta_{nm} \tag{9}$$

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 with:

$$\delta_{nm} = \begin{cases} 1 & n = m \\ 0 & otherwise \end{cases}$$

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.[30] as:

$$\bar{K}_n(x;p,N) = K_n(x;p,N)\sqrt{\frac{w(x;p,N)}{\rho(n;p,N)}} \tag{10}$$

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:

$$Q_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \bar{K}_n \left(x; p_1, N-1\right) \bar{K}_m \left(y; p_2, M-1\right) f\left(x, y\right) \tag{11}$$

According to Yap et al. [30], "the lower order weighted Krawtchouk polynomials have relatively high spatial frequency components. This, together with the fact that Krawtchouk polynomials are polynomials of discrete variable, contributed to the ability of the Krawtchouk moments to represent edges (sharp changes of the image intensity values) more effectively". In case of silhoutte (binary image), this can capture the information of ahape represented by the edges. Combining that information with the parameters p_1 and p_2, which can be viewed as a translation factor, it's possible to extract local information of edges of the silhouette. Indeed, if $p = 0.5 + \Delta p$, the weighted Krawtchouk polynomials are shifted by about $N\Delta p$. The direction of shifting relies on the sign of Δp, with the polynomials shifting along the positive x direction when Δp is positive and vice versa. we'll descibe in next subsection, how shape informations are extracted via Krawtchouk moments.

4.2.2. Feature extraction

For a given image of a human, the silhouette is projected in Krawtchouk polynomial basis and the moment are extracted to describe the shape of the human. A feature vector of the image is then formed by different orders of moments. Thanks to the ability of Krawtchouk moment to extract feature of specific regions of the image, we divided each silhouette into 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:

$$\mathcal{F}_{Kr} = \left[Q_{nm}^{bottom}, Q_{nm}^{top}\right]^T$$

with $m \in [0:M]$ and $n \in [0:N]$

Each human silhouette extracted is converted to a common space 48 x 128 to

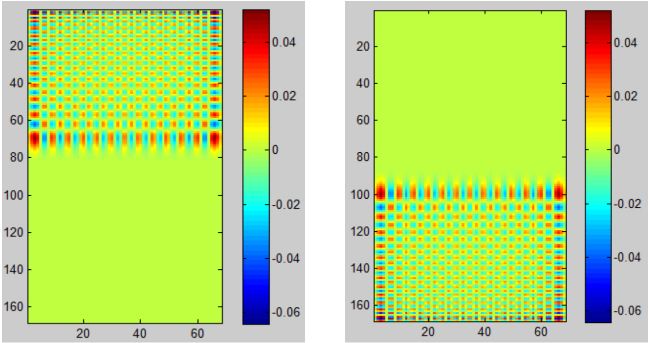


Figure 4: Krawtchouk polynomial for up and bottom

get the invariance to translation and scale. For rotation invariance, we supposed that the verticality of the silhouette is preserved.

According to some related works, we choose to compute Krawtchouk moments with parameter (m = n). In order to find the suitable value of N, we used a database with simulated silhouettes and done cross validation over all. From order (N = M = 24), we got a stable and best accuracy for pose recognition, so, the final feature vector has 48 dimensions.

$$\mathcal{F}_{Kr} = \left[Q^{bottom}_{0,0} \ldots Q^{bottom}_{23,23}, Q^{top}_{0,0} \ldots Q^{top}_{23,23}\right]^T$$

P.T Yap et al. [30], further, argued that the lower order of Krawtchouk moments store information of a specific region-of-interest of an image and the higher order moments store information of the rest of the image. The better way to evaluate the powerful of geometric moment is in its capacity to reconstruct an image with less square error. In the same paper, P.T showed that Krawtchouk moments present advantage on some well known orthogonal geometric moment.

4.3. Hahn shape descriptor

4.3.1. Hahn Polynomial and moments

As shown in [31], Hahn moments are a generalization of Krawtchouk and Chebyshev moments. This implies that Hahn moments encompass all their

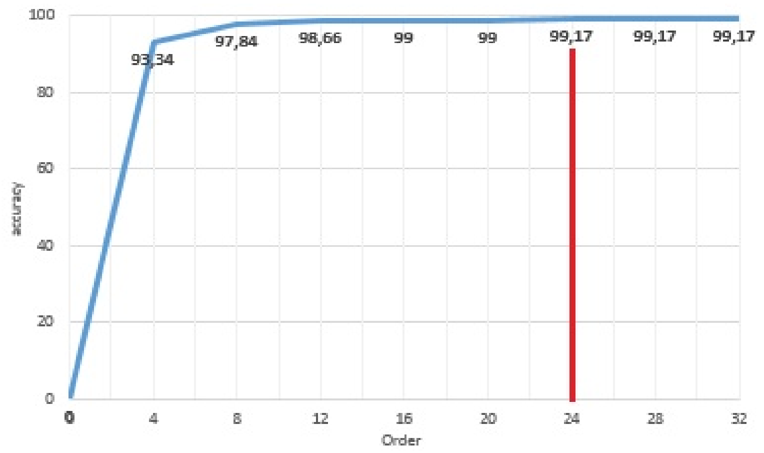


Figure 5: Accuracy of Krawtchouk descriptor with different orders

properties. The aims of using hahn moment in our work is to comapare its result in our framework to the result of Krawtchouck. We expected similar or better result from this moment in spirit to confirm the theorical analysis that Hahn encompass most properties of the Krawtchouk one. The n^{th} order of Hahn polynomial is also based on the hypergeometric function and is defined as:

