Silhouette-based Human Action Recognition using Sequences of Key Poses

Alexandros Andre Chaaraoui[a,*], Pau Climent-Pérez[a], Francisco Flórez-Revuelta[b]

[a] *Department of Computing Technology, University of Alicante, P.O. Box 99, E-03080, Alicante, Spain*
[b] *Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Road, KT1 2EE, Kingston upon Thames, United Kingdom*

Abstract

In this paper, a human action recognition method is presented in which pose representation is based on the contour points of the human silhouette and actions are learned by making use of sequences of multi-view key poses. Our contribution is two-fold. Firstly, our approach achieves state-of-the-art success rates without compromising the speed of the recognition process and therefore showing suitability for online recognition and real-time scenarios. Secondly, dissimilarities among different actors performing the same action are handled by taking into account variations in shape (shifting the test data to the known domain of key poses) and speed (considering inconsistent time scales in the classification). Experimental results on the publicly available

*Corresponding author: Alexandros Andre Chaaraoui, Department of Computing Technology, University of Alicante, P.O. Box 99, E-03080, Alicante, Spain. Phone: +34 965903681, Fax: +34 965909643

Email addresses: `alexandros@dtic.ua.es` (Alexandros Andre Chaaraoui), `pcliment@dtic.ua.es` (Pau Climent-Pérez), `F.Florez@kingston.ac.uk` (Francisco Flórez-Revuelta)

URL: `http://www.dtic.ua.es` (Alexandros Andre Chaaraoui)

Weizmann, MuHAVi and IXMAS datasets return high and stable success rates, achieving, to the best of our knowledge, the best rate so far on the MuHAVi *Novel Actor* test.
Keywords: human action recognition, key pose, key pose sequence, Weizmann dataset, MuHAVi dataset, IXMAS dataset

1. Introduction

Human action recognition has been of great interest in recent years due to its direct application and need in Surveillance, Ambient Intelligence, Ambient-Assisted Living (AAL) and Human-Computer Interaction systems. While it is still a recent field of research, huge advances have been made in classification of human actions (Poppe, 2010; Turaga et al., 2008; Weinland et al., 2011), recognition based on context and scene understanding (Kjellström , Sidenbladh; Bremond, 2007), as well as enhancement of traditional tracking and motion analysis systems with semantics about human activities (Moeslund et al., 2006; Hu et al., 2004). In this paper, a simple but yet very effective approach is presented in order to support accurate human action recognition at the level of basic human motion, like *walking, jumping, running, falling*, etc. Based on human silhouettes, a scale and location invariant feature is computed which shows to be a powerful discriminating signal, especially when considering its variation over time. At the training stage, the method learns the per class features that make up the most characteristic poses, the so called *key poses*. These can be acquired from single- or multi-view data, which makes the method suitable for scenarios with one or more cameras without any explicit constraints about the point of view (POV).

Using the ground truth data, the *sequences of key poses* corresponding to the labelled videos are obtained. These sequences are matched later with the current test sequence based on Dynamic Time Warping (DTW).

Our system has been designed so as to run at a frame rate close to real-time and to support online recognition. Since our target application is human monitoring at home for AAL services, these were both essential premises. Experimentation on three popular benchmarks (Weizmann from Blank et al. (2005), MuHAVi from Singh et al. (2010) and IXMAS from Moeslund et al. (2006)) shows that our approach outperforms state-of-the-art methods with similar conditions.

The contributions to the literature of this paper are two-fold. On the one hand, an efficient human action recognition method is presented which can be applied in a wide spectrum of application scenarios due to its performance in real-time and the absence of requirements as camera calibration or specific POVs. On the other hand, in this work human action recognition is carried out based on sequences of key poses. This achieves to filter noise and outliers from the training instances while at the same time it models the temporal evolution between key poses.

The remainder of this paper is organised as follows: section 2 summarises the most relevant and recent related works in human action recognition. In section 3 the chosen pose representation is analysed briefly. Our model learning approach is broken down into steps in section 4, and the final action recognition stage is presented in section 5. Section 6 gives a detailed analysis about the experimental results obtained and compares them with other state-of-the-art references. Finally, section 7 presents some conclusions and discussion.

2. Related Work

When analysing human action recognition approaches based on vision techniques, classification can be made with respect to different semantic levels. Common criteria are: 1) the structural layout of the recognition method (Aggarwal and Ryoo, 2011); 2) the learning approach, for instance, exemplar-based vs. model-based, where we find generative models like Hidden Markov Models (HMM) and discriminative models like Conditional Random Fields (CRF) (Poppe, 2010); 3) the type of input features used for the classification (Poppe, 2010; Weinland et al., 2010).

Attending to the latter, *global* (also known as *dense* or *holistic*) representations and *local* (also known as *sparse*) representations of the images can be obtained. The first require a region of interest (ROI) and therefore the human body needs to be detected in the image, usually with background subtraction and blob extraction techniques. While this additional step of pre-processing is a disadvantage, it is usually overcome by the significant reduction of both image size and inherent complexity of its content. Bobick and Davis (2001) used such a global representation in their Motion History- and Energy-Images (MHI, MEI), which encode the temporal evolution of the movement of the image and its spatial location respectively over a sequence of frames. Weinland et al. (2006) extended the work of Bobick and Davis (2001) to a 3D Motion History Volume in order to combine images from multiple cameras and to obtain a free-viewpoint representation. While Bobick and Davis (2001) use seven Hu Moments for description and classification, Weinland et al. (2006) use Fourier analysis in cylindrical coordinates. Space-time volumes are constructed in Blank et al. (2005) by means of obtaining the solution to the Poisson equation for a sequence of binary silhouettes. Global space-time features (composed of the weighted moments of local space-time saliency and orientation features) are employed to achieve action recognition, detection and clustering. More recently, MHI templates have been clustered in a Self-Organising Map in order to represent image viewpoint and movement in a principal manifold (Martinez-Contreras et al., 2009). Each sequence of MHI is projected onto the map and the coordinates of activation are modelled with an HMM. Maximum Likelihood classifier is used for the final recognition.

