Exemplar-based Human Action Recognition with Template Matching from a Stream of Motion Capture

Daniel Leightley ⋆1, Baihua Li[1], Jamie S. McPhee[2], Moi Hoon Yap[1], and John Darby[1]

[1]School of Computing, Mathematics and Digital Technology,
[2]School of Healthcare Science,
Manchester Metropolitan University, M1 5GD, UK.
d.leightley@ieee.org,{b.li,j.s.mcphee,m.yap,j.darby}@mmu.ac.uk
http://www.mmu.ac.uk

Abstract. Recent works on human action recognition have focused on representing and classifying articulated body motion. These methods require a detailed knowledge of the action composition both in the spatial and temporal domains, which is a difficult task, most notably under real-time conditions. As such, there has been a recent shift towards the exemplar paradigm as an efficient low-level and invariant modelling approach. Motivated by recent success, we believe a real-time solution to the problem of human action recognition can be achieved. In this work, we present an exemplar-based approach where only a single action sequence is used to model an action class. Notably, rotations for each pose are parameterised in Exponential Map form. Delegate exemplars are selected using k-means clustering, where the cluster criteria is selected automatically. For each cluster, a delegate is identified and denoted as the exemplar by means of a similarity function. The number of exemplars is adaptive based on the complexity of the action sequence. For recognition, Dynamic Time Warping and template matching is employed to compare the similarity between a streamed observation and the action model. Experimental results using motion capture demonstrate our approach is superior to current state-of-the-art, with the additional ability to handle large and varied action sequences.

Keywords: human action recognition, motion capture, exponential map, online recognition, template matching, dynamic time warping

1 Introduction

Human action recognition is an active and challenging field of research that has received wide attention in computer vision due to the number of applications. There is an ever-increasing demand for more robust, accurate and practical solutions for recognising human action online in areas such as healthcare, human-computer interaction, security and gaming.

⋆ This research is funded by the John Dalton Institute.

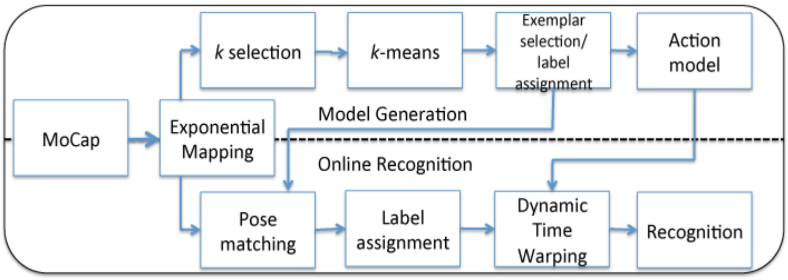


Fig. 1: Flowchart of the proposed approach. Top row denotes the process for generating an action model. Bottom row denotes the online recognition framework.

Recognising and classifying human action is a difficult task. This is in a large part due to large variations among subjects, overlap of poses between actions and data noise [1]. It is important to have a sufficient number of training samples to allow for each action to be represented effectively. Thereby, a large number of training samples are common for many recognition approaches. However, it is not always practical to utilise a large number of training samples due to the size and complexity in computing classification. Therefore, we propose the use of the exemplar paradigm as an effective approach to modelling human action. For our approach, we use the exemplar paradigm to reduce an action sequence to its most descriptive *key* elements. Elgammal *et al.* [2] sought to capture the dynamics of human gesture with an exemplar-based recognition system. The approach represented each gesture as a sequence of key exemplars. The exemplars were selected based on a non-parametric estimation framework. Then, a probabilistic framework is employed for matching testing sequences with the exemplar model. Most recently, for exemplar-based approaches, Barnachon *et al.* [3] proposed an integral histogram approach for human action recognition evaluated against motion capture data (MoCap). They extend the concept of histograms to characterise action sequences, where similar poses are clustered to characterise the action phase with multiple exemplar histograms. Finally, for recognition, they decompose a stream of MoCap into integral histograms based on Dynamic Time Warping (DTW) to compute a confidence score based on distance. There are limitations with this approach, such as manual selection of k for k-medoids, ability to classify individual poses and recognition is undertaken on a limited dataset with *known* subjects and *known* test sequences.

