# Action Classification using the Average of Pose Changes

*Janto F. Dreijer and Ben M. Herbst*

Applied Mathematics
Stellenbosch University
Private Bag X1, 7602 Matieland, South Africa
{janto, herbst}@dip.sun.ac.za

## Abstract

This article briefly discusses some of our ongoing work on the problem of human action recognition. We evaluate a simple and intuitive technique, based on the changes in human pose, against publicly available behaviour datasets. We achieve results comparable to many other state of the art techniques, while also being much simpler and potentially faster.

## 1. Introduction

### 1.1. Problem Statement

Action recognition has interesting applications such as detecting falls [1] and indexing movies [2], and has received increased attention in recent years.

Our goal is to create a system capable of classifying actions in a live video stream, using lightweight techniques. We evaluate a simple and intuitive technique, based on the changes in human pose, against publicly available behavior datasets. We achieve results comparable to many other state of the art techniques, while also being much simpler and potentially faster.

We find that the average of pose changes are surprisingly discriminative for these datasets and conclude that this simple approach is sufficient for action types that have stereotypical poses, at least while the library of poses remain small.

### 1.2. Related Work

State of the art approaches to action recognition can roughly be grouped into three: Pose transition models, collections of quantized space-time interest points ("bag of features") and template images. Also of interest is the motion of key points, which is often used in gesture recognition applications.

#### *1.2.1. Pose transitions*

In this approach actions are regarded as transitions over a sequence of observations of body pose. Individual poses are usually represented as a location in a feature space and a model constructed of the motion through this space. Actions are then classified based on how well they fit the learnt model.

Pose observations have been encoded in terms of their contours [3, 4], optical flow [5, 6], geometric moments [7], and various others. These transitions are then represented in graphical models such as hidden Markov models [6] and Monte Carlo random walks through graphs [4].

It should be noted that separation of the pose from the background is often not ideal, and therefore may introduce significant noise in the pose encodings.

#### *1.2.2. Bag of features*

This approach is inspired by recent advances made in recognising generic objects and textual understanding. Actions are seen as collections of specific space-time interest points or cubelets. These techniques involve extracting interesting features from the space-time volume [8, 9, 10]. These discrete feature points are usually summarized in the form of multidimensional histograms. Segments of videos are then compared via a comparison of their histograms [8, 9].

For example, Laptev *et al.* [11] represent interesting points, found with a Harris corner detector, by a Histogram of Gradients descriptor. These features are quantized into words using k-means clustering. Video segments are then classified based on their histogram of words using support vector machines.

Others also consider the space-time volume, but instead try to characterize its properties by using, for example, the solution to the Poisson equation [2].

#### *1.2.3. Motion of points*

Lange *et al.* [12] have investigated the human ability to recognise a moving human figure from no more than a few key points. They found a high correlation between their simulation results and psychophysical data. This news might be promising to those that believe the path of various body parts such as hand, head and feet, may be a major component in interpreting human behaviour.

Much work has been done on hand-gesture recogni-

Figure 1: Average of poses (AME) for Weizmann dataset (source: [16])

tion in this regard. Bobick and Wilson [13] represent the trajectory of a hand as a sequence of fuzzy states in a configuration space to capture the repeatability and time-scale variability of a gesture. Nam and Wohn [14] described a hidden Markov based method for recognising the space-time hand movement pattern of various basic gestures by first projecting 3D motion data onto a 2D plane. They then employ a chain encoding scheme and construct a HMM network from simpler left-to-right discrete HMMs.

Song *et al.* [15] have addressed the problem of detecting humans from their motion pattern. They model the joint position and velocity probability density function of triplets of moving features.

#### *1.2.4. Template based*

Techniques such as Average Motion Estimates (AMEs, [16] ) represent the average of a subject's poses as a single image. Although this is much simpler than the above methods, Lu *et al.* [16] reported surprisingly high performance on the Weizmann dataset. AMEs have, however, only been tested on this relatively simple dataset, partly because poses need to be made translation invariant first.

AMEs emphasize body parts that do not vary (see Figure 1). Indeed, although AMEs represent the motion with regard to the image background, it does not represent the changes in the pose itself.

Davis and Bobick [17] have examined motion-energy images (MEI) and motion-history images (MHI). MEIs are binary images which represent where motion has occurred spatially and MHIs are grayscale images where intensity indicates recent motion. Examples are shown in Figures 2 and 3. MEIs and MHIs are made scale and translation invariant by comparing their Hu moments [18] when classifying actions.

An attractive property of template techniques is that motion can be represented by a single intuitive image. They do, however, also rely on tracking and segmentation of a subject from its background.

