Skeleton-Based Human Action Recognition with Profile Hidden Markov Models

Wenwen Ding[1,2], Kai Liu[1(✉)], Fei Cheng[1], Huan Shi[1], and Baijian Zhang[1]

[1] School of Computer Science and Technology, Xidian University, Xi'an, China
[2] School of Mathematical Sciences, Huaibei Normal University, Anhui, China
{dww2048,chengfei8582}@163.com, kailiu@mail.xidian.edu.cn,
shihuan.xidian@gmail.com, zhangbaijian0307@126.com

Abstract. Recognizing human actions from image sequences is an active area of research in computer vision. In this paper, a novel HMM-based approach is proposed for human action recognition using 3D positions of body joints. First, actions are segmented into meaningful action units called dynamic instants and intervals by using motion velocities, the direction of motion, and the curvatures of 3D trajectories. Then action unit with its spatio-temporal feature sets are clustered using unsupervised learning, like SOM, to generate a sequence of discrete symbols. To overcome an abrupt change or an abnormal in its gesticulation between different performances of the same action, Profile Hidden Markov Models (Profile HMMs) are applied with these symbol sequences using Viterbi and Baum-Welch algorithms for human activity recognition. The experimental evaluations show that the proposed approach achieves promising results compared to other state of the art algorithms.

Keywords: View-invariant representation · Skeleton joints · Human activity recognition · Profile HMM · Self-organizing map

1 Introduction

Recognizing human activity is a key component in many applications, such as Video Surveillance, Ambient Intelligence, Human-Computer Interaction systems, and even Health-Care. Despite remarkable research efforts and many encouraging advances in the past decade, accurate recognition of the human actions is still a quite challenging task.

Many recent state-of-the-art techniques for human action recognition rely on: Bag-of-Word (BoW) [1] representations extracted from Spatio-Temporal Interest Points (STIP) [2], Dynamic Time Warping (DTW)[3] algorithm derived from exemplar-based approaches, Eigenjoints [4] stem from skeleton-based approaches, etc. Despite these good results were achieved by state of the art activity recognition approaches, these still have some limitations.

To address these issues and enhance human action recognition performance, time-sequential representation is more appropriate for these problem. Frame by

H. Zha et al. (Eds.): CCCV 2015, Part I, CCIS 546, pp. 12–21, 2015.
DOI: 10.1007/978-3-662-48558-3_2

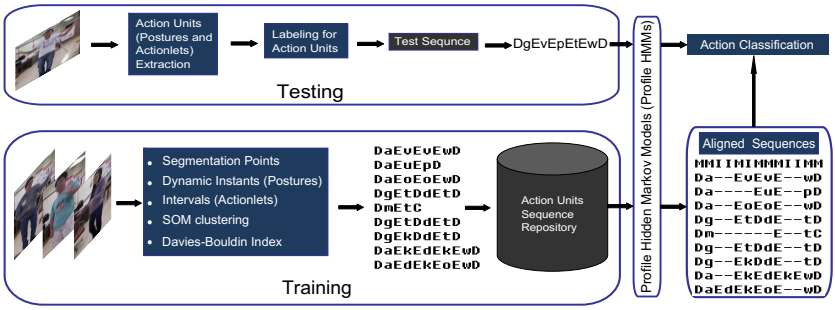


Fig. 1. The general framework of the proposed approach.

frame representations suffer from redundancy. Therefore segmenting video into states and handling unaligned video sequences are two main problems. In this paper, we use action-units and novel probabilistic methods (Profile HMM [5]) to handle unaligned video sequences. The principle is illustrated in Fig. 1. First, trajectories of action, also referred to as discrete curves, can be drawn by several 3D joint points. The segmentation points S, splitting actions into meaningful action-units, can be captured by the direction of motion and curvature of the trajectory having maximum velocity. Then, the features of action units, consisting of dynamic instants (postures) $\boldsymbol{\xi}_p$ and intervals (actionlets [6]) $\boldsymbol{\xi}_a$, are extracted from these segmented trajectories and then are mapped into two Self Organizing Mappings (SOMs) [7] recorded as T_{ξ_p} and T_{ξ_a}, respectively. Unlike actions that have labels showing on, postures and actionlets do not have such labels. Therefore, T_{ξ_p} and T_{ξ_a} can be scattered in plots according to the Davies-Bouldin Index (DBI) value [8] which decide the number of labels of postures and actionlets. These plots in SOM can be named with upper-case letters and lower-case letters respectively referred as the labels of postures and actionlets. Finally, capturing the sptio-temporal relationships between action-units of a given action, Profile HMMs are generated by sequences of discrete symbols of each action. With these profile HMMs, each action represented by time-series is trained and aligned, thus elevating classification performance.

