Action Recognition Using MHI Based Hu Moments with HMMs

Elit Cenk Alp
Department of Computer Engineering
Ankara University
Ankara, Turkey
elitcenkalp@gmail.com

Hacer Yalim Keles
Department of Computer Engineering
Ankara University
Ankara, Turkey
hkeles@ankara.edu.tr

***Abstract*—Action recognition from video streams is among the active research topics in computer vision. The challenge is on the identification of the actions robustly regardless of the variations imposed by appearances of actions performed by different people. The challenge increases when the data is gathered from an outdoor environment, i.e. background and illumination variations. This paper proposes a Hidden Markov Model (HMM) based approach to model actions using Hu moments that are computed using a modified Motion History Images (MHI). The experiments performed using Weizmann dataset show that the proposed method generates 99% classification accuracy, which is comparable to many state of the art techniques that use MHI and HMMs separately.**

Keywords—Action recognition; MHI; HMM;

I. INTRODUCTION

Action recognition is an active research area in computer vision with many important applications, including human–computer interfaces, content-based video indexing, video surveillance and robotics, among others [1]. This paper details a method to cope with temporal features that are needed for the detection of human actions using Hidden Markov Models; we extend the Motion History Image (MHI) [2]. The MHI is an appearance based template generation approach for motion analysis. The advanced motion in a temporal template is summarized into a gray scale image, where the most recent motion information is preserved. In other words, MHI provides a history of temporal changes of a time sequence image that decay over time [3].This method is not sensitive to silhouette based noises, like shadows, missing parts and holes. In this research, we aim to find an answer to this question: "Although MHIs represent the evolution of motion in time, can we improve it further by tracing hidden elements in their flow in time?" Our experiments show that the answer is positive.

Combining the robust temporal representation of MHIs with the learning capability of time sequential data of the Hidden Markov Models (HMMs), we employed a new approach to model and recognize actions. MHI representation makes our approach robust to variances that occur at the outdoor environments, and enables the method focus more on the characteristics of motion. We extract 7 Hu moments from the MHI representation at each frame and use these features to encode the patterns captured from different actions with scale stability using HMMs.

The remainder of this paper is organized as follows. Section 2 simply reviews related work and Section 3 details feature extraction. Section 4 describes activity modeling and recognition. Experimental results and discussion are explained in Section 5.

II. RELATED WORK

Weinland et al. [1] provide a survey of the literature on action recognition. They have classified approaches with respect to how they represent the spatial and temporal structure of actions, how they segment and recognize actions from a continuous video stream, and how they handle variations in camera viewpoint. Bobick and Davis [2] evolve the temporal template representation. This representation takes both shape and motion, represented as evolving silhouettes extracted using background subtraction. They compute statistical descriptions of MHI and MEI images using moment-based features. Their choice is 7 Hu moments; yet they utilized these features as temporal templates.

Yamato et al. [4] conducted a study using discrete HMMs to describe movements in time sequential tennis images. Each feature vector of the sequence is assigned to a symbol by vector quantization. But, some information is lost in the quantization process. They obtained promising results using HMMs. Kale et al. [5] developed a human recognition approach using the gaits that are computed based on width of their binary silhouettes. The observation probability function is modeled as a continuous distribution. Using HMMs, they made tests on 3 different data sets that recognize gaits. Sundaresan [6] uses continuous HMMs to recognize gaits. They also used sums of silhouettes as their features. They obtain best results using inner product distances.

Mendoza and Blanca [7] describe a robust contour feature, i.e. shape-context, which is based on continuous HMMs. The shape-context feature vector is built from the histogram of a set of non-overlapping regions in the frame. Using the KTH dataset, they obtained human motion recognition accuracies up to 94%. Martinez-Contreas et al. [8] proposed a framework for recognition of action recognition from a silhouette based feature set. Their model relies on 2D modeling of human action based on motion history image (MHI). These templates are projected on a Kohonen self-organizing feature map (SOM). After they got MHIs, they used discrete HMM by lowering the features with SOM. In the MuHaVi-MAS dataset they report 98% success rate, and in the ViHaSi 99% success rate.

