IWNEST PUBLISHER

Journal of Industrial Engineering Research

(ISSN: 2077-4559)

Journal home page: http://www.iwnest.com/AACE/

Adaptive sequence of Key Pose Detection for Human Action Recognition

[1]T. Sindhu and [2]C. Nagavani

[1]PG Student, Kamaraj College of engineering and technology, Virudhunagar, Tamilnadu, India.
[2]Assistant professor, Kamaraj College of engineering and technology, Virudhunagar, Tamilnadu, India.

ARTICLE INFO

Article history:
Received 22 February 2015
Accepted 20 March 2015

Keywords:
Human action recognition, Key pose, Key pose sequence, Silhouette, Weizmann dataset.

ABSTRACT

Vision based human action recognition allows to detect and understand meaningful human motion. In this method, human silhouette extraction technique like background subtraction is applied. The extracted human silhouettes are processed in order to obtain the contour based feature. Finding the most characteristic poses among the training data returns the key poses. K-means clustering is an efficient technique to cluster poses based on attributes like distance. With reference to Weizmann dataset, six actions such as running, walking, jumping, bending, waving one hand and waving two hands are considered. The actions are learned by making use of sequences of key poses. The sequence of key poses model the temporal evolution between key poses with respect to the original training sequences. Action recognition is done by means of Minimum Distance Classifier. Action recognition of unknown key poses is matched with known classes of cluster center, then the correct action is recognized. Weizmann dataset deals with single view scenarios with an average success rate of 93%.

To Cite This Article: T. Sindhu and C. Nagavani., Adaptive sequence of Key Pose Detection for Human Action Recognition. J. Ind. Eng. Res*., **1(3),** 10-15, 2015*

INTRODUCTION

Human action recognition has recently become of important interest in recent years due to its direct application and need in Surveillance, Ambient Intelligence, Ambient-Assisted Living (AAL) and Human-Computer Interaction systems. In this paper, a simple but yet very effective approach is presented in order to support accurate human action recognition at the level of basic human motion, like walking, running, jumping, bending, waving one hand and waving two hands. Human action recognition has recently become of important interest due to its wide variety of applications. Improvements in vision-based recognition of short-temporal human behaviors have led to advanced visual surveillance systems [1], as well as sophisticated human-computer interaction (HCI) techniques [2], which are applied to gaming or intelligent environments, among others. Although visual interpretation of human motion, like actions or gestures, has been studied extensively[3], specific requirements of dynamic environments have only sparingly been taken into account [4], [5] Especially at home, robots can be useful supporting several safety and health care scenarios as, for instance, monitoring or mobility assistance [6]. These care services among others, can potentially improve the independent living of the elderly, or serve senior assisted living facilities. Reliable support can only be ensured if robots are intelligent enough to analyse and understand the scenario they perform in and the events that occur, in order to be able to interact appropriately. The application of human action recognition (HAR) to this specific case of HCI comes along with several additional hurdles: 1) since human behaviours are subject to change depending on the specific scenario and actor, and moreover, the behaviours can vary over time. The system has to adapt its knowledge dynamically to recognise new scenarios as, for instance, new actions or new actors. This process needs to happen incrementally, as the system should be able to learn continuously over time without requiring to start from scratch. Furthermore, the recognition capabilities need to evolve and adapt to the new data that needs to be discriminated. Therefore, a simple camera setup should be employed and real time recognition algorithms are required. Constraints are considered choosing a low-cost feature extraction and a state-of-the-art real-time HAR recognition method.

Proposed Method:

Key poses are defined as a set of frames that uniquely distinguishes an action from others. Therefore, the goal of using key poses is to model an action by its most characteristic poses in time. This makes it possible to significantly reduce the problem scale in exemplar-based recognition methods and, at the same time, to avoid

Corresponding Author: T. Sindhu, PG Student, Kamaraj College of engineering and technology, Virudhunagar, Tamilnadu, India.

redundant or superfluous learning. The under-lying idea is that if the human brain is able to recognise what a person is doing based on a few individual images, why should not action recognition methods be able to sustain only on pose information. In this regard, Baysal *et al*. and Cheema *et al*. [4] use no temporal information at all, Thurau and Hlavac model the short-term temporal relation between consecutive key poses with n-grams (trigrams showed good results at acceptable computational cost), and Eweiwi *et al*. take into account the temporal context of a small number of frames by means of obtaining temporal key poses based on MHI. While our approach is very similar to these works at the training stage when applied to a single view, our contribution considers long-term temporal relation between key poses and thus takes advantage of the known temporal evolution of key poses over a whole sequence To achieve an adaptive human action recognition by evolving bag of key poses in order to provide optimized solution with less processing time. Human action recognition can be obtained based on sequences of key poses. A complete overview of the involved stages of the learning process can be seen in figure 1.