The use of MoCap as a data medium presents a number of challenges, namely high degrees-of-freedom (DOF) and data consisting of complex non-linear relations, singularities that are difficult to model and interpret. To handle these challenges, parametrisation of the rotations in Exponential Map form (EMP) has been proposed [4]. EMP is more robust than other forms of rotation parametrisation, such as Euler angles or Quaternions. Bregler and Malik [5] were among the first to utilise traditional Exponential Maps for gradient-based recognition of complex human action based on complete MoCap sequences. They propose a novel method for computing the product of an Exponential Map to solve a simple linear problem. Their proposed method is capable of handling multiple DOF that may contain noise, singularities and discontinuities. More recently, Taylor *et al.* [6] proposed a non-linear generative approach by pre-processing MoCap poses into EMP form. Then, learn the local constraints and global dynamics of each sequence by a conditional Restricted Boltzmann Machine. The approach is able to capture complex non-linearities in the data and distinguish between different motion styles, as well as the transition between action classes.

We introduce a combined template-based exemplar approach for online action recognition evaluated by using streamed motion sequences provided by two popular MoCap datasets (as demonstrated in Fig. 1). To construct an exemplar-based action model we select a single action sequence to represent each action class. Where each pose is represented in EMP form. k-means clustering is used to group action poses based on similarity. An exemplar is selected for each cluster based on a ranking scheme. As a product of the clustering process, we are able to order the exemplars into a time-ordered sequence of labels, to reflect the different *phases* of the action sequence. For recognition, we evaluate and classify each pose in the observed motion sequence individually. Each pose is represented in EMP form, and then compared against the exemplar model based on the City Block metric, which returns a prediction of the most similar exemplar. Finally, we utilise DTW to perform template matching and recognition by analysing the labels associated with each classified pose and a corresponding exemplar.

The remainder of this paper is organised as follows; Section 2 describes the selection criteria to construct a action model. Section 3 outlines the approach for comparing exemplars and observations based on a distance and label matching. Experimental results and conclusion are provided in section 4 and 5 respectively.

2 Exemplar-based Action Model

Human motion, typically captured by a marker-based MoCap system, is modelled using a *kinematic chain*. A kinematic chain consists of *body segments* that are connected to various body *joints*. Whereby a sequence can be seen as a time-sequential sequence of 3D joint coordinates that relate to the fixed kinematic chain. In our work, a motion sequence is a series of frames (otherwise denoted as poses), with each frame specifying the 3D coordinates of the joints at a certain time period. Thus, a sequence is denoted as $\mathcal{P} = \{p_t^j | t = 1, \ldots, T; j = 1, \ldots, J\}$, where t denotes the time and j is the joint index.

2.1 Pose Representation

There is no single solution to parametrisation of rotation that is suitable for all application domains. In our approach, we take each pose of a MoCap sequence and parametrise the 3D angle of each joint in Exponential Map form, as proposed in [4]. This allows us to avoid “gimbal” lock, discontinuities and ball-and-socket joints complications that are associated with using MoCap data. The Exponential Map is formulated from the corresponding Quaternion representation of each joint rotation. For clarity, the parametrisation can be expressed as

$$EMP(p_t^j) \mapsto \bar{p}_t^j \tag{1}$$

Thus, a set of transformed poses is denoted as $\bar{\mathcal{P}}$.

2.2 Exemplar Selection

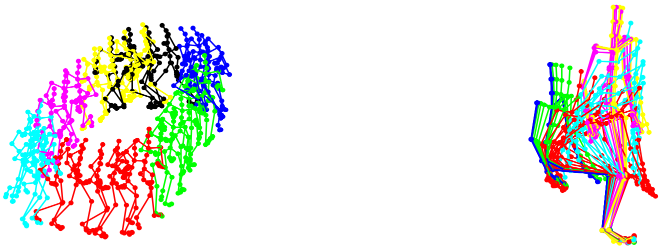


Fig. 2: Example clustering of MoCap data, where each cluster is denoted by a specific colour. Left: *Running in a circle*. Right: *Sitting in a chair*.

A challenge to overcome is that certain poses may be semantically similar but not necessarily numerically, yet, represent the same time instance within an action sequence. To further complicate matters, to identify and group similar poses manually is a time consuming and arduous task. In our approach, we have selected a centroid method, namely k-means clustering [7] to group $\bar{\mathcal{P}}$ into k cluster based on similarity. The selection of k-means over other approaches, is due to the way in which the selection criteria of the median element is performed. Further, k-means is computationally faster when dealing with a large number of observations when compared to hierarchical clustering methods making it ideal for clustering MoCap data.