Vezzani et al. [9] models the emission probability function as Mixture of Gaussians and used the projection histogram of the foreground masks as their features. Each histogram contains the number of moving pixels for each row and each column of the image frame. HMM framework improve the temporal evolution of postures. Using the Weizmann dataset,

978-1-5090-3843-5/17/$31.00 ©2017 IEEE

they achieved a 96% success. Hsieh et al. [10] extracted silhouette mapped into three different polar coordinate systems that characterize three parts of human figure. Each polar coordinate system is quantized by partitioning some cells with different angles. After decreasing the number of features with PCA, they achieved 98% success rate in the Weizmann dataset. Kuehne et al. [11] proposed a framework that models human activities as temporally structured processes. They model action units using HMMs, similar to speech recognition systems. These action units form the building blocks to model complex activities as sentences using an action grammar.

Based on simple silhouettes, Wang and Suter [12] have described a probabilistic framework for identifying human activities in video streams. The method combines kernel principal component analysis (KPCA) and factorial conditional random field (FCRF). In education; HMM, CRF, Factorial CRF, and various models as data and two different data sets are used. 98% in HMM, 95% in CRF, and 100% in factorial CRF. Fathi and Mori [13] have chosen low-level optical flow features and selected some of them using AdaBoost to create medium-level features. Later, these mid-level features were also trained by AdaBoost; they report 90% success in the KTH dataset, 99% success in the Weizmann dataset, 71% success in the Soccer dataset, and 51% success in the Ballet Dataset. Ahad et al. [3] developed the Directional Motion History Image using the MHI concept; MHI and MEI are calculated according to the Optical Flow angle. They used the features taken from the MEIs and the Hu Moments that are extracted from the MHIs in training. With this method, 93%success was reported.

Recently, Afsar et al. [15] presented a method to automatically detect human from videos and recognizing actions. They used human boundary information as a main feature. These features have been used in HMM by translating symbols with Vector Quantization (VQ). Lei et al. [16] proposed a hierarchical framework that combines a convolutional neural network (CNN) and HMM to recognize and segment continuous actions. In their work, the Gaussian mixture model was replaced by a CNN to model the emission distribution of the HMMs. They report 89.2% success in Weizmann dataset and 94.43% in KTH dataset.

III. Feature Selection

Simple features have a critical value for motion identification, due to (1) fast extraction, (2) easy interpretation. After we generated the MHI of a frame, we computed 7 Hu Moments using it and used the moments to model different actions with HMMs. The proposed system is very simple in that the model is based only on these seven features. Although there are works in the literature that utilizes MHIs, Hu moments and HMMs separately; to the best of our knowledge, our proposal is novel in combining them together in the action recognition domain.

The flow chart of the proposed system is depicted in Fig. 1. For each category, the MHI and Hu Moments are extracted after the background subtraction operation. For training, we normalize the features using all the features from all video categories. Then, we use the Baum-Welch algorithm for training HMMs. During testing, the same features (i.e. MHI + Hu Moments) are extracted from the test videos, and the features are normalized using the same normalization vectors in training. For each category, the model that gives the largest log-likehood value is selected for classification.

A. Motion History Image

MHI is a scalar-valued image such that the pixels that have moved more recently are brighter (i.e., pixels with high intensity values belong to the recent motion) and vice versa (i.e. pixels with lower intensity values reveal a previously occurring motion over time). The intensity value of zero depicts that there is no detected movement in that particular place in the history of the sequence. MHI [1] expresses the motion flow or sequence using the temporal density of each pixel. To define motion shape and space distribution, Weinland et al. [1] also introduced the Motion Energy Image (MEI). One of the advantages of MHI representation is that the time frame up to a few frames can be coded in a single frame, and the time scale of MHI's human motion is covered. MHI, i.e. H_τ (x, y, t), can be calculated using the update function given in Eqn (1) as:

$$H_\tau(x,y,t) = \begin{cases} \tau & if\ \Psi(x,y,t) = 1 \\ \max(0, H_\tau(x,y,t-1) - \delta) & otherwise. \end{cases} \quad (1)$$