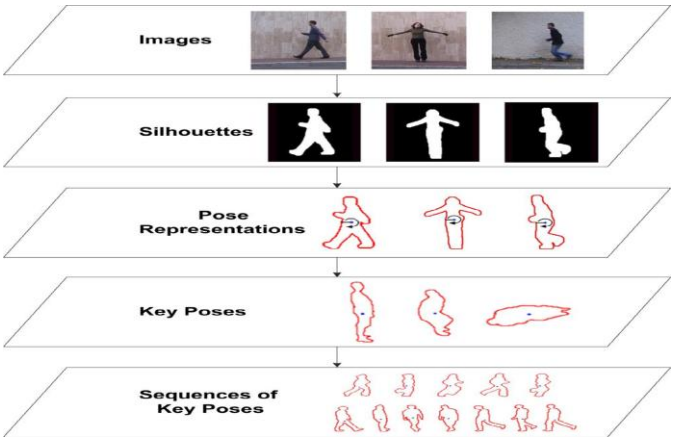

Fig. 1: Overview of the learning process.

Human silhouette extraction technique like background subtraction needs to be applied. Then the extracted human silhouettes are processed in order to obtain the contour-based feature. The silhouette's contour leads to the used pose representation, by means of a distance signal feature which, in conjunction with the model learning approach and the action classification, shows to be a highly efficient technique. At the training stage, the method learns the per class features that make up the most characteristic poses, the so called key poses. Using the ground truth data, the sequences of key poses corresponding to the labeled videos are obtained. These sequences are matched later with the current test sequence based on Minimum Distance Classifier.

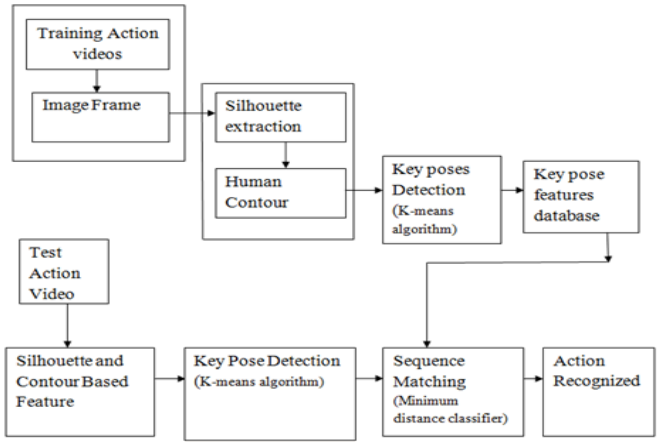

Fig. 2: Block Diagram of the Proposed System.

Methodology:

Action Video into Image Sequences:

The action video of Weizmann dataset consists of different actions acted by various actors. These videos are decomposed into number of image sequences. Each image is composed of RGB Components.

Journal of Industrial Engineering Research, 1(3) May 2015, Pages: 10-15

Silhouette Extraction:

The method relies on a global pose representation based on the contour points of the silhouette. Binary silhouette is obtained by human silhouette extraction techniques, e.g. background subtraction. Using only the contour points and not the whole silhouette is motivated by getting rid of the redundancy that introduces the inside part of the human silhouette, leading therefore to a less expensive feature extraction. In addition, usage of contours avoids the need of morphological pre-processing steps and reduces the sensitivity to small viewpoint variations or lighting changes. Specifically, the contour-based feature has been chosen, which is described briefly in the following.

First, the contour points $P = \{p_1, p_2, \dots, p_n\}$ of the silhouette need to be obtained. For this purpose, contour extraction is applied. Second, the centre of mass $C_m = (x_c, y_c)$ of the silhouette's contour points is calculated with respect to the n number of points:

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, \quad y_c = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1}$$

Third, the distance signal $DS = \{d_1, d_2, \dots,$ is generated by determining the Euclidean distance between each contour point and the centre of mass. Contour points should be considered always in the same order. For instance, the set of points can start at the most left point with equal y-axis value as the centre of mass, and follow a clockwise order

$$d_i = \|\mathrm{C_m} - \mathrm{p_i}\|, \quad \forall \mathrm{i} \in [1 \dots \mathrm{n}] \tag{2}$$

Finally, scale-invariance is achieved by fixing the size of the distance signal, sub-sampling the feature size to a constant length L, and normalising its values to unit sum.

$$\hat{D}S[i] = DS\left[i * \frac{n}{L}\right], \ \forall \mathrm{i} \in [1 \dots \mathrm{L}] \tag{3}$$

$$\bar{D}S[i] = \frac{\hat{D}S[i]}{\max \hat{D}S[i]}, \quad \forall \mathrm{i} \in [1 \dots \mathrm{L}] \tag{4}$$

Key pose Detection:

The first step of the learning process is to process all the frames of the video sequences in order to obtain their pose representation. The per class key poses are learned by means of K-means clustering with Euclidean distance. Hence, the extracted features of all available images of the same action class $samples = \{s_1, s_2, \dots, s_n\}$ are grouped into K clusters; where each cluster $center\ of\ centers = \{c_1, c_2, \dots, c_k\}$ represents a key pose kp as it is a characteristic pose among the training data.

