González, L., Velastin, S.A y Acuña, G. (2018). Silhouette-based human action recognition with a multi-class support vector machine. In *9th International Conference on Pattern Recognition Systems (ICPRS 2018).*

DOI: https://doi.org/10.1049/cp.2018.1290

Silhouette-based human action recognition with a multi-class support vector machine

Luis González*, Sergio A Velastin[+], Gonzalo Acuña*

*Universidad de Santiago de Chile, Santiago, Chile
[+]University Carlos III Madrid, Spain, Cortexica Vision Systems Ltd.
and Queen Mary University of London, UK

Keywords: Bag of key poses, MuHAVi, SVM, Computer vision, Human Action Recognition

Abstract

Computer vision systems have become increasingly popular, being used to solve a wide range of problems. In this paper, a computer vision algorithm with a support vector machine (SVM) classifier is presented. The work focuses on the recognition of human actions through computer vision, using a multi-camera dataset of human actions called *MuHAVi*. The algorithm uses a method to extract features, based on silhouettes. The challenge is that in *MuHAVi* these silhouettes are noisy and in many cases include shadows. As there are many actions that need to be recognised, we take a multiclass classification approach that combines binary *SVM* classifiers. The results are compared with previous results on the same dataset and show a significant improvement, especially for recognising actions on a different view, obtaining overall accuracy of 85.5% and of 93.5% for leave-one-camera-out and leave-one-actor-out tests respectively.

1 Introduction

Ageing statistics both in developing and developed countries indicate a tendency of increases in average age and, consequently, a growth of the elderly population. Given also the trends on family "nuclearisation" and the reductions in extended family support, many of the elderly also prefer to live independently, when possible, so technologies that enable assisted living become desirable. In particular, the use of automated video monitoring has the potential to increase levels of safety and security while maintaining acceptable privacy by reducing human observation, limiting it only to cases of emergencies. Furthermore, added to technologies such as the Internet of Things, smartphones and global connectivity, the promise of ubiquitous intelligent homes becomes foreseeable.

Some countries such as Spain are already working on this (e.g. the AAL Joint Program), having 600 million euros between 2008 and 2013, being the first major investment in research in this line. There is also a framework of the European Union (EU) called "Horizon 2020", which points to the development of technology that allows assisted monitoring in intelligent environments to better support personnel in care facilities for the elderly. Central to those efforts is to emulate the ability of human beings of understanding what a person is doing (on their own, with others and with objects in their environment). In the context of computer vision, this has been called "Human Action Recognition" (HAR). It is important to point out that this ability is not only useful in the context of assisted living but also in many other areas such as public space surveillance, retail services, media handling (e.g. to automatically summarise video material), entertainment and so on. The last decade has seen many advances, but there is still significant progress to be made before solutions can be used in real environments. The use of standard bench marking datasets such as *MuHAVi* is part of the effort to make progress. The investigation of the state of the art on the use of this dataset,revealed the need to compare the features proposed by [1], but using an SVM classifier, versus the method proposed by [2] to have identify the better method in future works. Therefore, this paper evaluates the computer vision algorithm first proposed in [1], originally tested with the MuHAVi-MAS [3] dataset, a small dataset of human actions where all silhouettes have been manually defined and therefore they are as noise-free as possible. In this paper we use a more realistic set of data known as MuHAVi-uncut [4] which contains much longer sequences also involving more actors and more camera views. In this case, silhouettes have been obtained automatically by a foreground estimation algorithm and therefore it contains noise and shadows likely to be encountered in a real application. Thus, here we assess how robust the action recognition algorithm is to these real imperfections. In addition, we propose the use of a multiclass Support Vector Machine classifier for the final stage in the process. The MuHAVi dataset contains 17 actions (table 1), carried out by 7 different people and seen from 8 different cameras[3], as shown in figure 1 (these are from MuHAVi-MAS that were segmented manually).

The rest of the paper is organised as follows: Section 2 presents a brief review, Section 3 describes the vision algorithm used, dataset and classifier. Section 4 reports the results obtained and section 5 concludes the paper.