There are also works which take advantage of image features that have not been originally designed for action recognition. Image gradients and optical flow have been widely and successfully used in tracking methods and their application to action recognition shows good results. In this sense, Tran and Sorokin (2008) designed a complex combination of shape and motion features. A 286-dimensional descriptor is obtained by encoding the binary shape of the silhouette, the vertical and horizontal optical flow and the context of 15 surrounding frames reduced with PCA. Nearest Neighbour classification is done by discriminative metric learning and data subsampling. Fathi and Mori (2008) use mid-level motion features (spatio-temporal cuboids) made up of weighted combinations of thresholded low-level features based on optical flow. A variant of Adaboost is applied and one binary classifier is learned for every pair of classes in order to obtain a multi-class classifier, which achieves highly accurate results on popular action recognition datasets (Weizmann from Blank et al. (2005) and KTH from Schuldt et al. (2004)). Main disadvantages of such global representations are the lack of resistance to viewpoint changes and partial occlusions; under these circumstances global representations suffer from high intra-class variance and are therefore difficult to learn accurately.

When using local representations, the image is regularly taken as it is and observed as a collection of patches or points. Commonly different types of salient points are obtained based on shape and gradient changes (like Harris and SUSAN corners, SIFT and SURF points; see Wu et al. (2010b); Juan and Gwun (2009) for more details). When considering the temporal evolution of the location or aspect of these points, space-time corners are applied. These encode 3D information of interest points "where the local neighbourhood has a significant variation in both the spatial and the temporal domain" (Poppe, 2010). Great effort has been made to extend traditional salient point detectors to 3D: Laptev (2005) used the Harris corner as basis, while Oikonomopoulos et al. (2005) extended the salient point detector from Kadir and Brady (2003), and Scovanner et al. (2007) created a 3D version of the popular SIFT points. A different approach is presented in İkizler and Duygulu (2007), where the human body is represented with oriented rectangular patches; then a histogram is obtained with the 15° orientations resulting in 12 circular bins. Spatial information is encoded using a 3x3 grid and concatenating the histograms of each individual bin. Among different recognition methods, DTW showed the best results achieving perfect accuracy with the Weizmann dataset. While local representations have achieved good recognition rates, great obstacles persist in attaining stable and constant features in cluttered environments.

For greater detail about these methods and exhaustive reviews about the state of the art, we refer to the popular works Poppe (2010) and Moeslund et al. (2006), or more recent ones, like Aggarwal and Ryoo (2011); Chaaraoui et al. (2012).

3. Pose Representation

As introduced in section 1, our method relies on a global pose representation based on the contour points of the silhouette. We assume that a binary silhouette is obtained previously by human silhouette extraction techniques, e.g. background subtraction. Using only the contour points and not the whole silhouette is motivated by getting rid of the redundancy that introduces the inside part of the human silhouette, leading therefore to a less expensive feature extraction. In addition, usage of contours avoids the need of morphological pre-processing steps and reduces the sensitivity to small viewpoint variations or lighting changes (Ángeles Mendoza and Pérez de la Blanca, 2007). Specifically, the contour-based feature from Dedeoğlu et al. (2006) has been chosen, which is described briefly in the following.

First, the contour points $P = \{p_1, p_2, ..., p_n\}$ of the silhouette need to be obtained. For this purpose, contour extraction is applied based on the border following algorithm from Suzuki and be (1985).

Second, the centre of mass $C_m = (x_c, y_c)$ of the silhouette's contour points is calculated with respect to the n number of points:

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, y_c = \frac{\sum_{i=1}^{n} y_i}{n}. \tag{1}$$

Third, the distance signal $DS = \{d_1, d_2, ..., d_n\}$ is generated by determining the Euclidean distance between each contour point and the centre of mass. Contour points should be considered always in the same order. For instance, the set of points can start at the most left point with equal y-axis value as the centre of mass, and follow a clockwise order.

$$d_i = \|C_m - p_i\|, \quad \forall i \in [1...n]. \tag{2}$$

Finally, scale-invariance is achieved by fixing the size of the distance signal, sub-sampling the feature size to a constant length L, and normalising its values to unit sum.

$$\hat{DS}[i] = DS\lceil i * \frac{n}{L} \rceil, \quad \forall i \in [1...L], \tag{3}$$

$$\bar{DS}[i] = \frac{\hat{DS}[i]}{\sum_{i=1}^{L} \hat{DS}[i]}, \quad \forall i \in [1...L]. \tag{4}$$

This type of global pose representation has a significant advantage over similar features presented in section 2. While the spatial information is preserved in greater detail than histogram- or grid-based representations, the feature still has a low dimensionality and its processing presents a very low computational cost (see section 6).

4. Model Learning

Lately, several works (Baysal et al., 2010; Cheema et al., 2011; Eweiwi et al., 2011; Thurau and Hlaváč, 2007) build upon key poses. Baysal et al. (2010) define key poses as "a set of frames that uniquely distinguishes an action from others". Therefore, the goal of using key poses is to model an action by its most characteristic poses in time. This makes it possible to significantly reduce the problem scale in exemplar-based recognition methods and, at the same time, to avoid redundant or superfluous learning. The underlying idea is that if the human brain is able to recognise what a person is doing based on a few individual images, why should not action recognition methods be able to sustain only on pose information. In this regard, Baysal et al. (2010); Cheema et al. (2011) use no temporal information at all, Thurau and Hlaváč (2007) model the short-term temporal relation between consecutive key poses with n-grams (*trigrams* showed good results at acceptable computational cost), and Eweiwi et al. (2011) take into account the temporal context of a small number of frames by means of obtaining temporal key poses based on MHI. While our approach is very similar to these works at the training stage when applied to a single view, our contribution considers long-term temporal relation between key poses and thus takes advantage of the known temporal evolution of key poses over a whole sequence.

A complete overview of the involved stages of the learning process can be seen in figure 1.