The rest of the paper is organized as follows: Section 2 presents the related work; Section 3 elaborates our method of features extraction, clustering and classification of action units consisting of postures and actionlets; Section 4 and discusses the parameters setting presents our experimental results; and Section5 concludes this paper.

2 Related Work

Action Recognition. In the past decade, video-based action recognition and detection has tremendous amount of background literature [9,10]. Recently, with the development of the commodity depth sensors like Microsoft Kinect [11],there has been a lot of interests in human action recognition from depth data. Several

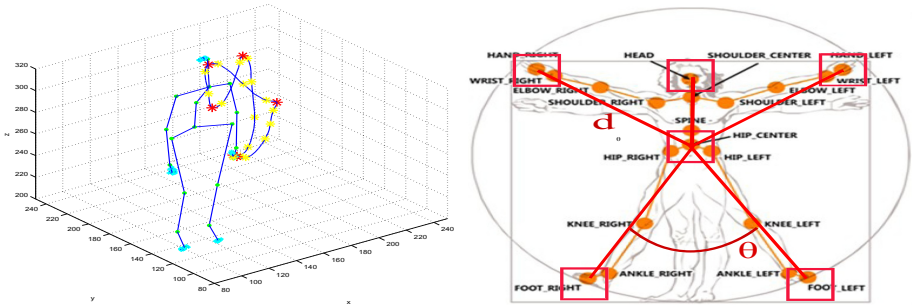
research utilize skeleton joint positions as features for action recognition. Li et al. [12] employed a bag-of-3D-points graph approach to encode actions based on 3D projection of body silhouette points. Xia et al. [13] mapped 3D skeletal joints to a spherical coordinate system and used a histogram of 3D Joint Locations (HOJ3D) to achieve view-invariant posture representation. The joints were then translated to a spherical coordinate system to achieve view-invariance.

Spatio-Temporal Alignment. Spatio-temporal alignment of human action has been a topic of recent interest due to its applications in animation and human activity recognition. Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) are two main approaches based on sequential representation of the activity for this problem. In [14], each action is modeled as a series of synthetic 2D human poses matched by using the Viterbi algorithm. Mapping poses or frames into symbols is the main challenge of HMM approaches. But these frame by frame representations suffer from redundancy. Furthermore, HMM structure must be adaptively designed for specific application domains. DTW is a method for temporally aligning multi-modal sequences from multiple subjects performing similar activities. DTW deals with sequence aligning by operations of deleting and inserting compression expansion, and substitution, of subsequences. Zhou and Torre [15] extended DTW to propose Canonical Time Warping (CTW) for finding the temporal alignment that maximizes the spatial correlation between two behavioral samples coming from two subjects.

3 Proposed Method

3.1 Representation of Meaningful Action Units

Action is represented as a sequence of dynamic instants and intervals, which are computed using the direction of motion and the spatio-temporal curvature of a 3D trajectory. It use depth cameras to track 3D trajectories that



(a) (b)

Fig. 2. (a) The trajectory of an action of *high hand wave* is segmented by *red stars*. (b) Illustration of human posture representation based on relative distance and angles of star skeleton.

each trajectory represents the evolution of one coordinate x, y, or z over time, and indicates the position of a specific joint of human. Motion trajectories provide rich spatio-temporal information about an object's behavior. To obtain meaningful action units, we must learn superior segmentation points $\mathcal{S} = \{s_1, ...s_i, ..., s_j, ..., s_m\}(1 < i < j < m)$ to segment 3D trajectory of an action, as shown in Fig. 2a. The problems of under-segmented and over-segmented trajectories will always lead to insignificant action units. Based on [6], superior segmentation points $\mathcal{S}$ for a trajectory can be obtained.