$$h_n(x;\alpha,\beta,N) = {}_3F_2\left(-n, n+\alpha+\beta+1, -x; \alpha+1, -N; 1\right) \tag{12}$$

where $\alpha > -1$ & $\beta > -1$

The set of (N+1) Hahn polynomial forms the complete set of discrete basis functions with the weight function:

$$w(x;\alpha,\beta,N) = \begin{pmatrix} \alpha+x \\ x \end{pmatrix} \begin{pmatrix} \beta+N-x \\ N-x \end{pmatrix} \tag{13}$$

and satisfies the orthogonality condition:

$$\sum_{x=0}^{N} w(x;\alpha,\beta,N) h_n(x;\alpha,\beta,N) h_m(x;\alpha,\beta,N) = \rho(n;p,N)\delta_{nm} \tag{14}$$

where $\rho(n;\alpha,\beta,N) = \frac{(-1)^n(n+\alpha+\beta+1)^{N+1}(\beta+1)^n n!}{(2n+\alpha+\beta+1)(\alpha+1)^n(-N)^n N!}$ and δ_{nm} is the Kronecher function.

In the same context as krawtchouk discret moment, a normalization process is applied. The computation of Hahn moment is the same as define for the Krawtchouk moment. Then, based of the Hahn polynomials, the Hahn moment can be defined as:

$$M_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \bar{h}_n(x; \alpha 1, \beta 1, N-1) \bar{h}_m(y; \alpha 2, \beta 2, M-1) f\,(x, y) \tag{15}$$

the couple of parameters (α, β) is to control the selection of a specific region in the image and then allow the using of Hahn moment as a local region-based descriptor. In the specific case of (0,0), we have a global descriptor.

4.3.2. Feature extraction

The feature extraction process for Hahn moment is the same for Krawtchouk moment. Then, for a given image, we compute the moment in the specific case for $m = n$. The local feature is also extracted from the image. We found the suitable values of couple (α, β) to cover the differents emphasis region of the silhouette (up and bottom) of the silhouette. We divided the silhouette into two regions because of when Hahn moments are set to be a global descriptor, a larger number of moments are needed [31]. The couple (α, β) can be compute as follow:

$$\alpha_1 = \frac{x_c}{N} t_1 \text{ and } \beta_1 = (1 - \frac{x_c}{N}) t_1 \text{ along x axis} \tag{16}$$

$$\alpha_2 = \frac{y_c}{M} t_2 \text{ and } \beta_2 = (1 - \frac{y_c}{M}) t_2 \text{ along y axis} \tag{17}$$

where (x_c, y_c) are the central points of the emphasis regions and the factor t_1 and t_2 define if the moment is local or global. $t = 0$ set the moment to become global and more t increase, the moment is set to be local. According to [31], t can be set to 20 N to obtain sufficiently close approximation. In our application, we used images of size 48 x 128 and extracted the half up and bottom separately. We set $t_1 = 0$ for global extraction along x axis and $t_2 = 1000$ for local along y

axis. During experiment, we have found the different suitable values of couple. Then, for the top region we have $(\alpha_1, \beta_1) = (0, 0)$ and $(\alpha_2, \beta_2) = (100, 900)$ and for the low region, we have $(\alpha_1, \beta_1) = (0, 0)$ and $(\alpha_2, \beta_2) = (900, 100)$ (see Fig.6).

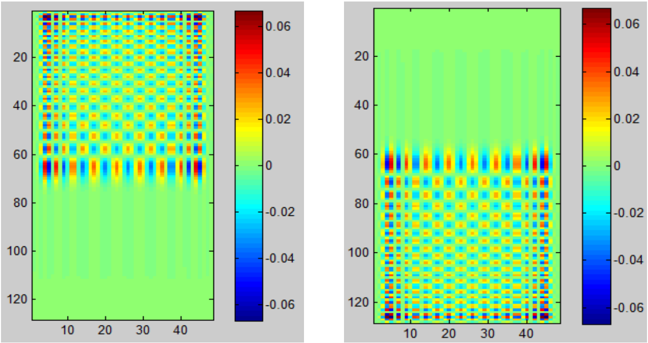


Figure 6: Hahn polynomial for up and bottom

Based on the suitable couple value to extract information of emphasis region, we have tested differents value of the order. We note that Hahn and Krawtchouk moment vary in the same direction. Then, in order to have the same length of the feature vector, we set the order to 23.