2 Related works

A wide range of different algorithms and systems have been proposed for Human Action Recognition. As the literature is

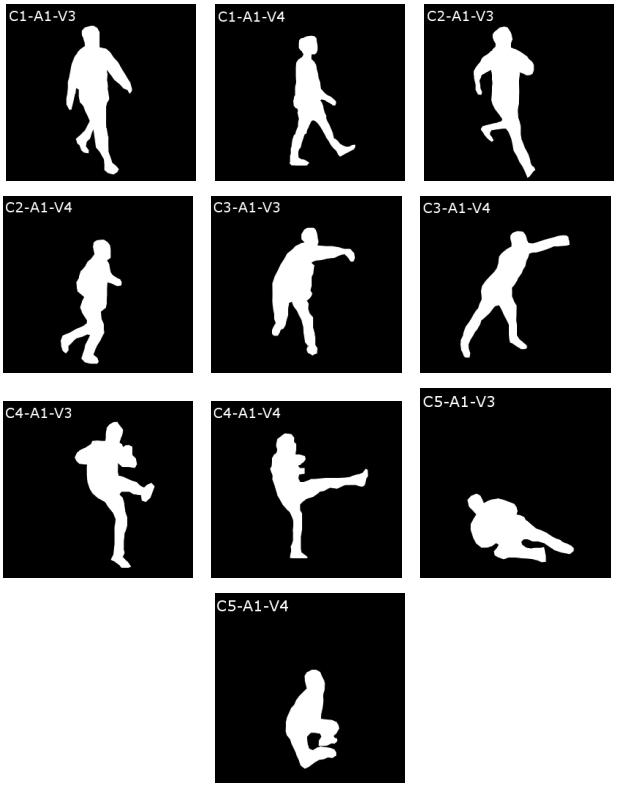


Figure 1. Images of some actions of Muhavi dataset.

getting rather large, here we focus on works that have experimented with the same dataset. The interested reader can consult a good review of the subject in [5],[6] and [7].

A multiview approach was proposed in [8], introducing the concept of *bag of key poses*, which serves to describe that an action can be represented by a sequence of key pose features extracted from silhouettes from different people and different views. A *clustering* process inspired by bag of words is used to group pose sequences that can later be used to classified previously unseen actions.

Returning to the idea of *bag of key poses* [1] integrates the K-means algorithm as a classifier, obtaining much better computational performance results than previous algorithms while having better accuracy and speed. This is the algorithm we will test and expand in this paper.

In a different approach, [2] presents a human action representation, consisting of *Motion History Images* (MHIs), which represent how often (in a given temporal window) a given pixel has been seen as foreground, thus encoding the amount of "motion" (or presence) of that pixel, hypothesising that that is related to actions. Then, it uses *Histograms of Oriented Gradients* (HOG) applied to the MHIs as features that can be used for training. A simple approach based on *Nearest Neighbour* (NN) is used as a classifier to determine the class of a previously unseen action sample by finding a trained sample to which it is the closest.

Action Class	Action
C1	WalkTurnBack
C2	RunStop
C3	PullHeavyObject
C4	PickupThrowObject
C5	Punch
C6	Kick
C7	ShotGunCollapse
C8	WalkFall
C9	LookInCar
C10	CrawlOnKnees
C11	WaveArms
C12	DrawGraffiti
C13	JumpOverFence
C14	DrunkWalk
C15	SmashObject
C16	JumpOverGap
C17	ClimbLadder

Table 1. Actions available in the MuHavi dataset.

3 Proposed approach

As pointed out above, in this paper we use the approach first proposed in [1] and test it under the much more stringent conditions of MuHaVi-uncut. For the sake of completeness, we outline part of the algorithm here.

3.1 Image segmentation and noise reduction

As outlined earlier, MuHAVi-uncut has been generated using what was at the time state-of-the-art foreground estimation using mixture of Gaussians. Although the method could also estimate shadows, not all shadows are removed and the silhouettes are also inherently noisy. Therefore, as proposed by [2] an image filtering process using a 15x15 median filter is performed, as illustrated in figure 2. Using the same filter allows to directly compare results with [2].

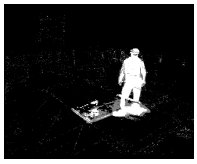
(a)  (b) 

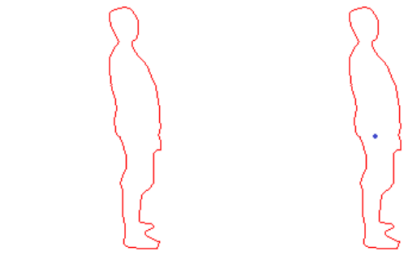
Figure 2. Images (a) without filter (b) with filter.