For dynamic instants of action, we can utilize human postures to represent in this moment. Human postures can be represented by relative distances d and angles θ from 3D star skeleton,as shown in Fig. 2b. For intervals of action, we can utilize actionlets to represent these intervals from paper [6].

3.2 Clustering Feature Using Unsupervised Learning

Action labels are easily labeled in real life, such as walk, sit down, stand up, throw, etc. Unlike actions with labels that are shown on a map grid, an actionlet or a posture is hardly labeled or highly generalized using our human language. Therefore, Self Organizing Map (SOM) [7] and the Davies-Bouldin Index (DBI) [8] value are used to cluster postures and actionlets.

Self Organized Mapping. A self-organizing map(SOM) is a type of artificial neural network for the visualization of high-dimensional data using unsupervised learning. It can project complex, nonlinear statistical relationships between high-dimensional patterns into simple geometric relationships on a low-dimensional topology map. The training process of SOM is an incremental learning algorithm. The weight vectors m of nodes are initialized either to small random values or sampled evenly from the subspace spanned by the two largest principal component eigenvectors, which is a good initial approximation.

Davies Bouldin Index. The feature of postures ξ_p and actionlet ξ_a map to SOM forming similar neural units in T_{ξ_p} and T_{ξ_a} need to be grouped and labeled later. To find initial partitioning, we use the Davies-Bouldin index value to scattered the T_{ξ_p} and T_{ξ_a} in plots. By definition, the lower the DBI, the better the separation of the clusters and the tightness inside the clusters.

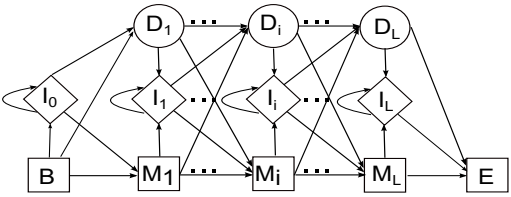
The number of plots of the T_{ξ_p} and T_{ξ_a} can be decided by the definiton of DBI. Each plot was symbolized by capital or lower-case letters according to the posture of actionlet. Therefore, The feature of postures ξ_p and actionlet ξ_a were transformed into symbols from a discrete alphabet so that an action can be represented by upper-case and low-case letters generated alternately in a individual sequence, for example, *DaEvEvEwD*. Similar actions will correspond a sequence family F for generating a Profile HMM or say a motif.

3.3 Profile HMMs for Temporal Alignment of Human Motion

In this section, we describe the design of general Profile HMMs and our Profile HMMs in greater detail. The classifiers we build for human action recognition are based on our Profile HMMs. Using the Forward-Backward algorithm [16], we can compute the total probability of a sequence being generated by Profile HMMs, i.e. can be used to classify unknown sequences as belonging to which model. Using the Viterbi algorithm [17], we can compute the most likely path through Profile HMMs that generates a sequence, i.e. the most likely alignment of the sequence against the model. Using initial parameters that assign uniform probabilities over all action units in each time step, we apply the well known Baum-Welch algorithm [18] to iteratively find new Profile HMM parameters which maximize the likelihood of the model for the sequences of action units in the training videos.

Profile Hidden Markov Models. Profile HMMs consist of several types of states: match states M_i, insert states I_i, and delete states D_i. For each position i in a Profile HMM, there is one match state, one insert state, and one delete state. A Profile HMM can thus be visualized as a series of columns, where each column represents a position i in the sequence as shown in Figure 3a. Any arbitrary sequence can then be represented as a traversal of states from column to column. Each state emits symbols with a probability distribution specific to its position in the chain.

Given a Profile HMM, how to align multiple sequences based on the model is the first problem to solve. Viterbi algorithm is used for seeking the most likely path of each sequence generated by the model. Multiple sequence alignment is mean to find Viterbi path of each sequence. Fig. 3b shows a small example of a set of human posture sequences. Profile HMM (Fig. 3a) can be constructed from the set of sequences by using the Baum-Welch algorithm. The result of aligned sequences can be showed in Fig. 3c.