As a result the descriptor extracted from Hans moments is given by:

$$\mathcal{F}_{Ha} = \left[M_{0,0}^{bottom} \dots M_{23,23}^{bottom}, M_{0,0}^{top} \dots M_{23,23}^{top}\right]^T$$

4.4. Zernike shape descriptor

4.4.1. Zernike Polynomial and moments

Broadly used in shape recognition through the geometric moment of Zernike, since introduced by Teague[29], the Zernike polynomial formed a complete orthogonal set over the interior of the unit circle. Let's Z_n^m be the Zernike polynomial of order n and repetition m. Z_n^m is defined by:

$$Z_n^m(\rho, \theta) = R_{nm}(\rho) \exp(jm\theta) \tag{18}$$

where n: positive integer or zero;

m: positive or negative integer subject to constraints $n \geq |m|$ and $n - |m|$ is even;

ρ: Radial normalized distance of pixel (x,y) relative to the center of mass of the object;

θ: Azimut angle of pixel (x,y) relative to the center of mass of the object.

The radial polynomial is defined by:

$$\begin{gathered} R_{mn}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s F(n,m,s,r), \\ F(n,m,s,r) = \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!}\rho^{n-2s} \end{gathered} \tag{19}$$

$R_{n,-m}(\rho) = R_{n,m}(\rho)$ and all polynomial are subject to the orthogonality condition:

$\int\int_{x^2+y^2\leq 1}[V_{nm}(x,y)]^* V_{pq}(x,y)dxdy = \frac{\pi}{n+1}.\delta_{np}\delta_{mq}$

with δ_{ab} is the Kronecher function

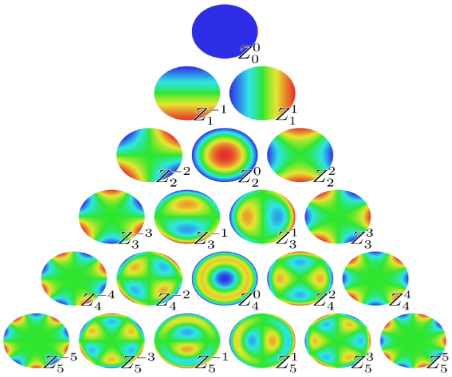


Figure 7: Zernike polynomial of unit circle for different order and repetition [32]

The 2D moment of Zernike are constructed by using the set of polynominal combined with the function intensity of the images. Let's A_{nm} be the Zernike moment of order n and repetition m. A_{nm} is defined by:

$$A_{nm} = \frac{n+1}{\pi}\int\int_{x^2+y^2\leq 1} f(x,y)V^*_{nm}(\rho,\theta)dxdy \tag{20}$$

In digital domain, A_{nm} is computed as:

$$A_{nm} = \frac{n+1}{\pi} \sum_{x} \sum_{y} f(x, y)[V^{*}_{nm}(\rho, \theta)] \tag{21}$$

where $x^2 + y^2 \leq 1$ and V^{*}_{nm} is the complex conjugate of the polynomial.

Zenike moments have the advantage of robustness to noise and minor variations in shape, invariant to rotation and have minimum information redundancy. However, its computation present some probleme such as coordinate space normalization, numerical approximation of continuous integrals and computational complexity [33].

4.4.2. Feature Extraction

For feature extraction of an image, we followed the same process for Krawtchouk descriptor. For a given order, we computed all possible moments with order less than the given order. That mean, for an order equal to n, we computed all possible moment for order from 0 to n. This way for extracting the feature vector allow us to get much information that can make the difference between two similar images. Many prior works try to find the best of order which is suitable to effectively characterize the shape. In their works, [34] showed that the suitable order for Zernike moment is in the range from 7 to 12. They determined this range by computing the reconstruction error of image for differents values of order. In our study, we tried to find the suitable order for our specific case. We did the same process as in Krawtchouk moment case and got an excellent accuracy from the order 8 to 16 and can confirm this study.

5. Experimental studies

In section 3.2 we have shown that for each 2D image of a silhouette in the database, we store both the feature vector and the associated 3D skeleton composed of 19 joints. Then, for each test image with its extracted silhouette, the similarity is computed between the processed feature vector and the stored features vectors in the database. For similarity computation, we compared different