3.2 Feature extraction

A relatively fast feature extraction is proposed in [9], which consists of a series of steps as described below:

This part of the algorithm follows what was proposed in [10]. In each frame of a given action sequence, from the silhouette a contour of N points $P = p_1, p_2, ..., p_N$, where $p_i = (x_i, y_i)$, is calculated. If N is sufficiently large, its choice is not too critical. From these N points the centroid $C = (x_c, y_c)$ is calculated (please also see figure 3) by:

$$x_c = \frac{\sum_{i=1}^{N} x_i}{N}, y_c = \frac{\sum_{i=1}^{N} y_i}{N} \quad (1)$$



(a) (b)

Figure 3. (a) Silhouette (b) with calculated centroid

Next, the silhouette (centred at C) is inscribed by the smallest circumference and divided up into n equal segments as shown in figure 4.

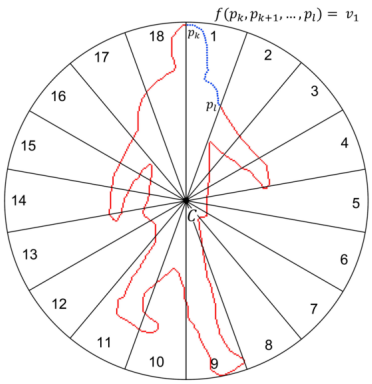


Figure 4. Silhouette inscribed by smallest circumference and division of the same.

A list is computed $B = [(\alpha_1, b_1), ..., (\alpha_n, b_n))]$ in a clockwise direction, and the number of elements on list B is called nB (set to 18 in our experiment, so the circle is divided up in equal segments of 20 degrees), where each element α_i represents an arc (in radians) of the corresponding circle.

$$\alpha_i = \begin{cases} arccos(\frac{y_i - y_c}{d_i}) \cdot \frac{180}{\pi} & \text{if } x_i \geq 0 \\ 180 + arccos(\frac{y_i - y_c}{d_i}) \cdot \frac{180}{\pi} & \text{otherwise} \end{cases} \quad (2)$$

$$b_i = \left\lceil \frac{nB \cdot \alpha_i}{360} \right\rceil, \forall_i \in [1...nB] \quad (3)$$

With reference to figure 5, a maximum distance is computed from the furthest and closest points from the centroid:

$$f_{range}(p_k, p_{k+1}, ..., p_l) = max(d_k, d_{k+1}, ..., d_l) - min(d_k, d_{k+1}, ..., d_l)$$

Where d is the distance between each point p and the centroid C of the circumference, the distance is calculated with the equation:

$$d_k = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (4)$$

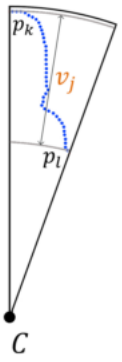


Figure 5. Calculating the distance of each point in the silhouette with respect to the centroid C.

Then, for each circle segment j, a value v_j is calculated as:

$$v_j = f_{range}(p_k, p_{k+1}, ..., p_l)/b_k, ..., b_l \quad (5)$$

which are then normalised:

$$\overline{v_j} = \frac{v_j}{\sum_{i=1}^{nB} v_i} \quad (6)$$

Finally, the feature vector $\overline{V}$ is formed by all the segment components:

$$\overline{V} = (v_1, v_2, .., v_n) \quad (7)$$

3.3 Action Recognition

We use supervised training using a one-against-one multiclass approach with RBF (Radial Basis Function) SVM (Support Vector Machine) classifiers. RBF is known for being better at capturing non-linear relations in the data. One-against-one decomposes the problem of k classes into $k(k-1)/2$ binary problems, where all the possible one-to-one combinations between all classes are created. Each new data is evaluated for all created classifiers, obtaining one vote for each winning class in each case, then the final result is the class with the most votes. In this case 136 binary classifiers are created that represent each combination of pairs of classes.

No additional data normalisation is required, as that is done by the feature extraction process described earlier.

The RBF kernel needs two parameters, C and γ, which are application dependent. We use the well accepted exhaustive grid approach that finds these parameters using cross-validation on a testing dataset and choose the combination that produces the smallest mean classification error. The testing dataset (and therefore the computation of optimal SVM parameters) depends on the evaluation being performed.

4 Results

We follow the LOCO and LOAO tests suggested by [2], as described below.

4.1 LOCO (Leave-one-camera-out)

In this test, one camera is left out and the algorithm is trained with the remaining cameras and tested with the camera that was left out. This evaluates view dependence. This is done for all seven cameras (one camera in MuHAVi-uncut is not usable as it provides much less data).

Table 2 shows the mean accuracy and standard deviation results of the human actions recognition by means of the LOCO test and compares it with the results reported in [2], which it outperforms significantly in all cases (its overall average accuracy is 85.52% vs. 49.28%), indicating its better resilience to point of view. It is also noted that this method is less variable from one camera to another.