4.1. Learning Key Poses

The first step of the learning process is to process all the frames of the video sequences in order to obtain their pose representation, as mentioned in section 3. Then, similar to Cheema et al. (2011); Baysal et al. (2010), the per class key poses are learned by means of *K-means* clustering with Euclidean distance. Hence, the extracted features of all available images of the same action class $samples = \{s_1, s_2, ..., s_n\}$ are grouped into K clusters; where

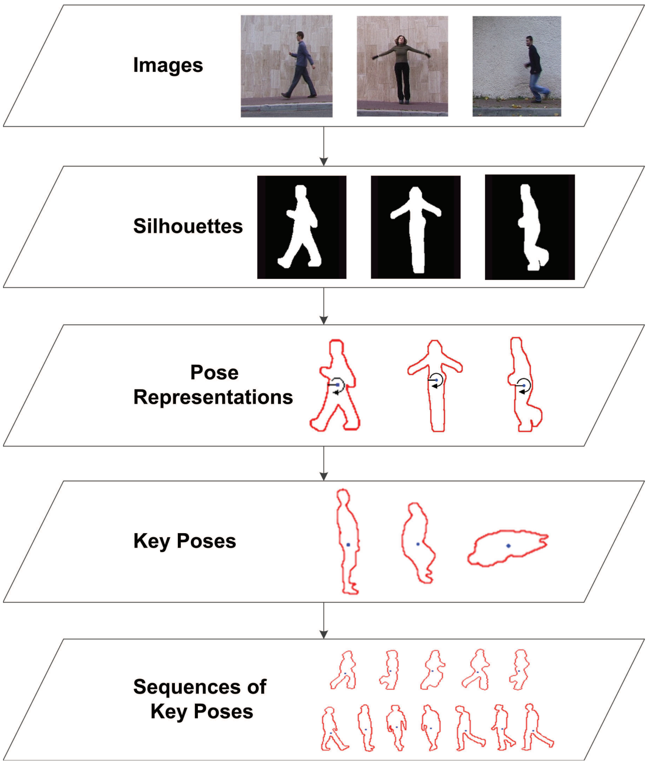

Figure 1: Overview of the learning process: first, a human silhouette extraction technique, like background subtraction, needs to be applied. Then the extracted human silhouettes are processed in order to obtain the contour-based feature. Finding the most characteristic poses among the training data returns the key poses. The sequences of key poses model the temporal evolution between key poses with respect to the original training sequences.

each cluster centre of $centres = \{c_1, c_2, ..., c_K\}$ represents a key pose kp as it is a characteristic pose among the training data. The process of clustering is repeated λ times, so as to avoid local minimum, and the best result is taken (the usage of more advanced clustering algorithms is being considered for future works). Given that the clustering process returns the corresponding label of each sample, $labels = \{l_1, l_2, ..., l_n\}$ in which l_i stands for the index of the cluster assigned to s_i, clustering results are evaluated with the following compactness metric C:

$$C = \sum_{i=1}^{n} |s_i - c_{l_i}|, \tag{5}$$

where the instance with the lowest value is taken as the final result.

This key pose learning process is repeated individually for the training samples of each action class. This way, a set of K key poses is obtained for each action class.

4.2. Learning Sequences of Key Poses

As stated beforehand, our goal is to learn the long-term temporal evolution of key poses. Consequently, our interest resides on the successive key poses that are involved in an action performance. As the training data is made up of sequences of labelled action performances, the corresponding sequences of key poses can be modelled. For the pose representation of each frame of a sequence, i.e. $S_{poses} = \{pose_1, pose_2, ..., pose_n\}$, the *nearest neighbour* key pose is found. The successive *nearest neighbour* key poses constitute the simplified sequence of known characteristic poses and their evolution: $S = \{kp_1, kp_2, ..., kp_n\}$. This way, a set of sequences of key poses is obtained for each action class. This decisive step significantly improves exemplar-based action recognition by shifting the training data to a common and known domain (the set of characteristic key poses), and therefore filtering out single examples with noise or partial occlusions.

4.3. Learning from Multiple Views

Nowadays, most application scenarios do have more than one camera available. Multiple views of the same environment help to avoid occlusions due to obstacles (like furniture or having several persons in the field of view) and make it possible to have multiple POV of the same event at our disposal. However, the task of dealing with several video streams, modelling 3D representations and targeting action recognition applications still has to overcome great difficulties, as dealing with richer data leads to high computational cost and burdensome systems (Moeslund et al., 2006; Holte et al., 2011).

Since the presented method shows successful results in single-view action recognition, one wonders if the approach is able to accurately model multi-view data. Among the different available approaches of combining multi-view data (Holte et al., 2011; Wu et al., 2010a) a *feature fusion* approach has been chosen, so as to test if the model based on sequences of key poses is able to learn from multiple views. In this sense, multi-view data is combined at the feature level and no changes are performed at the modelling or recognition levels.

Assuming that v video streams of the same scenario are available, first each frame is individually processed to its pose representation. Then the multi-view pose representation $\bar{DS}_{mv}$ is obtained by frame-by-frame con-

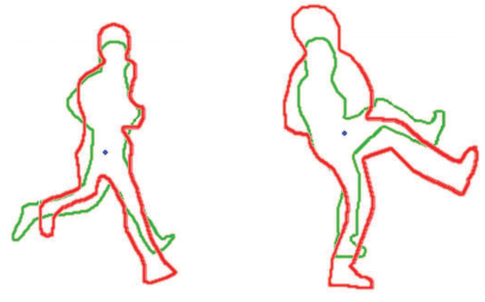

Figure 2: Multi-view key poses: *RunLeftToRight* (left) and *KickRight* (right) from MuHAVi.

catenation of single-view pose representations $\bar{DS}_{sv}$:

$$\bar{DS}_{mv} = \bar{DS}_{sv_1} \circ \bar{DS}_{sv_2} \circ ... \circ \bar{DS}_{sv_v}. \tag{6}$$

This step is identically performed with train and test instances, using multi-view pose representations at the succeeding stages. As a result, when feeding the model with multi-view pose representations, sequences of multi-view key poses (see figure 2) are inherently obtained.