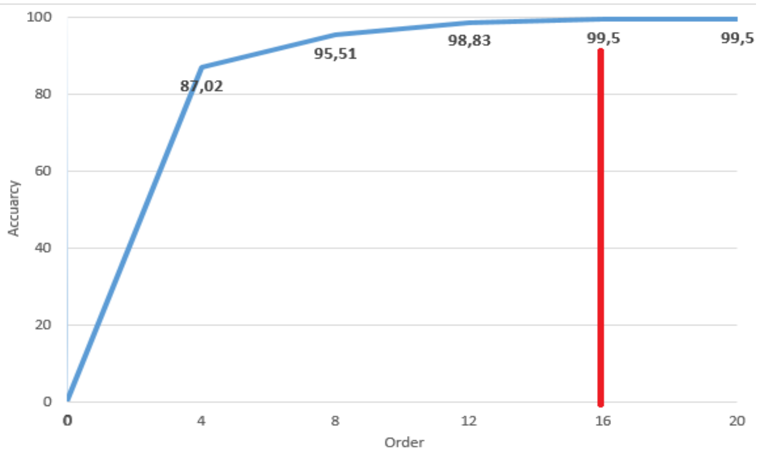


Figure 8: Accuracy of Zernike descriptor with different orders

metrics (MSE, MAE, Cosine) and finally choose the Euclidian distance which had the best performance. Note that the approach does not only give the more suitable silhouette but gives in a classified way the n^{th} most probable silhouettes assumptions (H_1 to H_n). The final pose can be either the first result returned by the system (winner-takes-all) or the mean pose among the n most probable result. One of the well-known problem in pose estimation is the ambiguity when many silhouettes matched, due to symetry. To resolve this ambiguity in human motion analysis system (not the purpose of this paper), one can keep the n ambiguous assumptions and by using multi-hypothesis approaches, can find the correct matching poses in window time (Δt).

In order to quantitatively evaluate the results, we used the simulation. By knowing the real skeleton of the test image, we can process the reprojection error of the estimated 3D joints. This criteria of reprojection error has been chosen over the 3D Euclidian distance between joints in order to be able to use in the futur manually labeled images as, for some database, the 3D ground truth is not available. According to the experimental result, when the mean error is less than 6 pixels, the pose of the result is considered similar to the pose of the requested silhouette. Under this empiric threshold, the difference between two silhouettes is hardly visible for a human.

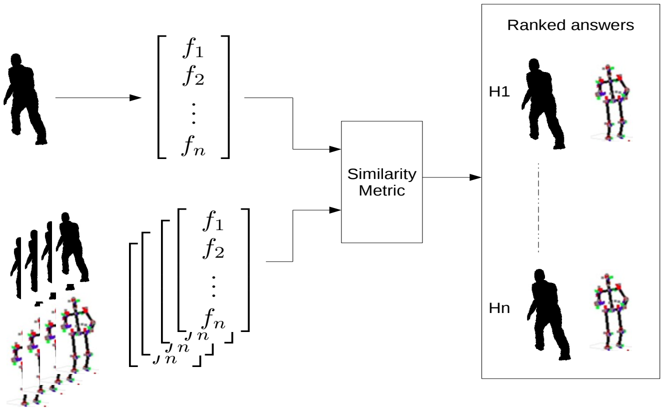


Figure 9: Overview of the experimental process for silhouette recognition

5.1. Pose Estimation

In this subsection, we'll show some results of pose estimation with the different shape descriptors presented in section 4. In order to make some comparison, we used both simulated images taken from our database and a not-simulated images taken from humanEva database [4] and some that we recorded. In order to first, make visual comparison, we'll show below some visual result for pose estimation.

In Fig. 10, we show a bad pose estimation result with Hu descriptor. For this silhouette, the descriptor can't find a good matching result in the database. Note that here, we used the first image return by the program. We can also note that the mean reprojection error is **24.84 px** which is over the empiric threshold.

In opposite, Fig. 11 shows skeletons estimation from a single monocular image with Krawtchouk descriptor. We obtain approximatively (without a visual difference) the same result with Zernike and Hahn descriptor. For this result, the reprojection error of the first image is **0.92 px**, of the second is **2.69 px**

[4] http://humaneva.is.tue.mpg.de/

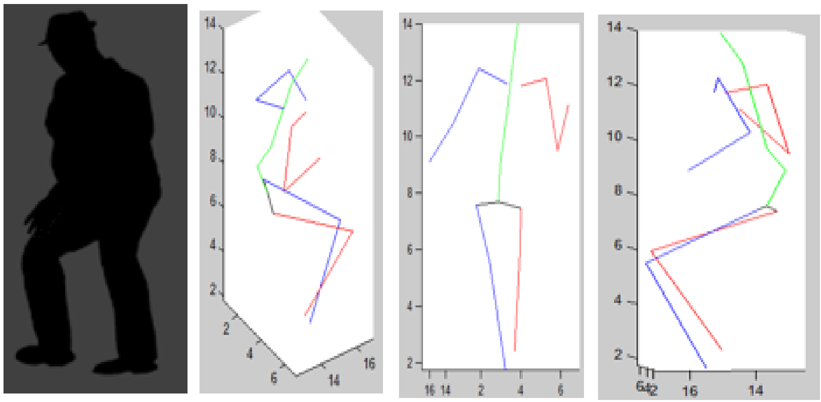


Figure 10: 3D pose estimation results with Hu descriptors: Left, the resquest silhouette and from left to right, the 3D estimated skeleton from various viewpoints

and of the last image is **3.04 px**. These means errors show that the retrieval pose is near to the original pose. Note that the reprojection error can be due to the scale difference between images as we normalize the bounding box and not the real human size.