In this equation, (x, y) and t show the spatial position and the time, respectively. τ parameter determines the temporal duration of the MHI; δ is the decay parameter. Ψ(x, y, t) signals presence of a motion in the current video image. Update function is called for each new video frame that is analyzed sequentially. A set of different image processing techniques are utilized in the literature to describe this update function, e.g. background subtraction, optical flow, image differences, etc. MHI is constructed from an image obtained from the frame subtraction using a threshold ξ:

$$\Psi(x,y,t) = \begin{cases} 1 & if\ \mathrm{D}(x,y,t) \geq \xi \\ 0 & otherwise. \end{cases} \quad (2)$$

where, D(x, y, t) is the frame difference.

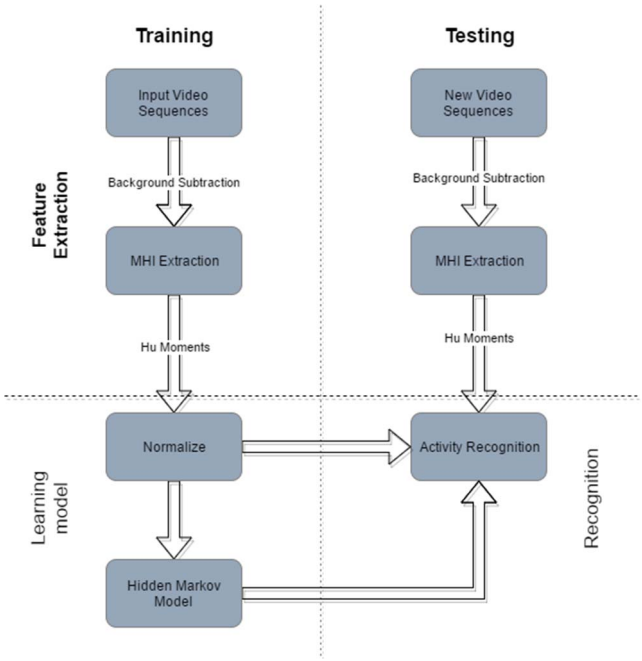


Fig. 1. The flow chart of the proposed action recognition system.

B. Modified Motion History Image

We modified H_τ such that the decay factor is scaled over time. This approach is advantageous in that instead of a constant linear decay factor, an exponential decay factor emphasizes the recent motion more effectively. The HMM models utilizing this time scaled decay factor increased recognition accuracy.

$$H_\tau(x,y,t) = \begin{cases} \tau & if\ \ \Psi(x,y,t) = 1 \\ H_\tau(x,y,t-1) * \beta & otherwise. \end{cases} \quad (3)$$

In this function β is a scalar scale value that should be set between 0 and 1; this setting reduces the MHI value with time.

In order to compute Ψ (x, y, t), we simply take the differences between consecutive image frames by subtracting the previous image from the incoming image.

In our experiments, we set β to 0.9; which results in 10% decrease in pixel values in the following frame. But the pixel value never drops lower than 4, due to truncation. This enables us to use MHI as motion energy images (MEI) which keeps the history of moving pixels in a separate process. Hence, in our approach, we do not need to extract MEIs explicitly, since MHI serves both purposes alone; a simple thresholding, e.g. using 4, gives us MEI. Sample MHIs are depicted in Fig. 2.