(a)

NFLS
NFLS
NKYLT
QWT

(b)

N – F L S
N – F L S
N K Y L T
Q – W – T

(c)

Fig. 3. (a) A general Profile HMM of length L. M_i is the ith match state, I_i is the ith insert state, Di is the ith delete state. B is the begin state, and E is the end state. (b) Illustration of human posture representation based on relative distance and angles of star skeleton.

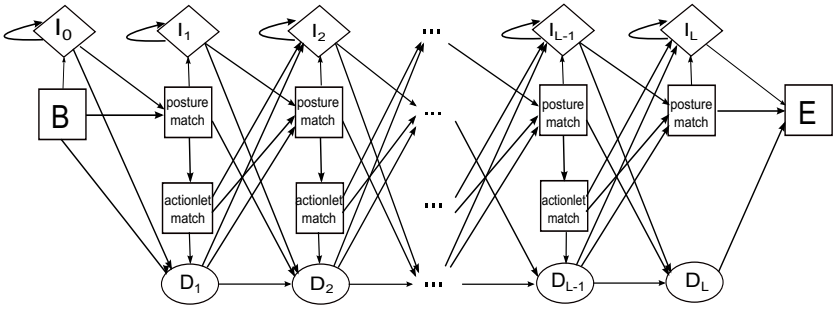


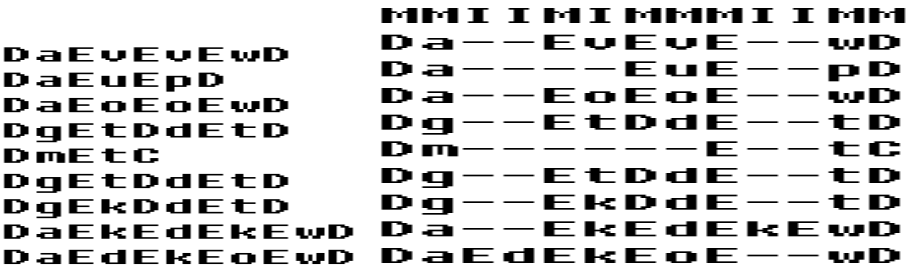
Fig. 4. The Profile HMM for human action recognition.

Adapting Profile HMMs for Human Action Recognition. In this section we describe the structure of our profile HMMs as shown in Fig. 4. The main difference between our profile HMM and others is that the Profile HMMs used in biology have only a single chain of Match states. In our case, the addition of a second match state per position was intended to allow the model to represent the correlation between action units in videos. In the context of human action recognition, actions are segmented into meaningful action-units called dynamic instants and intervals being incarnated in human postures and actionlets, which labels of postures and actionlets are upper-case and lower-case letters respectively. Therefore, an action can be represented by a string, for example, *DaEvEvEwD*. Pay attention to the first and the end symbol which is upper-case letters meaning that an action begin or end with a posture in our observation. This change is necessary as postures and actionlets obviously alternated in an action. To allow for variations between the observed action-units in the same action sequences, the model has two additional states for each position in the chain. One is insert states I_i representing one or more extra abrupt or abnormal action-units inserted in a sequence between two normal parts of the chain. The other is Delete states D_i allowing period action-units to be omitted from the action sequences.

We now explain the design and use of profile HMMs Λ of k classes with models $\Lambda_1, \Lambda_2, ..., \Lambda_k$ which employ to capture characteristics exhibited by every kind of actions. If we already have a set of action-unit sequences (Fig. 5a) belonging to a family, a profile HMM $\Lambda_c (1 < c < k)$ as shown in Fig. 4 can be constructed from the set of unaligned sequences after using the Baum-Welch algorithm. The length L of the $\Lambda_c (1 < c < k)$ must be chosen, and is usually equal to the average length of the unaligned action unit sequences in the training set. The transition and emission probabilities are initialized from Dirichlet distributions.

Once Profile HMMs Λ have been constructed, we then construct a classifier $\mathcal{C}_1$ for the task of choosing the best model $\Lambda_c (1 < c < k)$ for new test sequence q of action-units

$$c = \mathcal{C}_1(q) = \underset{c}{\operatorname{argmax}}\, P(q|\Lambda_c). \tag{1}$$



(a) (b)

Fig. 5. (a) A set of action-unit sequences of action *high arm wave*. (b)The alignment generated via the Profile HMM method for the set of action-unit sequences of action *high arm wave*. The match and insert columns are marked with the letters M and I respectively in the first line.