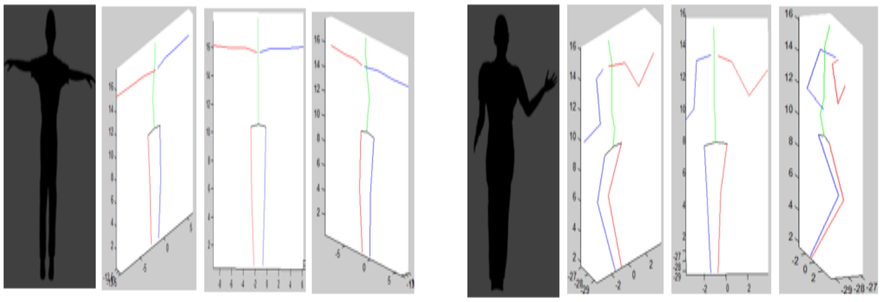


Figure 11: 3D pose estimation results with Krawtchouk descriptors: Left, the resquested silhouette followed from left to right, by the 3D estimated skeleton from differents viewpoints

The test on simulated images dataset shows very accurate results. In case of a realistic image, we used images that we recorded and also from the publicy dataset ”humaneva”. Others publicy dataset in human pose estimation, can’t be exploit in our framework, due either to the lack of motion capture file that can be import in the graphic software or to the lack of 3D ground truth. This make a quantitave evaluation and comparision on publicy dataset difficult. At this step, we choose to present such experiment in a qualitative way to have a visual validation.

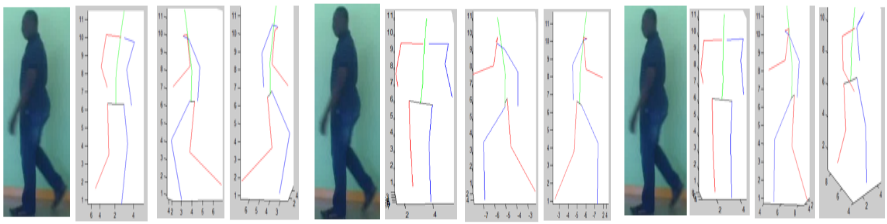


Figure 12: Realistic image tested (back, left and right view of the skeleton) by respectively Hahn, Krawtchouk and Zernike descriptors

In figure 12, we submitted an image extracted from a walking action video with a complex pose. Let's note that, as the approach is based on silhouette only, the video doesn't require to have a very good quality. In this experiment, the tested poses aren't in the dataset, the system will try to find the nearest existing pose. So, we don't expect to get an exact 3D pose as a result, but an approximative pose. Visually, we can note that the result of Hahn is closest to the original pose than Krawtchouk and Zernike results. The most difference is the spreading of foot and arm. However, as the approach is based on silhouette only, the descriptors is not able to totaly make difference between the right side from the left side for foot and arm. This kind of confusion due to the point of view can be solved for action recognition system by using a multi-hypothetical tracking. This is not treat in this paper and will be investigate in future work.

To obtain quantitative result, it's necessary to have dataset with 3D pose as ground truth. Some publicy dataset doesn't provide those ground truth. In case of HumanEva, ground truth are avaible but due to the format, an interface is necessary to convert in CMU format, on which our system is train. Another way, which is more general, is to train the system with data from the CMU dataset used for simulated image and made a visual test on other dataset without quantitative evaluation. Previous results on simulated data and on our won recorded video is confirmed on "Humaneva" dataset. In 13, we present the