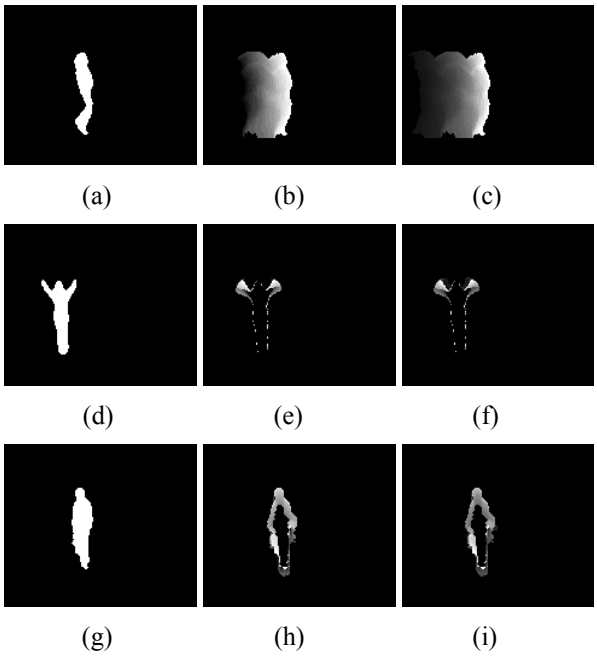


Fig. 2. The first column (a, d, g) contains silhouette images. The images in the second column (b, e, h) are the MHIs that are extracted using classical method. The images in the third column (c, f, i), are the MHIs that are obtained using our scale-based method. The ones on the first row (a, b, c) belong to jumping, the second row images (d, e, f) belong to two-hands wave, and the last row (g, h, i) contains jump in place.

IV. Motion Modelling and Recognition

Hidden Markov Models (HMM) are very convenient for human motion analysis; each motion class is trained separately using Baum and Welch algorithm with a subset of the motion image streams. Then the probability of each test image sequence is calculated by the Viterbi algorithm to determine the HMM Model that gives the highest likelihood probability.

A. Description of Hidden Markov Model

The classical HMM model is shown in Fig 3. HMM can be defined by the following parameters.

- N is number of state. We denoted as S = $\{S_1, S_2, S_3, \ldots, S_N\}$ and the hidden state at time t as $q_t \in S = \{s_1, s_2, s_3, \ldots, s_N\}$
- Π is probability of initial state. $\pi = \{\pi_1, \pi_2, \pi_3, , , , \pi_N\}$, where π_i is probability of initial state s_i
- A is the transition probabilities between states $A = (a_{ij})_{N*N}$, where $a_{ij} = p(q_{t+1} = s_j | q_t = s_i)(1 \le i, j \le N)$ a_{ij} is the probability of reaching state s_j at time t+1 from state s_i at time t.
- B is the emission probability distribution, $B = \{b_i(o_t)\}$, where $b_i(o_t) = p(o_t|s_t)(1 \le i \le N), b_i(o_t)$ is the probability of observation symbol o_t from state s_i at time t.
- T is the number observation symbol in the sequence. We denote this sequence as $O = \{o_1, o_2, o_3, \ldots, o_N\}$

We represent a HMM as $\lambda = \{A, B, \pi\}$. Here, π describes the embedded stochastic process, A describes the Markov chain and B describes the relation between observation symbols and states.

The probability of generating an observation symbol in each case can be calculated by the Gaussian probability density function Eq. 4.

$$b_i(o_t) = \frac{1}{\sqrt{(2\Pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(o_t - \mu_i)^T \Sigma_i^{-1} (o_t - \mu_i)\right) \quad (4)$$

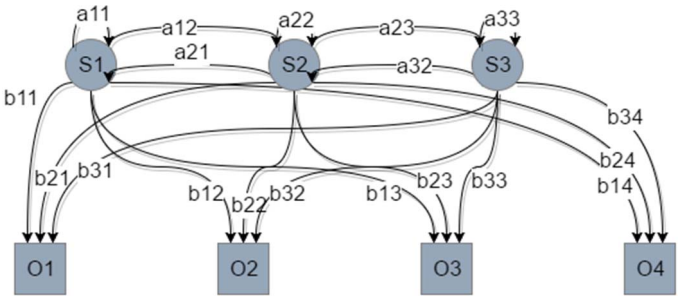


Fig. 3. The classical HMM model

Where, μ_i, Σ_i are the mean and covariance matrices of observations; T operator takes the transpose of a matrix; Σ_i^{-1} is the inverse of Σ_i; d is the dimension of observation symbol o_t.