### 1.3. Practical considerations

There are two important factors that have to be taken into account when designing a system capable of performing action recognition on live video streams: amount of processing resources and type of background information

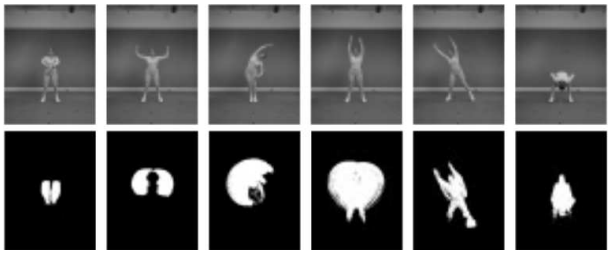

Figure 2: Examples of MEIs for aerobic exercises (source: [17])

Figure 3: Examples of MHIs for waving and crouching (source: [17])

available.

If realtime performance is required, a lightweight strategy has to be used, especially when multiple cameras are involved. Lightweight algorithms allow one to process input from multiple cameras with a single server or push the action recognition algorithm onto smart cameras that typically employ weaker processors.

There are, however, few reports that provide the computational costs involved with existing techniques that would make them applicable to realtime action classification. Those that do report their costs are, in the best cases, in the order of a frame per second for low resolutions on modern consumer hardware [2, 5, 6]. We assume that those that do not report on their efficiency are much slower.

A sophisticated background model is also not always available, depending on the application. It might be good enough to separate subjects, but not to provide error-free body silhouettes. We assume some degree of segmentation of a subject from its background and sufficient inter-subject separation can be obtained.

## 2. Our Approach

We have investigated various background models, but have decided to use a naive technique to demonstrate our action classifier. Because the datasets (discussed later) contain only one subject we do not require a tracker or inter-subject separation that may be needed in real world applications such as surveillance.

By assuming any motion within the video is primarily of the subject we can use a simple technique to determine the *changes in the subject's pose*, i.e. consecutive frames are subtracted and the difference thresholded:

$$\Delta'_{pose}(n) = |I(n+1) - I(n)| > k \qquad (1)$$

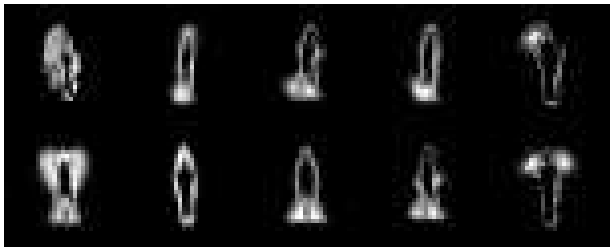

Figure 4: Average of pose changes for Weizmann dataset.

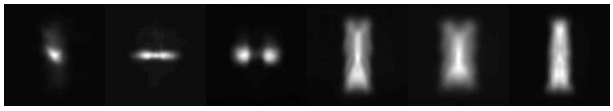

Figure 5: Average of pose changes for KTH dataset.

where $I(n)$ is a specific frame in the sequence and $k = 20/255$ in pixel intensity. We also apply median filtering to remove minor noise.

To obtain a translation and scale invariant representation of the change in pose, we shift and scale the contents of $\Delta'_{pose}(n)$ so its immediate bounding box is centered and encompasses the entire image. We call this new image $\Delta_{pose}(n)$.

By taking the average of *changes* in pose in a video, i.e.

$$T_{video} = \frac{1}{N} \sum_{n=0}^{N-1} \Delta pose(n), \qquad (2)$$

we can obtain Average Pose Changes as shown in Figures 4 and 5.

For classification we determine a template for each video in the testing set. We represent each template as a vector by concatenating its rows. We then estimate a query video's associated action through either a k-nearest neighbour lookup or using a Support Vector Machine (SVM).

Our approach is related to AMEs that represent the average pose (and indirectly including some motion information) and MEIs that are a binary indication of motion. However it is our opinion that it pays to emphasize exactly those body parts that vary and how often they vary.

Note the difference between Figures 1 and 4. In our technique changing body parts are emphasised instead of the static body. We believe that this is an important distinction for two reasons:

- the pose *change* is obtained through simple subtraction and thresholding between frames and is thus, unlike the pose itself, readily available,
- the AME cannot adequately address actions where the average pose may be the same, but the amount of activity of body parts are important.

| | bend | jack | jump | pjump | run | side | skip | walk | wave1 | wave2 |
|---|---|---|---|---|---|---|---|---|---|---|
| bend | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| jack | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| jump | 0 | 0 | 89 | 0 | 0 | 0 | 11 | 0 | 0 | 0 |
| pjump | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| run | 0 | 0 | 0 | 0 | 90 | 0 | 10 | 0 | 0 | 0 |
| side | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| skip | 0 | 0 | 0 | 0 | 10 | 0 | 90 | 0 | 0 | 0 |
| walk | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| wave1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 |
| wave2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

96.9% class average

Table 1: Confusion matrix for Weizmann dataset using pose change templates and nearest neighbour for classification. Provided silhouettes used as pose images.