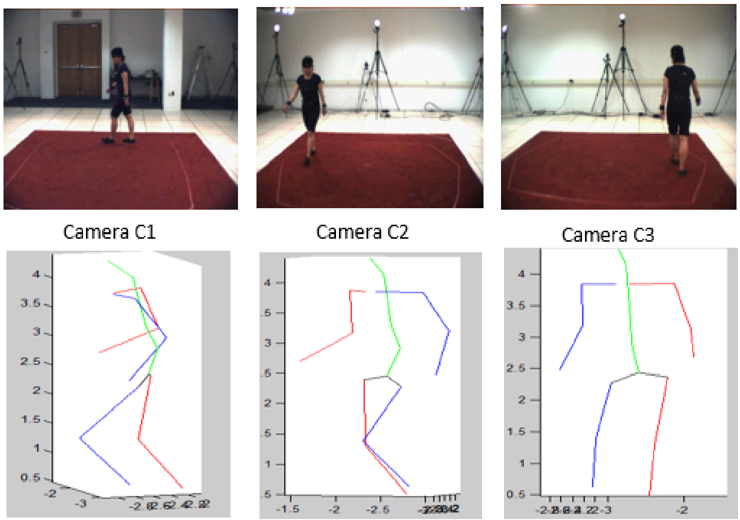
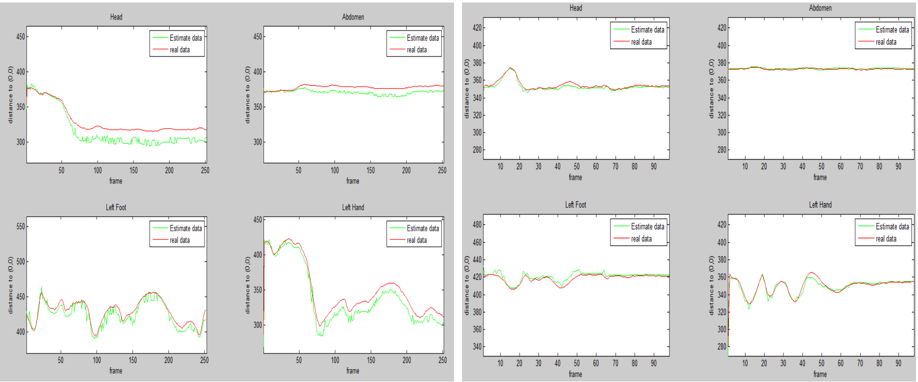


Figure 13: Real world data tested from Humaneva dataset tested by Hahn descriptors

result from Hahn descriptor by testing with the image at the position frame number 21 recorded from the camera ”C1”. The result present here is also an approximation of the 3D pose. We can observe the result for differents viewpoint captured by cameras C2 and C3. By this way we also tested the generalization of our approach in spirit to be free from the dataset that we used for training. The main difficult is when some pose don’t have equivalent in the training data. In this case, the system failed trying to find the closet pose. To solve that, motion capture that cover almost configuration is need for training. It’s impossible in pratice and still a challenge task for machine learning scientist.

Another way to evaluate the consistency, statiblity and therobustness of our approach, one can estimate the 3D pose of person for differents motions in time domain by considering the successive detections during a complete movie of movement. Figure 15 (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. It seems that the red and green curves have the same appearance, which means that the successive detections are stable in time and that the method is reliable.

An offset due to shape scaling, however, exist. Note that there is no use of the time line and each frame is processed independently.



(a) Climb motion (b) Jump motion

Figure 14: Tracking result with Hahn descriptor

5.2. Representativity and descriptor robustness to noise

Silhouette extraction is still an active research field. It is well known that extraction is subjected to noise. The first point was to check the descriptors robustness to noise. For this, we conducted some experiments with the dataset introduced in previous subsection. This dataset contains **2404** unlearned data. By unlearned data, we mean image of a human avatar that wasn't used for the training database construction. We added gaussian noise on the image in the database in order to perturb the originally extracted silhouette. The aims of these experiences are to evaluate the capacity of shape descriptors to encode various shapes with different values of the standard deviation of the 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 $std = \{0, 1, 2, 3\}$.

For the experiment, the training dataset was composed of **11700** silhouettes and the testing dataset was composed of **2404** silhouettes. Since we extracted the silhouettes from motions recorded in videos, the difference between image at

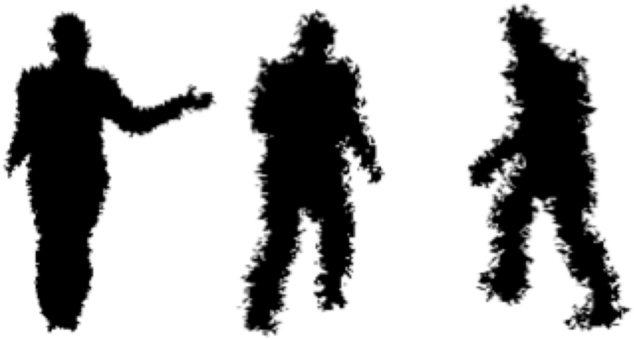


Figure 15: Example of noised silhouettes

frame t and image at frame $t+1$ is hardly perceptible. When $std = 0$, we have the original silhouette and when $std > 0$, the Gaussian white noise is added on the silhouette. The histograms in Fig. 16, on page 29 show that more the std increases, more the recognition accuracy decreases. For a single neighbour ($N = 1$), with $std = \{0, 1, 2, 3\}$, the recognition rate is respectively $RR = \{35.44, 35.44, 24.95, 18.96\}$ for Hu descriptor, $RR = \{98.67, 97, 80.53, 58.56\}$ for Krawtchouk descriptor, $RR = \{98.67, 97.67, 82.86, 66.38\}$ for Zernike descriptor and $RR = \{99.67, 96, 81.85, 61.4\}$ for Hahn descriptor. This accuracy grows up quickly when we augment the number of N assumption returned by the program. Aside the result of Hu descriptor, we can note that the others descriptors has good and interesting accuracy. A ranking of different descriptors can be made based on these results by calculating for the descriptor, the mean accuracy over the entire dataset of noised silhouettes. Mean accuracy over the three datasets is presented in table 1.