B. Recognition

Assuming that $\lambda^1, \lambda^1, \lambda^2, \lambda^3, \ldots, \lambda^N$ are HMMs of act_1, act_2, act_3, …, act_N , respectively; and $O = \{o_1, o_2, o_3, \ldots, o_N\}$ is the observation sequence; we calculate the probability of the test sequence using the forward and backward algorithm for each given trained action model [13] as:

$$p(O|\lambda^1), p(O|\lambda^2), p(O|\lambda^3), \ldots, p(O|\lambda^N) \quad (5)$$

The maximum likelihood corresponding to an action is chosen as the best probable action:

$$\text{test\_number} = \underset{1 \le n \le N}{\text{argmax}}(p(O|\lambda^N) \quad (6)$$

C. Training

As we stated above, we used Baum and Welch algorithm to train models. The initial values of the parameters, which are explained in Section A, are determined using the K-Means algorithm. In the K-Means algorithm, we use the Euclidean

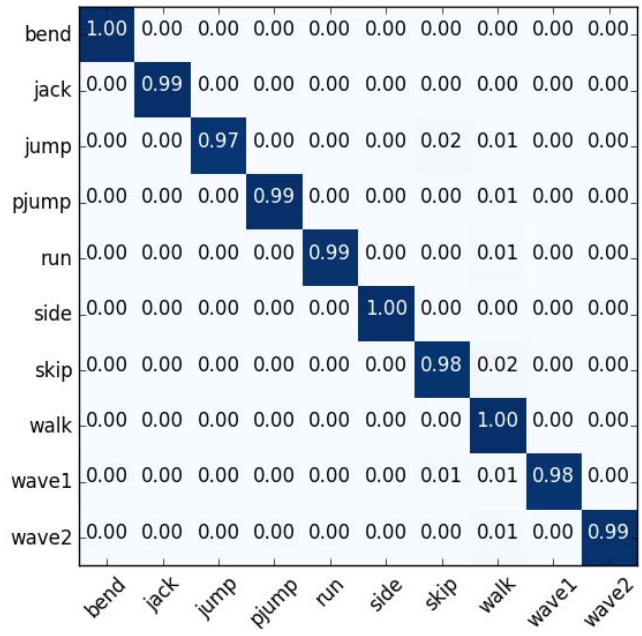
distance between the data and the cluster mean to update the parameters (Eq. 7).

$$d_i(o_t) = \sqrt{(o_t - u_i)^T (o_t - u_i)} \quad (7)$$

Where, d_i is the distance between observation (o_t) and mean (u_i) in cluster i.

The detected cluster centers are used as the initial values of the HMMs. Multiple training sequences are used by the Baum-Welch algorithm to train each action model. In the training stage, in order to solve the problem of very small values, i.e. values close to 0, that are mainly caused by cumulative production; and to avoid the underflow, we utilized the scale factor method [13]. This method helps resolving these problems.

(a)



(b)

Fig. 4. Results on Weizmann dataset: (a) sample frames for different motions, (b) confusion matrix for per-frame classification; the average accuracy is 99%.

V. Results and Discussion

In order to evaluate the proposed method, we used well known Weizmann dataset. In this dataset, there are 93 low resolution (180 x 144 pixels) videos that belongs to 9 different people. In Fig 4(a) we showed some sample frames of the dataset.

In the pre-processing stage of our system, we used the silhouette data that come along with the dataset. During the setup of the experiments, we configured different HMM models by changing the number of *hidden states* to different values among 3 to 8 and evaluated the recognition performance of the system separately for each model. Test results show that the best classification accuracy is obtained when the number of hidden states are configured using 5,6, and 7.

Because the number of training samples for each action is low, the separation of the data into distinct train, test and validation sets is not reasonable. Therefore, we applied *leave-one-out* technique in to evaluate the performance of the proposed solution. In this method, we set one video stream aside for testing and used the rest of the dataset for training. The obtained model is tested with the test stream. This process is repeated for all individual video streams in the dataset. The average accuracy is reported here.

Since the number of test data is limited, i.e. around 10 for each class, we followed a resampling technique like [8] to evaluate the accuracy of the classifications. For this purpose, we resampled a set of 15 frame sequences from each test stream; each is obtained by a 1 frame shift. For example, the first test frame package includes first 15 frames (i.e. frames 1, 2, ..., 15), the second package includes frames 2 to 16 etc. This approach helps us to evaluate the model performance thoroughly with resampling of different parts from a test stream that is not included in the training phase. The recognition performance is depicted with a confusion matrix in Fig. 4(b). The number of hidden states was 5 in the depicted experiment. In this experiment, the average recognition performance is 99%. The performance of the proposed approach is very promising considering the data complexity and very limited number of train data and a lot of distinct resampling of the same motion classes.