## 3. Evaluation

### 3.1. Datasets

We test the average pose change templates against the Weizmann and KTH datasets.

The Weizmann dataset [2] contains examples of 10 actions performed by 9 subjects giving a total of just more than 90 videos. Segmented translation invariant silhouettes are provided with this dataset. As many have achieved near perfect results on this dataset, we only use it as a demonstration of acceptable results, rather than a measure of relative accuracy.

The KTH dataset [19] contains examples of 6 actions performed by 25 subjects, totaling 593 videos. These videos where designed to contain significant camera motion and zooming effects. Since the backgrounds are relatively uniform, it is easy to isolate the subject from the background.

We used similar cross-validation techniques as used in other studies: leave-one-person-out cross validation (LOOCV) for the Weizmann dataset and three-way cross validation for the KTH videos.

### 3.2. Discussion

We used the differences in provided foreground as pose changes in one test (Table 1), and immediate frame subtraction in another (Table 2). The near perfect results that were achieved on the Weizmann dataset, are similar to those of the AMEs [16]. Table 2 shows that even without a sophisticated background model, significant performance can still be achieved with an immediate foreground detection scheme.

Tables 3 and 4 show the performance against the KTH dataset using a nearest neighbour classifier and linear SVM. Actions with similar poses (jogging and walking, jogging and running) account for most of the loss in performance. It should be reiterated that no foreground mask was provided with the KTH dataset and hence is to be compared to Table 2 and not 1.

The results of some related studies are reported in Table 5. Note that many of these use different cross vali-

| | bend | jack | jump | pjump | run | side | skip | walk | wave1 | wave2 |
|---|---|---|---|---|---|---|---|---|---|---|
| bend | 89 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 |
| jack | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| jump | 0 | 0 | 78 | 0 | 0 | 0 | 22 | 0 | 0 | 0 |
| pjump | 0 | 0 | 0 | 89 | 0 | 0 | 0 | 0 | 11 | 0 |
| run | 0 | 0 | 0 | 0 | 90 | 0 | 10 | 0 | 0 | 0 |
| side | 0 | 0 | 0 | 0 | 0 | 89 | 11 | 0 | 0 | 0 |
| skip | 0 | 0 | 10 | 0 | 30 | 10 | 50 | 0 | 0 | 0 |
| walk | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 90 | 0 | 0 |
| wave1 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 78 | 0 |
| wave2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

85.2% class average

Table 2: Confusion matrix for Weizmann dataset using pose change templates and nearest neighbour for classification. Pose changes extracted from videos.

| | boxing | handclapping | handwaving | jogging | running | walking |
|---|---|---|---|---|---|---|
| boxing | 85 | 8 | 1 | 2 | 1 | 4 |
| handclapping | 6 | 90 | 4 | 0 | 0 | 0 |
| handwaving | 0 | 4 | 95 | 0 | 0 | 1 |
| jogging | 0 | 0 | 0 | 77 | 13 | 9 |
| running | 1 | 0 | 0 | 20 | 78 | 1 |
| walking | 0 | 0 | 0 | 6 | 2 | 91 |

86.0% class average

Table 3: Confusion matrix for KTH dataset using pose change templates and nearest neighbour for classification.

dation techniques and are strictly not comparable. E.g. LOOCV allows one to use approximately three times more videos for training than 3-way split. Still, we can say with reasonable confidence that the accuracy of our approach is comparable to many state-of-the-art algorithms.

A few remarks are in order:

- These datasets, specifically, contain actions mostly differentiable through pose analysis alone. i.e. these actions have stereotypical poses. This is in line with the recent analysis by Weinland and Boyer [23] of the Weizmann dataset.
- Some interesting real world actions are distinguishable through pose analysis alone.
- The datasets do not adequately represent interesting actions that are different primarily due to the speed at which they are executed. Jogging vs running, falling down vs sitting/bending, handing over an item vs punching another person in the stomach, are actions that contain similar poses, but should be treated as different actions due to their speed. This is especially important for applications such as fall detection, as with higher speeds comes higher risk of injury.