Action	Mean (%)	St. dev.(%)	[2]
1	77,11	15,07	26,2
2	94,46	13,96	61,6
3	85,26	14,23	64,7
4	89,95	9,93	54,2
5	91,37	12,14	39,7
6	85,91	15,67	55,4
7	87,19	14,64	64,3
8	78,07	22,74	1,4
9	81,17	14,67	46,9
10	90,10	11,16	48,5
11	85,59	15,36	59,6
12	88,13	14,07	12,2
13	80,24	13,81	77,6
14	82,60	10,74	52,4
15	81,51	17,46	87,1
16	93,86	9,76	51,7
17	81,24	18,13	34,7
Overall	85,52	-	49,28

Table 2. LOCO Mean and standard deviation of the recognition for each of the 17 actions.

4.2 LOAO (Leave-one-actor-out)

In this test, one actor is left out and the algorithm is trained with the remaining actors and tested for the actor that was left out. This evaluates actor dependence. This is done for all the actors.

Table 3 shows the mean accuracy and standard deviation results of the human actions recognition for the LOAO test and compares it with the results reported in [2], which it outperforms in many cases (its overall average accuracy is 93.52% vs. 83.11%). As expected, accuracy is higher and variation smaller than for the LOCO test in all cases as this test is less stringent (there is less variability between actors than between camera views).

Action	Mean (%)	St. dev.(%)	[2]
1	92,67	3,77	90,5
2	94,37	4,16	98,2
3	95,87	4,30	83,0
4	92,14	3,66	60,7
5	94,21	3,65	65,6
6	91,73	4,10	71,9
7	94,63	4,51	95,7
8	93,60	4,96	95,2
9	93,52	4,89	76,9
10	93,10	4,24	98,0
11	92,80	3,96	87,5
12	94,45	4,09	34,7
13	93,27	3,89	91,8
14	95,10	4,37	87,7
15	93,32	4,26	97,3
16	93,13	4,65	82,3
17	91,85	4,19	95,9
Overall	93,52	-	83,111

Table 3. LOAO Mean and standard deviation of the recognition for each of the 17 actions.

5 Conclusion

In this paper we have shown improvements on human action recognition on the MuHAVi-uncut dataset, particularly challenging because of salt-pepper noise and shadows. We have used silhouette features as proposed by [1] in combination with a multiclass one-against-one SVM classifier. This combination has proven particularly good compared to previous results [2], in the case of a LOCO test reaching an average accuracy of 85.5% and of 93.5% for LOAO. Performance has also been observed to be more stable to changes than the compared algorithm. Although feature extraction is relatively fast, the proposed method is slow to fine-tune to find optimal SVM parameters. Future work will look at datasets with more complex multi-person actions and consider deep learning methods (although they might be tricky to work with relative small datasets and so data augmentation techniques would need to be explored). We also want to explore better techniques to extract silhouettes that do not rely on foreground/background separation so that they can be use in non-static cameras.

Acknowledgements

Sergio A Velastin has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 600371, el Ministerio de Economía, Industria y Competitividad (COFUND2013-51509) el Ministerio de Educación, cultura y Deporte (CEI-15-17) and Banco Santander.

References

[1] A. A. Chaaraoui and F. Flórez-Revuelta, "Adaptive human action recognition with an evolving bag of key poses," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 2, pp. 139–152, 2014.

[2] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2d motion templates based on mhis and their hog description," *Iet Computer Vision*, vol. 10, no. 7, pp. 758–767, 2016.

[3] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS '10, (Washington, DC, USA), pp. 48–55, IEEE Computer Society, 2010.

[4] J. Sepúlveda and S. Velastin, "F1 Score Assesment of Gaussian Mixture Background Subtraction Algorithms Using the MuHAVi Dataset," *6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15)*, pp. 8 (6 .)–8 (6 .), 2015.

[5] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[6] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[7] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[8] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "An efficient approach for multi-view human action recognition based on bag-of-key-poses," in *Human Behavior Understanding* (A. A. Salah, J. Ruiz-del Solar, Ç. Meriçli, and P.-Y. Oudeyer, eds.), (Berlin, Heidelberg), pp. 29–40, Springer Berlin Heidelberg, 2012.

[9] A. A. Chaaraoui, "Vision-based Recognition of Human Behaviour for Intelligent Environments," 2014.

[10] S. Suzuki and K. Abe, "Topological Structural Analysis of Digitalized Binary Images by Border Following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.