Based on the mean accuracy (Table 1), we note that the Hahn descriptor outperformed the other descriptors when we consider more than one neighbour. Zernike outperformed when we consider the first result return by the program. Of course, the difference between the mean accuracy of the three last descriptors is very small (**less than 1%**), so we can't conclude which one is the best descriptor. Another interesting analysis is the run time of each descriptor. The

Table 1: **Mean accuracy in percent for each descriptor**

Descriptors	**N = 1**	**N = 3**	**N = 5**	**N = 7**
Hu	28.69	41.38	47.83	51.78
Krawtchouk	83.69	89.93	91.89	93.34
Zernike	**86.39**	90.63	92.71	93.88
Hahn	84.73	**90.84**	**92.71**	**94**

table 2 show the mean run time of each descriptor in the feature extraction step in second.

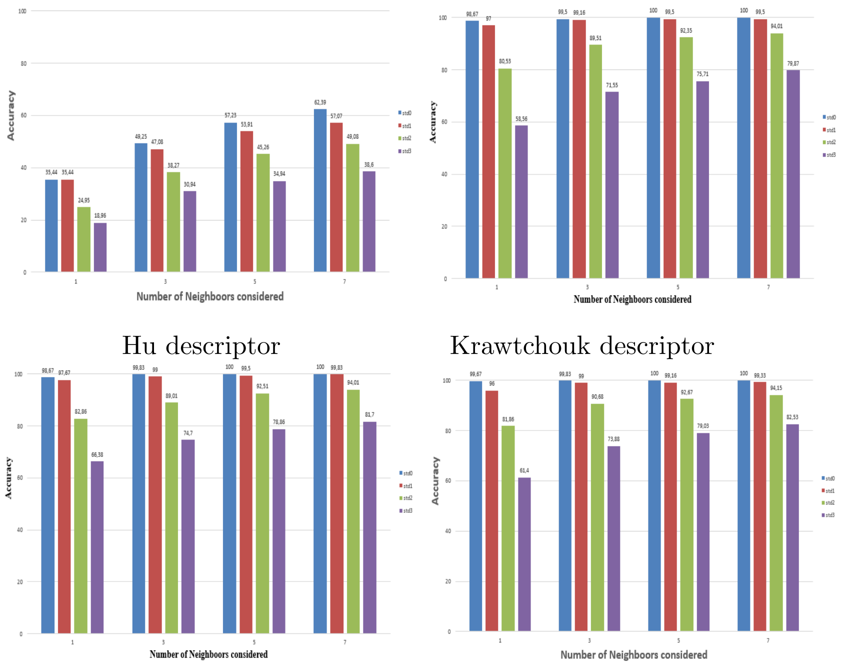
Table 2: **Run time of each descriptor in (s)**

Descriptors	**Hu**	**Kraw**	**Zernike**	**Hahn**
Run time	0.047	0.031	0.149	0.036

According to the run time of each descriptor to extract a feature vector for images, Krawtchouk descriptor, and Hahn are faster than Zernike descriptor. By combining the accuracy and runtime factors, we can choose Hahn descriptor as the best one for our approach.

6. Conclusions

In this paper, we presented a very simple framework for 3D human pose estimation from a single image. In particular, we referred to a scenario where the environment is equipped with a simple low-cost passive camera without the need of any depth information or field of view intersection. The mains novelty of the approach are the use of open source Softwares as Blender and Makehuman in order to easily generate the learning database and the proof that orthogonal moments are able to encode the shape of silhouette pose estimation purpose. We proved that using a very simple framework based on silhouettes comparisons, a full accurate 3D pose estimation was possible in real-time using a single image. In order to match learned and test silhouettes, we compared



Hu descriptor

Krawtchouk descriptor

Zernike descriptor

Hahn descriptor

Figure 16: Histogram of accuracy for unlearned data : colors represent the noise amplitude resp. $\{0, 1, 2, 3\}$ pixels. The abscisses represent the number N of neighboors considered $\{1, 3, 5, 7\}$.

Hu geometric moment and three orthogonal moments for shape description: Zernike, Krawtchouk and Hann moments. Moreover, we tested different moment orders and selected the best suitable for our approach. The proposed posture recognition method gives very promising results in real-time allowing to detect the pose with an accuracy between 84% and 94% depending the number of assumption chosen. As expected, the main limitation of our system is the non-detection of the symmetry of the human body as the left and the right part cannot be differentiated in a single silhouette view. In this regard, future work can concern the use of multiple hypotheses tracking for a video sequence in order to deal with this ambiguity.

References

[1] C. Wang, Y. Wang, A. Yuille, An approach to pose-based action recognition, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 915–922. `doi:10.1109/CVPR.2013.123`.