In order to compare the effectiveness of the modified MHI with silhouette based approach with HMMs, we performed additional experiments using the same features that are extracted using only the silhouettes and the modified MHIs. We also evaluated the method with a varying number of frames while forming the resampled packages. Table 1 shows the recognition accuracy depending on the number of frames used in forming the test streams and the selected feature extraction method. Our modified, i.e. scale-based, MHI results in better recognition accuracy in each test scenarios. When the video stream is provided as a whole, i.e. without resampling (Table 1, last row), the classification performance reduces approximately 8%; yet the performance of the proposed method is still better than the silhouette based method. The drop in the classification accuracy is due to the insufficient number of test examples, when resampling is not used; even if only one example from each class is misclassified during the tests, the accuracy drops 10%. That is the reason that resampling is usually applied to fairly evaluate the performance of the classifiers when the dataset is not large [8].

Number of Frames Used in Resampling	Accuracy Using only Silhouettes in HMMs	Accuracy Using Modified MHIs
5 frames	74%	98%
10 frames	80%	98%
15 frames	82%	99%
Whole video	86%	91%

Table 1: Recognition rates under different test configurations.

VI. CONCLUSION

In this paper, we present a method that uses a modified MHI as the feature for human motion classification. Hu Moments are calculated on the generated MHIs and HMM was used as the classification method. Using the Weizmann dataset, the proposed method can achieve up to 99% classification accuracy. The results confirm that our approach is comparable with many state of the art methods.

REFERENCES

[1] D. Weinland, R. Ronfard and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Underst.,* pp. 224-241, 2 2011.

[2] 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.

[3] M. A. R. Ahad, J. K. Tan, H. S. Kim and S. Ishikawa, "Temporal Motion Recognition and Segmentation Approach," *Int. J. Imaging Syst. Technol.,* pp. 91-99, June 2009.

[4] J. Yamato, J. Ohya and K. Ishii, "Recognizing Human Action in Time-Sequential Images," in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, 1992.

[5] A. Kale, A. N. Rajagopalan, N. Cuntoor and V. Kruger, "Gait-based Recognition of Humans Using Continuous HMMs," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, Washington, DC, USA, 2002.

[6] A. Sundaresan, A. RowChowdhury and R. Chellappa, "A hidden Markov model based framework for recognition of humans from gait sequences," in *Image Processing*, 2003.

[7] M. A. Mendoza and N. Perez de la Blanca, "HMM-Based Action Recognition Using Contour Histograms," in *Pattern Recognition and Image Analysis*, Girona, Spain, 2007.

[8] F. Martinez-Contreras, C. Orrite-Urunuela and H. Ragheb, "Recognizing Human Actions Using Silhouette-based HMM," in *Advanced Video and Signal Based Surveillance*, 2009.

[9] R. Vezzani, D. Baltieri and R. Cucchiara, "HMM based action recognition with projection histogram features," in *International Conference on Recognizing Patterns in Signals*, 2010.

[10] C.-H. Hsieh, P. S. Huang and M.-D. Tang, "Human Action Recognition Using Silhouette Histogram," in *Proceedings of the Thirty-Fourth Australasian Computer Science Conference - Volume 113*, Perth, Australia, 2011.

[11] H. Kuehne, A. Arslan and T. Serre, "The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities," in *Conference on Computer Vision and Pattern Recognition*, 2014.

[12] L. Wang and D. Suter, "Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[13] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Readings in Speech Recognition*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 1990, pp. 267-296.

[14] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[15] P. Afsar, P. Cortez and H. Santos, "Automatic Human Action Recognition from Video", *IEEE 18th International Conference on Computational Science and Engineering*, Porto, 2015.

[16] J. Lei, G. Li, J. Zhang, Q. Guo and D. Tu, "Continuous action segmentation and recognition using hybrid convolutional neural network-hidden Markov model model", *IET Computer Vision*, vol. 10, no. 6, pp. 527-544, 2016.