For our application, the objective is to cluster an action sequence in to k clusters, with each cluster containing similar poses. Hence, as a by-product of k-means process, each cluster characterises a phase of the action. Fig. 2 demonstrates the clustering process for two distinctively difference action sequences. Observe that for each cluster, similar poses are grouped together, making it ideal for our application. However, the difficulty with k-means is the selection of k. The Elbow method [8], which uses percentage of variance determined as a function is implemented to select the appropriate k based on automatic selection of the Elbow criterion.

In order to select a delegate for each cluster, we deploy a ranking scheme for each pose according to the City Block metric (also referred to as Manhattan distance). The equivalence D between any two poses $\bar{p}_m$ and $\bar{p}_n$ in a cluster is measured using the total distance amongst corresponding joints, defined as

$$D(\bar{p}_m, \bar{p_n}) = \sum_{j=1}^{J} |\bar{p}_m^j - \bar{p}_n^j| \quad (2)$$

The delegate exemplar, denoted as $E_{e,k}$, is a pose which has the lowest distance average between a cluster grouping of poses. k-means provides an elegant method for segmenting an action sequence. For our proposed method, a unique label is assigned to each exemplar based on where it appears in the action sequence. This enables our recognition framework to perform template matching based on the action phase. Thus, for each action model we retain k number of exemplars and a time-ordered unique label sequence to represent each action class. For simplicity, let $C = \{c_e | e = 1, \ldots, E\}$ be the action set, where $c_e = \{E_{e,k} | k = 1, \ldots, K_e\}$ be the action model which is composed by K number of exemplars for action class e.

3 Recognition

Given a test sequence, $\mathcal{A} = \{a_t^j | t = 1, \ldots, T; j = 1, \ldots, J\}$, where t can be of any length. We treat each t-th pose individually and assume that the sequence forms an action in time order. Further, we expect that each pose correlates to an action class in our action set.

Firstly, we parameterise each pose in exponential form, as described in Eq. 1. Secondly, using the distance function (Eq. 2) to determine the similarity between the current observation and the delegate exemplars for each action model. We compute the minimum of the summation of distances defined by

$$L_t := min\{D(\bar{a}_t, E_{e,k}) | e = 1, \ldots, E; k = 1, \ldots, K\} \quad (3)$$

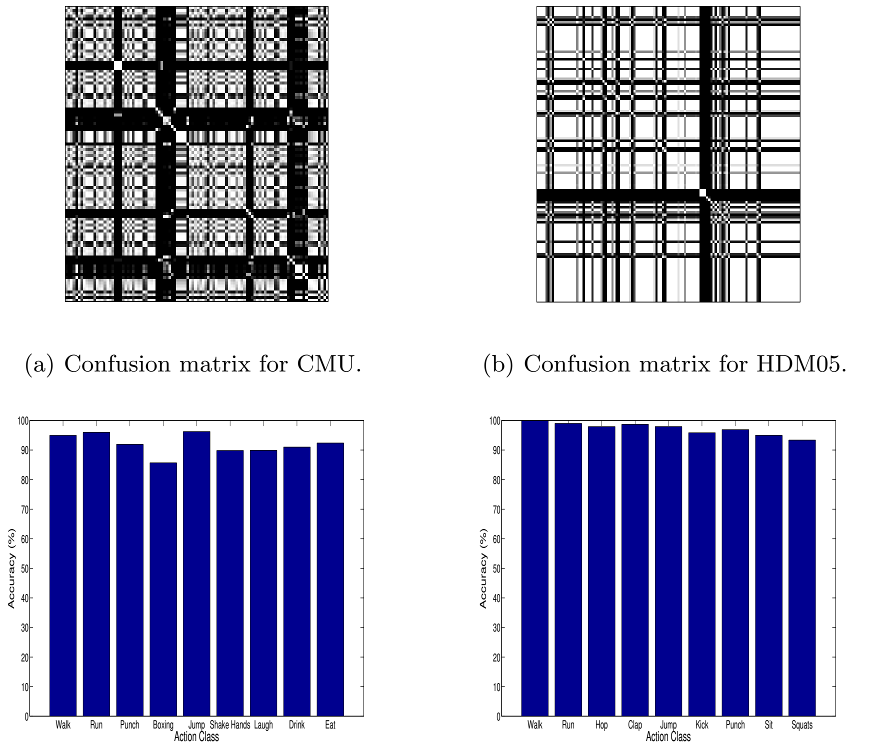
Therefore L_t is the winning exemplar label assigned to the observed pose. Recall, for each delegate exemplar we assign a unique label, using the selected exemplar and the associated label for each time instance we perform template matching to determine the action class of the current observation. Thus, over time a sequence of unique labels describing the action is constructed. We consider the temporal domain, as classifying poses individually will not provide the context to allow for robust recognition. An observed sequence can be of any length, to handle this we employ DTW as the action unfolds to match the observed label sequence to one of our previously learnt label sequences for each action model.