### 3.3. Efficiency

Relatively little attention has been given by others to make existing algorithms work on live video streams. Be-

| | boxing | handclapping | handwaving | jogging | running | walking |
|---|---|---|---|---|---|---|
| boxing | 88 | 8 | 1 | 0 | 0 | 2 |
| handclapping | 2 | 94 | 4 | 0 | 0 | 0 |
| handwaving | 0 | 9 | 91 | 0 | 0 | 0 |
| jogging | 1 | 0 | 0 | 76 | 10 | 13 |
| running | 2 | 0 | 0 | 15 | 79 | 4 |
| walking | 1 | 1 | 0 | 1 | 0 | 97 |

87.3% class average

Table 4: Confusion matrix for KTH dataset using pose change templates and SVM for classification.

| method | accuracy |
|---|---|
| Our method | 87.3% |
| Laptev *et al.* [11] | 91.8% |
| Rodriguez *et al.* [5] | 88.7% |
| Ahmed and Lee [6] | 88.3% |
| Wong [20] | 86.6% |
| Dollar *et al.* [21] | 85.9% |
| Niebles [10] | 81.5% |
| Schuldt [19] | 71.7% |
| Ke *et al.* [22] | 63.0% |

Table 5: Reported accuracies of related studies on the KTH dataset. Note that many of these use different cross validation techniques and, strictly speaking, are not comparable.

cause we compare the templates directly without extracting any features or moments, we gain a significant advantage in runtime speed. Ahmad and Lee [6], for example require calculating Zernike moments on 160x120 images, which take 0.69-0.82 seconds a frame (approximately 1.4fps) on their 1.7GHz machine.

Our implementation of pose change templates (using a SVM for classification), can currently run at approximately 16fps for a 400x400 video stream on a 3GHz processor.

The effects of tracking and framerate also need to be analysed. Higher frame rates will improve the detection of small motions, but will adversely affect our bounding box model. We therefore plan on using a more complex background model to determine the bounding box.

## 4. Conclusion

We have investigated a simple method of classifying human actions from a sequence of images.

Even though we have used a very simple approach, our performance is comparable to other existing techniques. At the same time, our approach holds the promise for action recognition requiring few computer resources. Several improvements can still be made to pose change templates, such as a temporal multiscale to detect actions that differ due to their speeds (e.g. running and jogging).

## 5. References

[1] H. Nait-Charif and S. J. McKenna, “Activity summarisation and fall detection in a supportive home

environment," in *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4*, (Washington, DC, USA), pp. 323–326, IEEE Computer Society, 2004.

[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2247–2253, December 2007.

[3] J. Hsieh and Y. Hsu, "Boosted string representation and its application to video surveillance," *Pattern Recognition*, vol. 41, pp. 3078–3091, October 2008.

[4] S. Xiang, F. Nie, Y. Song, and C. Zhang, "Contour graph based human tracking and action sequence recognition," *Pattern Recognition*, vol. 41, no. 12, pp. 3653–3664, 2008.

[5] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR08*, pp. 1–8, 2008.

[6] M. Ahmad and S.-W. Lee, "Human action recognition using shape and clg-motion flow from multi-view image sequences," *Pattern Recognition*, vol. 41, no. 7, pp. 2237–2252, 2008.

[7] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[8] L. Zelnik-Manor and M. Irani, "Statistical analysis of dynamic actions," *PAMI*, vol. 28, pp. 1530–1535, September 2006.

[9] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Computer Vision and Image Understanding,*, vol. 108, pp. 207–229, 12 2007.

[10] J. Niebles, H. Wang, and F. Li, "Unsupervised learning of human action categories using spatial-temporal words," *IJCV*, vol. 79, no. 3, 2008.

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2008.

[12] J. Lange, K. Georg, and M. Lappe, "Visual perception of biological motion by form: A template-matching analysis," *Journal of Vision*, vol. 6, pp. 836–849, 7 2006.

[13] A. Bobick and A. Wilson, "Configuration states for the representation and recognition of gesture," in *In International Workshop on Automatic Face and Gesture Recognition*, pp. 129–134, 1995.

[14] Y. Nam and K. Wohn, "Recognition of space-time hand gestures using hidden markov models," *In ACM Symposium on Virtual Reality Software and Technology*, 1996.

[15] Y. Song, X. Feng, and P. Perona, "Towards detection of human motion," in *In CVPR*, pp. 810–817, 2000.

[16] L. Wang and D. Suter, "Informative shape representations for human action recognition," *ICPR*, vol. 2, pp. 1266–1269, 2006.

[17] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, (Washington, DC, USA), p. 928, 1997.

[18] M.-K. Hu, "Pattern recognition by moment invariants," in *IRE*, vol. 49, p. 1428, Sep. 1961.

[19] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3*, (Washington, DC, USA), pp. 32–36, IEEE Computer Society, 2004.

[20] S. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *ICCV07*, pp. 1–8, 2007.

[21] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks*, (Washington, DC, USA), pp. 65–72, IEEE Computer Society, 2005.

[22] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *ICCV05*, pp. I: 166–173, 2005.

[23] D. Weinland and E. Boyer, "Action recognition using exemplar-based embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Anchorage), pp. 1–7, 2008.