5. Action Recognition

At the recognition stage, a final class label output needs to be given. To that end, two steps have to be taken: 1) in the same way as with our training sequences, silhouette contour points are processed and their corresponding pose representations are obtained; 2) for each test sequence, the pose representation of each frame is used to find the *nearest neighbour* key pose and build the analogous sequence of *nearest neighbour* key poses. This shift to our known data domain acts as filtering and simplification process, and introduces the needed stability when dealing with test data with meaningful differences to the training data, like action performances of different actors (see section 6).

Due to the temporal intra-class variance, a suitable distance metric is needed in order to compare the sequences of key poses. Different actors can perform the same actions on very different ways and they can do so faster or slower than others. While some motions are indispensable when performing an action, like moving one leg and then the other while walking, these can still appear with a considerable time shift, especially when dealing with elderly people. Dynamic Time Warping is particularly suitable when dealing with the comparison of sequences that can present inconsistent time scales, but without changing the temporal order. It is able to align two time series of different lengths even if there are accelerations or decelerations.

Given two sequences of key poses $S_{train} = \{kp_1, kp_2, ..., kp_n\}$ and $S_{test} = \{kp'_1, kp'_2, ..., kp'_m\}$ we compute the DTW distance $S_{train} - S_{test}$ as:

$$S_{train} - S_{test} = dtw\left(n, m\right), \tag{7}$$

$$dtw\left(i, j\right) = \min \left\{ \begin{array}{l} dtw\left(i-1, j\right), \\ dtw\left(i, j-1\right), \\ dtw\left(i-1, j-1\right) \end{array} \right\} + d(kp_i, kp'_j), \tag{8}$$

where $d(kp_i, kp'_j)$ is the Euclidean distance used for feature comparison between two key poses.

This way, using DTW, the nearest neighbour sequence of key poses is found and its label supplies the final result.

6. Experimentation

In order to test the accuracy and stability of the presented approach, three human action recognition datasets have been used as benchmarks. In the case of the Weizmann dataset, a *leave-one-sequence-out* cross validation procedure has been applied. This way, the system is trained with all but one video sequence, which is the one that evaluates the accuracy score. Iterating over all the sequences, the average success rate is used as final result. In the case of the MuHAVi dataset, its authors introduced an evaluation scheme based on view- and actor-invariance tests which we repeat so as to compare our results. And in the IXMAS dataset we used the usual *leave-one-actor-out* cross validation. Finally, a temporal evaluation is made in order to confirm the suitability for real-time applications. A comparison of the presented results with similar state-of-the-art approaches is given in section 6.5.

The three constant parameters of the presented method have been chosen based on empirical testing. The number of clustering attempts $\lambda = 3$ for all results shown, while the length of the distance signal feature L and the number of key poses per action class K are detailed for each test.

6.1. Weizmann Dataset

The Weizmann dataset presented in Blank et al. (2005) is a single-view (static front-side camera) outdoor dataset. It provides 180x144 px resolution images of 10 different actions performed by 9 actors. It has a relatively simple background, provides automatically extracted silhouettes (we use the version without post-alignment), and has become a reference in human action recognition. Actions include *bending (bend), jumping jack (jack), jumping*

	bend	jack	jump	pjump	run	side	walk	wave1	wave2
bend	9/9								
jack		9/9							
jump			9/9						
pjump				9/9					
run					7/10		3/10		
side			1/9			8/9			
walk							10/10		
wave1		1/9						8/9	
wave2		1/9							8/9

Figure 3: Confusion matrix of the Weizmann dataset without the *skip* action. *Leave-one-sequence-out* cross validation with 83 sequences.

forward (jump), jumping in place (pjump), running (run), galloping sideways (side), skipping (skip), walking (walk), waving one hand (wave1) and *waving two hands (wave2).* It is worth mentioning that several works exclude the *skip* action, as it commonly shows higher error rates and also weakens the recognition of other actions.

Figure 3 shows the result of the cross validation test without the *skip* action. At an average success rate of 92.77% (achieved with $L = 120$ and $K = 96$), it can be seen that the confusions made are coherent. As seen in the works from Saghafi and Rajan (2012); Shao and Chen (2010), *walk* and *run* present a high inter-class similarity, and therefore the difference between their key poses is minimal. In *jack* hands are risen, similarly to *wave1* and *wave2*.

Taking a closer look to the misclassifications of sequences from the *run* action class, it can be seen that the running or walking speed of the actors varied significantly. In addition, some of the actors do not move their arms along when running, which increases even more the similarity between running and walking. We have analysed a specific misclassification of a *run*

Table 1: Ten closest key pose sequences for a specific misclassification of a *run* sequence.

Index	**Action class**	**DTW distance**
1	walk	3,264716
2	run	3,795877
3	walk	4,116315
4	side	4,722770
5	run	4,869563
6	run	5,224457
7	run	5,319681
8	run	5,458966
9	run	6,019087
10	run	6,206304

sequence (see table 1). The ten closest sequences include seven sequences of the right class, which means that, for instance, a K-Nearest Neighbour (KNN) approach could have worked better in this case. The sequence number 2 is the closest sequence that would have produced a successful match. A 100% of its key poses proceed from the training instances of the *run* class. Surprisingly, only ~14% of the frames of the tested sequence have matched with a key pose from this class, which explains why this sequence has been misclassified.

When including the *skip* action, the success rate decreases to 90.32% (achieved with $L = 200$ and $K = 96$). Interestingly, this action is recognised perfectly, but the stability of the other actions is still affected because of the rise of inter-class similarity which occurs when adding this action class. It has been observed that the *skip* key poses get hit very frequently in several action classes as *jump*, *pjump*, *run*, *side* and *walk*. Similar conclusions have been obtained in Saghafi and Rajan (2012); Shao and Chen (2010).