Given an observed label $L_1, \ldots, L_t$ derived above, we compare the sequence with each action model c_e to find the best pattern match. The winning class $\mathcal{L}_t$ is determined by the minimum DTW cost with respect to the Itakura constraint, given as

$$\mathcal{L}_t := min\{DTW(L_{1\sim t}, c_e)|_{e=1,\ldots,E}\} \quad (4)$$

Finally, by having each pose classified, recognition of the sequence up to any time period is defined based on the accumulative voting indicated in $\mathcal{L}$. The recognition rate is computed by the total number of correctly classified frames divided by the total number of frames up to point t.

4 Experimental Results



(a) Confusion matrix for CMU.

(b) Confusion matrix for HDM05.

(c) Action class recognition rate for the CMU dataset.

(d) Action class recognition rate for the HDM05 dataset.

Fig. 3: Confusion Matrix for two popular motion capture datasets using our proposed template-based exemplar approach.

All experiments were conducted on *unknown subject actions*, meaning that we include no data from the subject being tested. All results presented were computed based on average accumulation of classification results. We have tested the proposed approach on CMU [9] and HDM05 [10] datasets. These two datasets consist of single action sequences making them suitable for evaluating our proposed approach. For the CMU dataset we use 9 trials, with each trial representing a single action class to generate our action model and 49 trials

Table 1: Pose classification accuracy, average k selection and recognition rate of our proposed template-based exemplar approach.

Dataset	Accuracy	Avg k	Per pose (ms)
CMU	91.89%	21 (±18)	17ms
HDM05	97.95%	9 (±4)	9ms

for testing (as used in [3]). For the HDM05 dataset 9 trials were used, each representing a single action to generate our action model and 144 trials for testing.

As a pre-processing task, six joint angles contained within CMU and five from the HDM05 consisted of constant values, so they were removed from our training and testing sequences. The remaining joints had between two and three DOF. For testing, the conversion process was undertaken on a per pose basis. We have achieved respectable recognition rates, inclusive of computational costs, as shown in Table 1. This demonstrates the ability of the proposed method to handle frames presented by a continuous sequence of MoCap in an online real-time setting at high-speed.

For the CMU dataset, we achieved a recognition rate of **91.89%** using our proposed framework. The average k obtained reflected complexity of each action class (as shown in Table 1). The difficulty is to overcome these variations, thus the exemplar paradigm is effective by generalisation of variations in the data. *Walking* and *Run* classes were relatively easy to distinguish between due to relatively small variation amongst subjects. However, difficulties were encountered, as demonstrated in Fig. 3(a) & Fig. 3(c) for *Laugh* class due to pose similarity between *Boxing* and *Punching* classes. In other work, Barnachon *et al.* [3] reported an accuracy of 90.92% which is comparable to our proposed method.

For the HDM05 dataset, a recognition rate of **97.95%** (as shown in Table 1) was achieved using our proposed framework. As with the CMU dataset, k was dependent on complexity and pose variation of each action class. *Walking* and *Running* classes were once again clearly identifiable. However confusion was observed for *Sit* and *Squat*, as demonstrated in Fig. 3(b) & Fig. 3(d) . Confusion observed was due to similar poses being performed in other action classes. Overall, interclass confusion remained limited reflecting the strength of our approach to correctly model action sequences using the exemplar paradigm. In other works tested on the HDM05 dataset, Muller *et al.* [11] reported an accuracy of 80% when using their *key-frames* approach. While this was a good recognition rate, the process cannot be performed online due to complex data processing.