This is done via a straightforward application of the forward-backward algorithm, i.e. to get the full probability of the given sequence q.

The second classifier $\mathcal{C}_2$ makes use of the well-known Viterbi algorithm for finding the most likely alignment of the sequence to the family, i.e. Viterbi path V. For a given output sequence q and the associated probability of the most likely Viterbi path V_c to each Profile HMM, the viterbi classifier $\mathcal{C}_2$ finds Viterbi paths for the sequence in each Profile HMM $\Lambda_1, \Lambda_2, ..., \Lambda_k$ and chooses the class c whose model produces the best Viterbi path V_c.

$$c = \mathcal{C}_2(q) = \underset{c}{\operatorname{argmax}}\, P_{viterbi}(q, \Lambda) = \max_{V_c} P(q, V|\Lambda). \tag{2}$$

In practical terms, the Viterbi classifier $\mathcal{C}_2$ finds each model's best explanation for how the action-units in the sequence were generated. We choice the Viterbi classifier $\mathcal{C}_2$ that provides the best explanation for the observed action-units.

4 Experimental Evaluation

The performance of the activity recognition was primarily evaluated based on its accuracy. In this section, we evaluate the proposed skeletal representation using three different datasets: MSR-Action3D [12], UTKinect-Action [13], and UCF Kinect Dataset [19].

4.1 Evaluation Settings

For MSR Action3D Dataset, in order to allow a fair comparison with the state of the art methods, we followed the test setting of [12], dividing the 20 actions into three subsets AS_1, AS_2 and AS_3 and using two experimental settings: one is non-cross-subject test setting and another is cross-subject test setting.

For UTKinect-Action Dataset, to allow for comparison with [13], we followed the same experimental set up using Leave One Sequence Out Cross Validation (LOOCV) on the 200 sequences. For UCF Kinect Dataset, we followed the same experimental set up using the Latency Aware Learning in [19].

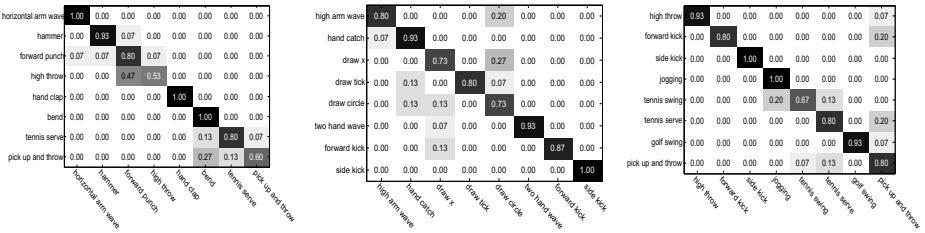
4.2 Experimental Results

We first evaluate the performance of the proposed approach on the three challenging 3D action datasets. The proposed method's primary advantage is robustness temporal misalignment. The experiment results on the three datasets are shown in Table 1. We can see that the proposed approach gives the best results on all datasets. In our experiments, the cross-subjects action recognition is conducted, which is more difficult than using the same subjects for both training and testing. From the results of MSR Action3D dataset on cross-subjects test, the recognition accuracy of our method on test three was 88.6% significantly outperforming the other joint-based action recognition methods, including Bag-of-3D-Points[12], Histogram of 3D joints[13], and EigenJoints [4], which achieved accuracies of 74.4%, 78.97%, and 82.3%, respectively. Specifically, it outperforms the state-of-the-art on UTKinect-Action dataset and UCT Kinect dataset. On the UTKinect-Action dataset, our approach has an accuracy of 91.7% which outperforms the HOJ3D feature in [13] (90.9%). Finally, we compare our result with all others on the UCF Kinect dataset. The results are shown in Table 1.

Fig. 6 shows the confusion matrices for MSRAction3D AS1, MSR-Action3D AS2 and MSR-Action3D AS3. We can see that most of the confusions are between highly similar actions like *forward punch* and *high throw* in the case of

Table 1. Human recognition accuracies on three datasets.