6.2. MuHAVi Dataset

The MuHAVi dataset (Singh et al., 2010) is a more recent and complex benchmark with multi-view images. It provides 720x576 px resolution images on a complex background with street light illumination. Its full version includes 17 different actions performed by 7 actors and has been recorded indoors with 8 CCTV cameras, each one at 45° to its neighbours. A manually annotated subset (*MuHAVi-MAS*) provides silhouettes for 2 of these views (front-side and 45°) and 2 actors, labelling 14 (MuHAVi-14: *CollapseLeft, CollapseRight, GuardToKick, GuardToPunch, KickRight, PunchRight, RunLeftToRight, RunRightToLeft, StandupLeft, StandupRight, TurnBackLeft, TurnBackRight, WalkLeftToRight* and *WalkRightToLeft*) or 8 (MuHAVi-8: *Collapse, Guard, KickRight, PunchRight, Run, Standup, TurnBack* and *Walk*) actions in its merged version.

6.2.1. Leave-one-sequence-out Cross Validation

As this dataset includes multi-view data, our method uses the proposed multi-view pose representations and learns sequences of multi-view key poses. Since two camera views are available, sequences are considered as pairs, each of which contains the images of the same action performance from a different view. Therefore, the 136 available sequences are taken as 68 different sequences when performing the *leave-one-sequence-out* cross validation test.

In figure 4, the confusion matrix for MuHAVi-14 shows very promising results with an average success rate of 91.18% (achieved with $L = 340$ and $K = 90$), misclassifying only 6 sequences.

	CollapseLeft	CollapseRight	GuardToKick	GuardToPunch	KickRight	PunchRight	RunLeftToRight	RunRightToLeft	StandupLeft	StandupRight	TurnBackLeft	TurnBackRight	WalkLeftToRight	WalkRightToLeft
CollapseLeft	4/4													
CollapseRight		4/4												
GuardToKick			6/8	2/8										
GuardToPunch				8/8										
KickRight					8/8									
PunchRight						8/8								
RunLeftToRight							3/4	1/4						
RunRightToLeft								4/4						
StandupLeft									1/2	1/2				
StandupRight										4/4				
TurnBackLeft			1/2								1/2			
TurnBackRight			1/4									3/4		
WalkLeftToRight													4/4	
WalkRightToLeft														4/4

	Collapse	Guard	KickRight	PunchRight	Run	Standup	TurnBack	Walk
Collapse	8/8							
Guard		16/16						
KickRight			8/8					
PunchRight				8/8				
Run					8/8			
Standup						6/6		
TurnBack		2/6					4/6	
Walk								8/8

Figure 4: Confusion matrices of the MuHAVi dataset: MuHAVi-14 (top) and MuHAVi-8 (bottom). *Leave-one-sequence-out* cross validation with 68 multi-view sequences.

In MuHAVi-8 only 2 sequences are misclassified and a success rate of 97.06% ($L = 250$ and $K = 90$) is achieved. In both tests it can be seen that *TurnBack* shows greater difficulty than other actions.

6.2.2. Identical Actors, Novel Camera

In this view-invariance test, all available sequences of one POV are used at training, whereas at testing, the same sequences but from the second POV are used. Hence, no multi-view learning can be applied. This test is executed twice, interchanging the training and testing groups, and the results are averaged.

Since view-invariance has not been explicitly considered, no exceptional robustness is expected in this sense. The test returns a result of 38.97% ($L = 220$ and $K = 70$) on MuHAVi-14 and 63.24% ($L = 370$ and $K = 50$) on MuHAVi-8.

6.2.3. Identical Cameras, Novel Actor

Similarly to the last test, all sequences of one actor are used at training, while the sequences of a different actor, unknown to the learning model, are used at testing (and vice-versa). As more than one view of the same action performance is available, multi-view learning is applied and 34 sequences with images of two views are used at training and another 34 at testing.

In contrast to the last test and as mentioned before, the presented method is designed to be robust to test data with meaningful differences to the train data (due to dissimilarities among actors or noise). For this reason, data is first shifted to the known domain of key poses and then matched to the corresponding train sequence.

Actor-invariance tests present an increased difficulty due to the singularity of multiple actor-dependant conditions. In this sense, parameters as size, body build, clothes, etc. are given by the actor, as well as the particular way in which each person performs an action. This can be seen, for instance, in gait analysis, where the involved dynamics even allow to perform person identification (Wang et al., 2010).

The *Novel Actor* test returns a success rate of 82.35% ($L = 450$ and $K = 110$) on MuHAVi-14 and 88.24% ($L = 250$ and $K = 110$) on MuHAVi-8. To the best of our knowledge, these are the highest results achieved so far.

6.3. IXMAS Dataset

With the purpose of extending the experimentation of our method to a more difficult dataset with more camera views, we have chosen the IXMAS dataset which is popular among human action recognition methods that are specifically designed for multi-view recognition. The INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset (Weinland et al., 2006) includes multi-view data and is especially aimed at view-invariance testing. It provides 390x291 px resolution images from five different angles including four sides and one top-view camera. A set of 12 actors have been recorded performing 14 different actions (*check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw over head* and *throw from bottom up*) 3 times each, resulting in a dataset with over 2 000 sequences. This benchmark presents an increased difficulty because subjects were asked to freely choose their position and orientation. Therefore, each camera has captured different viewing angles, which makes

	check watch	cross arms	scratch head	sit down	get up	turn around	walk	wave	punch	kick	pick up
check watch	28/36	4/36							4/36		
cross arms		32/36	1/36						3/36		
scratch head	1/36	8/36	22/36						2/36	2/36	1/36
sit down				35/36							1/36
get up					36/36						
turn around		1/36				35/36					
walk						6/36	30/36				
wave	3/36		6/36					26/36	1/36		
punch	3/36	1/36	1/36					1/36	30/36		
kick		1/36							4/36	30/36	1/36
pick up											36/36

Figure 5: Confusion matrix of the IXMAS dataset. *Leave-one-actor-out* cross validation with 11 actors and 396 multi-view sequences.

methods which rely on fixed camera views (front, side, etc.) unsuitable.