Our approach achieved overall recognition results that are superior to other approaches, when using the CMU (91.89%) and HDM05 (97.95%) datasets. This is significant advance on current approaches, with the added benefit of being a more straight-forward approach for analysing highly complex datasets and also the small number of exemplars retained (avg. k = 15). For online application, there is a need to detect actions early on in a sequence. Fig. 4 demonstrates the

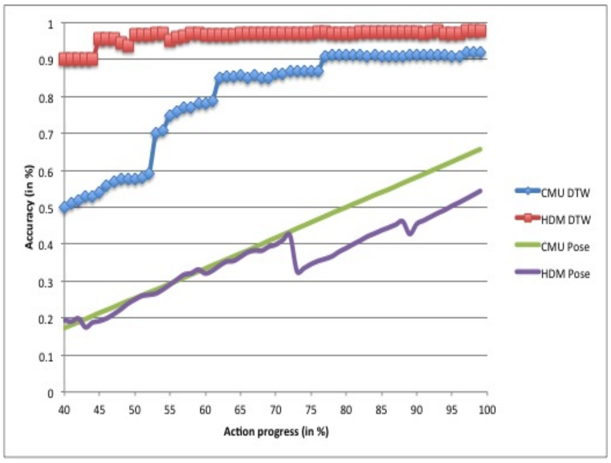


Fig. 4: Early recognition rate using our proposed method compared with individual pose results.

potential of our approach for early action recognition. In addition, the use of the exemplar paradigm has reduced the need to use complete action sequences by approximately 98% while remaining efficient in representing each action class. Further, observe in Fig. 4 that by using the proposed approach we have achieved superior results when compared with classifying poses individually without template matching.

This work has focused on pose classification and template matching of real number sequences, which has presented a number of challenges. It has been difficult to distinguish certain actions due to poses being indistinguishable for a period resulting in class confusion. One solution may be to focus on adjacent poses in the temporal domain to aid in identification of the correct action class as they progress through time. The results we have achieved, as shown in Table. 1, present our baseline results for future exemplar-based approaches. Finally, the results are comparable to current exemplar-based approach, namely Barnachon *et al.* [3], in terms of accuracy and recognition time per pose ($17ms$). There is scope for further improvement in stabilising class confusion, classification and recognition of action(s) as they unfold.

5 Conclusion

In this paper, we propose a template-based exemplar approach for human action recognition evaluated using a stream of MoCap. The use of an action model to represent each action class, where each model is characterised by a small number of exemplars, has reduced the need to use whole motion sequences for action representation by an average of 98%. Ultimately this leads to a significant computational saving for recognition which then transforms the problem into a relativity simple distance/template matching task. Yet emphasis is placed on the importance of selecting *ideal* actions for each action class. By using MoCap action sequences, our work seeks to present a baseline study for future exemplar-based recognition approaches. To our knowledge, this paper is the first to propose an exemplar approach for MoCap sequences represented in exponential map form. By utilising EMP we have been able to effectively model key characteristics for each action sequences but also handle singularities and discontinuities. Further work is required to address issues related to similarity in poses shared between multiple action classes, possibly analysing temporal occurrence. With further experiments undertaken on a wider and varied testing set while handling more complex action(s) and diverse motion datasets (*e.g.* Microsoft Kinect).

References

1. R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, June 2010.
2. A. Elgammal, V. Shet, Y. Yacoob, and L. S. Davis, "Learning dynamics for exemplar-based gesture recognition," in *CVPR.* Washington, DC, USA: IEEE Computer Society, 2003, pp. 571–578.
3. M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, "Ongoing human action recognition with motion capture," *Pattern Recognition*, 2013.
4. F. S. Grassia, "Practical parameterization of rotations using the exponential map," *Journal of Graphics Tools*, vol. 3, no. 3, pp. 29–48, 1998.
5. C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *CVPR*, 1998, pp. 8–15.
6. G. W. Taylor, G. E. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," in *NIPS*, 2006, p. 2007.
7. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematrical Statistics and Probability.* University of California Press, 1967, pp. 281–297.
8. D. Ketchen and C. Shook, "The application of cluster analysis in strategic management research: An analysis and critique," *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, 1996.
9. Carnegie Mellon University Motion Capture Dataset. The data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.
10. M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database HDM05," Universität Bonn, Tech. Rep. CG-2007-2, 2007.
11. M. Müller, A. Baak, and H.-P. Seidel, "Efficient and robust annotation of motion capture data," in *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, New Orleans, LA, August 2009, pp. 17–26.