MSR Action3D(Test Three)	**Accuracy**
Bag of 3D Points[12]	74.7
Histogram of 3D Joints[13]	78.9
Eigenjoints[4]	82.3
Spatio-temporal Feature Chain [6]	84.4
Random Occupancy Patterns[**?**]	86.2
Proposed Method	88.6
UTKinect-Action	**Accuracy**
HO3DJ[13]	90.9
Spatio-temporal Feature Chain [6]	91.5
Proposed Method	91.7
UCF Kinect	**Accuracy**
LAL[19]	95.9
Eigenjoint[4]	97.1
Spatio-temporal Feature Chain [6]	98.04
Proposed Method	97.6



(a) AS1CrSub (b) AS2CrSub (c) AS3CrSub

Fig. 6. Confusion matrix in AS1, AS2 and AS3 under Cross Subject Test using STFC.

MSR-Action3D AS1, *draw X*, *draw tick*, and *draw circle* in the case of MSRAction3D AS2, and *tennis swing*, *tennis serve*, and *pick up and throw* in the case of MSR-Action3D AS3.

5 Conclusions and Future Work

In this paper, we obtain meaningful action-units through take advantage of segmentation points. With labeling these action-units, an action can be represented by discrete symbol sequences. To overcome an abrupt change or an abnormal in its gesticulation between different performances of the same action, Profile Hidden Markov Models (Prifile HMMs) are applied with these symbol sequences using Viterbi and Baum-Welch algorithms for human activity recognition. These methods eliminate the noise and the periodic motion problems experienced by methodologies that either solve it only by hand setup or else ignore it. Applying action sequences to Profile HMMs resulted in our approach to significantly outperform other state of the art methods. The next step is to understand and predict human activities and object affordances combining more contextual information, and more importantly, of human interactions with the objects in the form of associated affordances.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant No. 61350110239, the Fundamental Research Funds for the Central Universities under Grant No. K5051203005, the Open Research Funds of State Key Lab. for novel software technology under Grant No.KFKT2012B16, and the Natural Science Foundation of the AnHui Higher Education Institutions of China under Grant No. KJ2014B14.

References

1. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**(3), 299–318 (2008)

2. Laptev, I.: On space-time interest points. International Journal of Computer Vision. **64**(2–3), 107–123 (2005)
3. Rabiner, L.R., Juang, B.-H.: Fundamentals of speech recognition, vol. 14. PTR Prentice Hall Englewood Cliffs (1993)
4. Yang, X., Tian, T.: Effective 3d action recognition using eigenjoints. Journal of Visual Communication and Image Representation **25**(1), 2–11 (2014)
5. Krogh, A., Brown, M., Mian, I.S., Sjolander, K., Haussler, D.: Hidden Markov models in computational biology: Applications to protein modeling. Journal of Molecular Biology **235**(5), 1501–1531 (1994)
6. Ding, W., Liu, K., Cheng, F., et al.: STFC: Spatio-temporal feature chain for skeleton-based human action recognition. Journal of Visual Communication and Image Representation **26**, 329–337 (2015)
7. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990)
8. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI **1**(2), 224–227 (1979)
9. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6), 976–990 (2010)
10. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding **115**(2), 224–241 (2011)
11. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**(2), 4–10 (2012)
12. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 9–14.IEEE (2010)
13. Xia, L., Chen, C.-C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27. IEEE (2012)
14. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (2007)
15. Zhou, F., Torre, F.: Canonical time warping for alignment of human behavior. In: Advances in Neural Information Processing Systems, pp. 2286–2294 (2009)
16. Ferguson, J.D.: Variable duration models for speech. In: Proceedings of the Symposium on the Application of HMMs to Text and Speech, pp. 143–179 (1980)
17. Viterbi, A.J.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory **13**(2), 260–269 (1967)
18. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics 164–171 (1970)
19. Ellis, C., Masood, S.Z., Tappen, M.F., Laviola Jr, J.J., Sukthankar, R.: Exploring the trade-off between accuracy and observational latency in action recognition. International Journal of Computer Vision **101**(3), 420–436 (2013)