Figure 5 shows the confusion matrix that has been obtained for this challenging dataset. As common in the state-of-the-art, we used a *leave-one-actor-out* cross validation test in which actor-invariance is tested by training with the instances from all but one actor and testing the sequences from the unknown one. This is repeated for all available actors and the average accuracy score is obtained. Following the test setup given by the publishers of the dataset, we excluded the *point* and *throw* actions. The test returns an average result of 85.86% ($L = 400$ and $K = 20$). As it can be seen in the confusion matrix, the actions that are performed with arms and hands present several misclassifications due to their similarity. *Walk* is matched with *turn around* because the proposed method does only rely on silhouette shape without explicitly learning action's kinematics. Turning around is essentially walking with a specific direction and this is not differentiated by our system.

6.4. Temporal Evaluation

When designing a human action recognition method intended to perform online, the temporal constraint is crucial. Even more when considering that this unit would be only one part of a complex distributed vision system which performs movement detection, tracking, background segmentation, person identification, privacy filtering, etc., and moreover needs to be executed on an embedded hardware device. For this reason, a human action recognition module needs to perform as fast as possible, and simple yet effective approaches are preferred over perfect yet unaffordable ones. Our evaluation system consists of a standard PC with an Intel Core 2 Duo CPU at 3 GHz, running Windows 7 64-bit and an implementation using the .NET Framework and the widely used Computer Vision library OpenCV (Bradski, 2000). Time evaluation has been performed using the hardware counter *QueryPerformanceCounter* with a precision of μs.

Executing the learning process for the 93 sequences of the Weizmann dataset, which contain 5687 frames of 180x144 px, takes $81.1s$. That is an average of $0.87s$ per sequence at $70.12FPS$. But more important is the speed of the testing process which takes $45.72s$, achieving an average speed of $0.49s$ per sequence at $124.38FPS$.

In MuHAVi-14, the training of 136 sequences made up of 7941 frames of 720x576 px takes $204.44s$, i.e. an average speed of $1.5s$ per sequence at $38.84FPS$. The testing process for this data takes $109.9s$, achieving an average speed of $0.81s$ per sequence at $72.25FPS$. As MuHAVi-8 has fewer action classes, the learning process speeds up to $53.76FPS$ and the testing process to $81.31FPS$.

Table 2: Comparison with similar state-of-the-art approaches on the Weizmann dataset.

Approach	**Input**	**Actions**	**Evaluation**	**Rate**	**FPS**
İkizler and Duygulu (2007)	Silhouettes	9	LOSO	100%	N/A
Tran and Sorokin (2008)	Silhouettes	10	LOSO	100%	N/A
Eweiwi et al. (2011)	Aligned sil.	10	LOSO	100%	N/A
Hernández et al. (2011)	Images	10	LOAO	90.3%	98
Cheema et al. (2011)	Silhouettes	9	LOSO	91.6%	56
Our method	Silhouettes	9	LOSO	92.8%	124

In the case of the IXMAS dataset these rates change to $155.52FPS$ for the training process and $26.48FPS$ for the testing process.

These tests were performed including all processing stages from the computing of the contour points to the actual recognition, and using the silhouette images as basis. The obtained performances correspond to the best test configurations shown in previous sections, without applying any further optimisation.

6.5. Comparison of Results

The comparison of different human action recognition approaches can be difficult and misleading because of diverse recognition goals (some only seek an action class label, and others need a reconstructed 3D environment), different kinds of input data (images, video streams, silhouettes, outputs of tracking systems, etc.) and even incompatible evaluation methods.

Table 2 shows a comparison of our result on the Weizmann dataset with other similar approaches. The success rates are obtained either with *leave-one-actor-out* (LOAO) or *leave-one-sequence-out* (LOSO) cross validations. Several works achieve perfect recognition on this dataset, but most of them

Table 3: Comparison with similar state-of-the-art approaches on the MuHAVi dataset. All use silhouettes as input data and LOSO as evaluation method.

	MuHAVi-14		**MuHAVi-8**	
Approach	**Rate**	**FPS**	**Rate**	**FPS**
Singh et al. (2010) (baseline)	82.4%	N/A	97.8%	N/A
Martinez-Contreras et al. (2009)	-	-	98.4%	N/A
Eweiwi et al. (2011)	91.9%	N/A	98.5%	N/A
Cheema et al. (2011)	86.0%	56	95.6%	56
Our method	91.2%	72	97.1%	81

Table 4: Comparison of results of the MuHAVi Novel Actor test.

Approach	**MuHAVi-14**	**MuHAVi-8**
Singh et al. (2010)	61.8%	76.4%
Cheema et al. (2011)	73.5%	83.1%
Eweiwi et al. (2011)	77.9%	85.3%
Our method	82.4%	88.2%

do not present any temporal evaluation and their suitability for real-time applications is arguable. It can be seen that, when comparing with methods that present temporal data, our performance improves state-of-the-art rates both in recognition accuracy and speed.

Table 3 presents similar comparisons for the MuHAVi dataset. Again the present method achieves state-of-the-art success rates and outperforms similar methods with real-time suitability in recognition accuracy, as well as in recognition speed.

We also want to point out the robustness of our method with respect to the *Novel Actor* test. Dissimilarities among action performances from

Table 5: Comparison with other multi-view human action recognition approaches of the state-of-the-art. The rates obtained in the *leave-one-actor-out* cross validation performed on the IXMAS dataset are shown (except for Cherla et al. (2008) where the type of test is not stated).

Approach	**Input**	**Actions**	**Actors**	**Views**	**Rate**	**FPS**
Wu et al. (2011)	Images	12	12	4	89.4%	N/A
Weinland et al. (2006)	Silhouettes	11	10	5	93.3%	N/A
Holte et al. (2012)	Images	13	12	5	100%	N/A
Cherla et al. (2008)	Silhouettes	13	N/A	4	80.1%	20
Weinland et al. (2010)	Images	11	10	5	83.5%	∼500
Our method	Silhouettes	11	12	5	85.9%	26

different actors lie in speed, shape and motion. As shown in table 4, our approach clearly outperforms latest results on both versions of the MuHAVi dataset. As seen in the results from Singh et al. (2010) and Cheema et al. (2011), this test presents a higher difficulty and the improvements achieved by our proposal constitute a significant benefit.