[2] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1385–1392.

[3] C. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, in: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, Vol. 1, 2000, pp. 677–684 vol.1. `doi:10.1109/CVPR.2000.855885`.

[4] M. W. Lee, R. Nevatia, Human pose tracking in monocular sequence using multilevel structured models, Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (1) (2009) 27–38. `doi:10.1109/TPAMI.2008.35`.

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

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

[7] L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1365–1372.

[8] J. Valmadre, S. Lucey, Deterministic 3d human pose estimation using rigid structure, in: Computer Vision–ECCV 2010, Springer, 2010, pp. 467–480.

[9] X. K. Wei, J. Chai, Modeling 3d human poses from uncalibrated monocular images, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1873–1880.

[10] E. H. H. Shum, Real-time physical modelling of character movements with microsoft kinect, Symposium on Virtual Reality Software and Technology (VRST 12) (18th) (2012) 17–24.

[11] S. Gaglio, G. L. Re, M. Morana, Human activity recognition process using 3-d posture data, IEEE Transactions on Human-Machine Systems 45 (5) (2015) 586–597. `doi:10.1109/THMS.2014.2377111`.

[12] Y. K. A. Jalal, Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data, International Conference on Advanced Video and Signal Based Surveillance (AVSS) (2014) 119–124.

[13] S. K. L. Piyathilaka, Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features, Conference on Industrial Electronics and Applications (ICIEA) (2013) 567–572.

[14] J. H. J. Chai, Performance animation from low-dimensional control signals, ACM Trans. Graph. (2005) 686–696.

[15] E. S. Ho, J. C. Chan, D. C. Chan, H. P. Shum, Y. ming Cheung, P. C. Yuen, Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments, Computer Vision and Image Understanding`doi:http://dx.doi.org/10.1016/j.cviu.2015.12.011`.

[16] A. G. D.F. Fouhey, V. Delaitre, People watching: Human actions as a cue for single-view geometry, Proceedings of the 12th European Conference on Computer Vision (2012) 732–745.

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

[18] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: In ICCV, 2005, pp. 1395–1402.

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

[20] M. de La Gorce, D. Fleet, N. Paragios, Model-based 3d hand pose estimation from monocular video, Pattern Analysis and Machine Intelligence, IEEE Transactions on 33 (9) (2011) 1793–1805. doi:10.1109/TPAMI.2011.33.

[21] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer vision and image understanding 104 (2) (2006) 90–126.

[22] M. Slembrouck, D. Van Cauwelaert, P. Veelaert, W. Philips, Shape-from-silhouettes algorithm with built-in occlusion detection and removal, in: International Conference on Computer Vision Theory and Applications (VISAPP 2015), SCITEPRESS, 2015.

[23] F. D. Atrevi, D. Vivet, F. Duculty, B. Emile, 3d human poses estimation from a single 2d silhouette, in: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2016, pp. 361–369. doi:10.5220/0005711503610369.

[24] S. B. P. Dollar, P. Perona, The fastest pedestrian detector in the west, in: In: Proceedings of the British Machine Vision Conference, 2010, pp. 1–11.

[25] M. E. Celebi, Y. A. Aslandogan, A comparative study of three moment-based shape descriptors, in: International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II, Vol. 1, 2005, pp. 788–793 Vol. 1. doi:10.1109/ITCC.2005.3.

[26] A. Derbel, D. Vivet, B. Emile, Access control based on gait analysis and face recognition, Electronics Letters 51 (10) (2015) 751–752. doi:10.1049/el.2015.0767.

[27] M.-K. Hu, Visual pattern recognition by moment invariants, information Theory, IRE Transactions on 8 (2) (1962) 179–187.

[28] R. Mukundan, K. Ramakrishnan, Moment functions in image analysistheory and applications, World Scientific, 1998.

[29] M. R. Teague, Image analysis via the general theory of moments, JOSA 70 (8) (1980) 920–930.

[30] P.-T. Yap, R. Paramesran, S.-H. Ong, Image analysis by krawtchouk moments, Image Processing, IEEE Transactions on 12 (11) (2003) 1367–1377. doi:10.1109/TIP.2003.818019.

[31] P.-T. Yap, R. Paramesran, S.-H. Ong, Image analysis using hahn moments, Pattern Analysis and Machine Intelligence, IEEE Transactions on 29 (11) (2007) 2057–2062.

[32] E. Pillu, Analyse et rgularisation spatio-temporelle: application l'criture manuscrite, Master's thesis, Universit Lille1 (2011).

[33] R. Mukundan, S. Ong, P. A. Lee, Image analysis by tchebichef moments, IEEE Transactions on image Processing 10 (9) (2001) 1357–1364.

[34] A. Khotanzad, Y. H. Hong, Invariant image recognition by zernike moments, Pattern Analysis and Machine Intelligence, IEEE Transactions on 12 (5) (1990) 489–497.