Last but not least, we compared the results obtained on the IXMAS dataset which presented a much higher degree of difficulty due to its increased number of actions, actors and views, as well as the different orientations that the subjects chose with respect to the cameras. Table 5 shows a comparison with other multi-view human action recognition approaches. The number of action classes, actors and views have been detailed because these vary among the approaches. Wu et al. (2011) obtained their highest rate excluding camera 4, whereas Cherla et al. (2008) excluded the top-view camera and reorganised the 4 side views into 6 viewing angles in order to achieve view consistency. Recently, Holte et al. (2012) achieved perfect recognition on this dataset relying on 4D spatio-temporal interest points. Nonetheless, the published recognition rates decrease when searching for methods which prove to be suitable for real-time applications. Once again, our method shows to be superior when regarding both action recognition accuracy and speed.

It can be seen that the improvements achieved for the MuHAVi dataset are more significant, and this is directly related to the quality of the input data. The silhouettes from the Weizmann and IXMAS datasets have been automatically extracted through background subtraction techniques. For this reason, the results present noise and incompleteness. Although, real-time silhouette extraction of an acceptable quality can be performed (Horprasert et al., 1999; Kim et al., 2005), silhouettes of a substantial higher quality can be obtained by recent advances in depth sensors which are able to apply markerless human pose recognition in real-time (Shotton et al., 2011). Furthermore, as the employed feature relies on the raw contour data and therefore presents sensitivity to these type of errors, image filters as border smoothing could be applied; or a more robust feature proposal could be used.

7. Conclusion and Discussion

In this paper, we have presented a human action recognition approach based on sequences of key poses. The human silhouette obtained, for instance, with background subtraction is used as initial input. The silhouette's contour leads to the used pose representation, by means of a distance signal feature which, in conjunction with the model learning approach and the action classification, shows to be a highly efficient technique. Accurate recognition results are obtained without compromising the method's suitability for real-time applications.

In contrast to exemplar-based methods, choosing a key pose-based approach leads to a simplified classification process in which the number of reference patterns is drastically reduced and noisy examples are filtered. The sequences of key poses allow us to model the long-term temporal evolution involved in action performances. Since the key poses themselves are non-temporal, introducing the temporal relationship between them at a superior level allows a higher semantic richness and improves classification with respect to strictly non-temporal key pose-based methods. Finally, an appropriate and efficient sequence matching algorithm, like DTW, enables to successfully classify sequences with inconsistent time scales. As section 6 shows, the presented method returns highly promising results on publicly available datasets, deals with both single- and multi-view scenarios successfully, and is especially robust to different ways in which actions are performed by different actors.

However, when considering sequences of key poses, we assume that the temporal order is always the same, limitation that could be overcome with the use of probabilistic graphical models like HMM. Moreover, as our method does not take into account location or optical flow, the system would have difficulty in distinguishing, for instance, walking forwards from walking backwards, because the involved poses and their relation are nearly identical. Other future lines include evaluating our method using images with occlusions and recognising a *null* or *unkown* action class which defines the normal human behaviour. The latter could be classified based on the distances to the learned action classes. If none of them is a good match, the *unknown* action class can be hit. Finally, view-invariance is not taken into account and different subject orientations need to be learned explicitly.

Acknowledgements.

This work has been partially supported by the Spanish Ministry of Science and Innovation under project "Sistema de visión para la monitorización de la actividad de la vida diaria en el hogar" (TIN2010-20510-C04-02) and by the European Commission under project "caring4U - A study on people activity in private spaces: towards a multisensor network that meets privacy requirements" (PIEF-GA-2010-274649). Alexandros Andre Chaaraoui acknowledges financial support by the Conselleria d'Educació, Formació i Ocupació of the Generalitat Valenciana (fellowship ACIF/2011/160). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Aggarwal, J., Ryoo, M., 2011. Human activity analysis: A review. ACM Comput. Surv. 43, 16:1–16:43.

Baysal, S., Kurt, M., Duygulu, P., 2010. Recognizing human actions using key poses, in: Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 1727 –1730.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., 2005. Actions as space-time shapes, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, pp. 1395 –1402 Vol. 2.

Bobick, A., Davis, J., 2001. The recognition of human movement using temporal templates. Pattern Analysis and Machine Intelligence, IEEE Transactions on 23, 257 –267.

Bradski, G., 2000. The OpenCV Library. Dr. Dobb's Journal of Software Tools .

Bremond, F., 2007. Scene Understanding: perception, multi- sensor fusion, spatio-temporal reasoning and activity recognition. Ph.D. thesis. Université de Nice-Sophia Antipolis.

Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F., 2012. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Expert Systems with Applications 39, 10873 – 10888.

Cheema, S., Eweiwi, A., Thurau, C., Bauckhage, C., 2011. Action recognition by learning discriminative key poses, in: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 1302 – 1309.

Cherla, S., Kulkarni, K., Kale, A., Ramasubramanian, V., 2008. Towards fast, view-invariant human action recognition, in: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on, pp. 1 –8.

Dedeoğlu, Y., Töreyin, B., Güdükbay, U., Çetin, A., 2006. Silhouette-based method for object classification and human action recognition in video, in: Huang, T., Sebe, N., Lew, M., Pavlovic, V., Kölsch, M., Galata, A., Kisacanin, B. (Eds.), Computer Vision in Human-Computer Interaction.

Springer Berlin / Heidelberg. volume 3979 of *Lecture Notes in Computer Science*, pp. 64–77.

Eweiwi, A., Cheema, S., Thurau, C., Bauckhage, C., 2011. Temporal key poses for human action recognition, in: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 1310 – 1317.

Fathi, A., Mori, G., 2008. Action recognition by learning mid-level motion features, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1 –8.

Hernández, J., Montemayor, A., José Pantrigo, J., Sánchez, A., 2011. Human action recognition based on tracking features, in: Ferrández, J., Álvarez Sánchez, J., de la Paz, F., Toledo, F. (Eds.), Foundations on Natural and Artificial Computation. Springer Berlin / Heidelberg. volume 6686 of *Lecture Notes in Computer Science*, pp. 471–480.

Holte, M., Chakraborty, B., Gonzalez, J., Moeslund, T., 2012. A local 3-d motion descriptor for multi-view human action recognition from 4-d spatio-temporal interest points. Selected Topics in Signal Processing, IEEE Journal of 6, 553–565.

Holte, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B., 2011. Human action recognition using multiple views: a comparative perspective on recent developments, in: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, ACM, New York, NY, USA. pp. 47–52.

Horprasert, T., Harwood, D., Davis, L., 1999. A statistical approach for real-time robust background subtraction and shadow detection, in: IEEE ICCV, pp. 256–261.

Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 34, 334 –352.

İkizler, N., Duygulu, P., 2007. Human action recognition using distribution of oriented rectangular patches, in: Elgammal, A., Rosenhahn, B., Klette, R. (Eds.), Human Motion Understanding, Modeling, Capture and Animation. Springer Berlin / Heidelberg. volume 4814 of *Lecture Notes in Computer Science*, pp. 271–284.

Juan, L., Gwun, O., 2009. A Comparison of SIFT , PCA-SIFT and SURF. International Journal of Image Processing (IJIP) 3, 143 – 152.

Kadir, T., Brady, M., 2003. Scale saliency: a novel approach to salient feature and scale selection, in: Visual Information Engineering, 2003. VIE 2003. International Conference on, pp. 25 – 28.

Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L., 2005. Real-time foreground-background segmentation using codebook model. Real-Time Imaging 11, 172 – 185. Special Issue on Video Object Processing.

Kjellström (Sidenbladh), H., 2011. Contextual action recognition, in: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (Eds.), Visual Analysis of Humans. Springer London, pp. 355–376.

Laptev, I., 2005. On space-time interest points. International Journal of Computer Vision 64, 107–123.

Martinez-Contreras, F., Orrite-Urunuela, C., Herrero-Jaraba, E., Ragheb, H., Velastin, S., 2009. Recognizing human actions using silhouette-based hmm, in: Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on, pp. 43 –48.

Ángeles Mendoza, M., Pérez de la Blanca, N., 2007. Hmm-based action recognition using contour histograms, in: Martí, J., Benedí, J., Mendonça, A., Serrat, J. (Eds.), Pattern Recognition and Image Analysis. Springer Berlin / Heidelberg. volume 4477 of *Lecture Notes in Computer Science*, pp. 394–401.

Moeslund, T.B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. 104, 90–126.

Oikonomopoulos, A., Patras, I., Pantic, M., 2005. Spatiotemporal salient points for visual recognition of human actions. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 36, 710 –719.

Poppe, R., 2010. A survey on vision-based human action recognition. Image and Vision Computing 28, 976 – 990.

Saghafi, B., Rajan, D., 2012. Human action recognition using pose-based discriminant embedding. Signal Processing: Image Communication 27, 96 – 111.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: a local svm approach, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, pp. 32 – 36 Vol.3.

Scovanner, P., Ali, S., Shah, M., 2007. A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th international conference on Multimedia, ACM, New York, NY, USA. pp. 357–360.

Shao, L., Chen, X., 2010. Histogram of body poses and spectral regression discriminant analysis for human action categorization, in: British Machine Vision Conference (BMVC), Aberystwyth, UK.

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

Singh, S., Velastin, S., Ragheb, H., 2010. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods, in: Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, pp. 48 –55.

Suzuki, S., be, K., 1985. Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing 30, 32 – 46.

Thurau, C., Hlaváč, V., 2007. *n*-grams of action primitives for recognizing human behavior, in: Kropatsch, W., Kampel, M., Hanbury, A. (Eds.), Computer Analysis of Images and Patterns. Springer Berlin / Heidelberg. volume 4673 of *Lecture Notes in Computer Science*, pp. 93–100.

Tran, D., Sorokin, A., 2008. Human activity recognition with metric learning, in: Forsyth, D., Torr, P., Zisserman, A. (Eds.), Computer Vision ECCV 2008. Springer Berlin / Heidelberg. volume 5302 of *Lecture Notes in Computer Science*, pp. 548–561.

Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O., 2008. Machine recognition of human activities: A survey. Circuits and Systems for Video Technology, IEEE Transactions on 18, 1473 –1488.

Wang, J., She, M., Nahavandi, S., Kouzani, A., 2010. A review of vision-based gait recognition methods for human identification, in: Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on, pp. 320 –327.

Weinland, D., Özuysal, M., Fua, P., 2010. Making action recognition robust to occlusions and viewpoint changes, in: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision ECCV 2010. Springer Berlin / Heidelberg. volume 6313 of *Lecture Notes in Computer Science*, pp. 635–648.

Weinland, D., Ronfard, R., Boyer, E., 2006. Free viewpoint action recognition using motion history volumes. Comput. Vis. Image Underst. 104, 249–257.

Weinland, D., Ronfard, R., Boyer, E., 2011. A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Underst. 115, 224–241.

Wu, C., Khalili, A.H., Aghajan, H., 2010a. Multiview activity recognition in smart homes with spatio-temporal features, in: Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras, ACM, New York, NY, USA. pp. 142–149.

Wu, X., Shi, Z., Zhong, Y., 2010b. Detailed analysis and evaluation of keypoint extraction methods, in: Computer Application and System Modeling (ICCASM), 2010 International Conference on, pp. V2–562 –V2–566.

Wu, X., Xu, D., Duan, L., Luo, J., 2011. Action recognition using context and appearance distribution features, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 489 –496.