The process of clustering is repeated λ times, so as to avoid local minimum, and the best result is taken .Given that the clustering process returns the corresponding label of each sample $labels = \{l_1, l_2, \dots, l_n\}$ in which l_i stands for the index of the cluster assigned to s_i, clustering results are evaluated with the following compactness metric C:

$$C = \sum_{i=1}^{n} |s_i - c_{l_i}| \tag{5}$$

where the instance with the lowest value is taken as the final result. This key pose learning process is repeated individually for the training samples of each action class. This way, a set of K key poses is obtained for each action class.

Key Pose Features Database:

As stated before, the goal is to learn the long-term temporal evolution of key poses. Consequently, our interest resides on the successive key poses that are involved in an action performance. As the training data is made up of sequences of labelled action performances, the corresponding sequences of key poses can be modelled.

For the pose representation of each frame of a sequence, i.e. $S_{poses} = \{pose_1, pose_2, \dots, pose_n\}$ minimum distance key pose is found. The successive key poses constitute the simplified sequence of known characteristic poses and their evolution. This way, a set of sequences of key poses $S = \{kp_1, kp_2, \dots, kp_n\}$ is obtained for each action class. This decisive step significantly improves exemplar-based action recognition by shifting the training data to a common and known domain.

Journal of Industrial Engineering Research, 1(3) May 2015, Pages: 10-15

Sequence Matching:

At the recognition stage, a final class label output needs to be given. To that end, two steps have to be taken: (1) in the same way as with training sequences, silhouette contour points are processed and their corresponding pose representations are obtained. Recognition techniques based on matching represent each class by a prototype pattern vector. An unknown pattern is assigned to the class to which it is closest in terms of predefined metric. The simplest approach is minimum distance classifier, which computes the Euclidean distance between the unknown. It chooses smallest distance to make a decision.

$$J = arg\ min_j \sum_{i=0}^{n} (x_i - u_{ij})^{\wedge 2} \quad (6)$$

where J – Class of human action label

i– key pose of action index

n- number of cluster for each action

u– Cluster center of human action classes

x- unknown class of action

An action recognition of unknown key poses x is matched with (trained) known classes of human action cluster center u. The distance metric (sum squared distance) computed between unknown key poses of action and known action cluster center and then summed distance between x and u. The summed distance process is repeated for each class of action in database which stores cluster center of known class's action. The unknown action x recognized from selecting minimum distance action among distances belongs to different human action as best matched with Jth action.

Experimental Results:

The Weizmann dataset presented is a single-view (static front-side camera) outdoor dataset. It provides 180 *144 px resolution images of different actions performed by many actors. It has a relatively simple background, provides automatically extracted silhouettes and has become a reference in human action recognition. Actions include bending (bend), jumping (jump), running (run), walking (walk), waving one hand (wave1) and waving two hands (wave2).

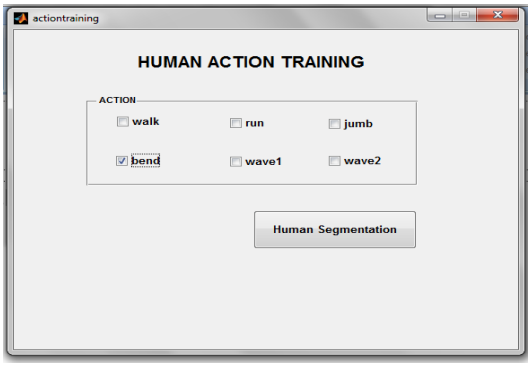

Fig. 3: GUI Design for Human Action Recognition.

In MATLAB GUIs are created using a tool called guide, the GUI Development Environment. This tool allows a programmer to layout the GUI, selecting and aligning the GUI components to be placed in it. Checkboxes may be added to a GUI by using the checkbox tool in the Layout Editor which are used to display on/off options. Figure 3 shows that six actions (Walk, Run, Jump, Bend, Wave1, Wave2) are provided for human segmentation where silhouette extraction take place. Here the actions are selected which will undergo background model and load every person's selected action in the dataset performs background subtraction.

In Figure 4, Binary silhouette is obtained by human silhouette extraction techniques e.g. background subtraction. The inside part of the human silhouette, leading therefore to a less expensive feature extraction. In addition, usage of contours avoids the need of morphological pre-processing steps and reduces the sensitivity to small viewpoint variations or lighting changes.

In Figure 5, selected action key pose will be obtained in the temporal evolution. The per class key poses are learned by means of K-means clustering with Euclidean distance. Hence, the extracted features of all available images of the same action class samples are grouped into K clusters; where each cluster centre of centres represents a key pose as it is a characteristic pose among the training data. The process of clustering is repeated

K times, so as to avoid local minimum, and the best result is taken. Given that the clustering process returns the corresponding label of each sample index of the cluster instance with the lowest value is taken as the final result. This key pose learning process is repeated individually for the training samples of each action class. K key poses is obtained for each action class where k refers to this six actions. Here the long-term temporal evolution of key poses for every single action is obtained. Successive key poses that are involved in an action performance in the training data is made up of sequences of labeled action performances, the corresponding sequences of key poses are modeled. At the recognition stage, a final class label output is obtained.

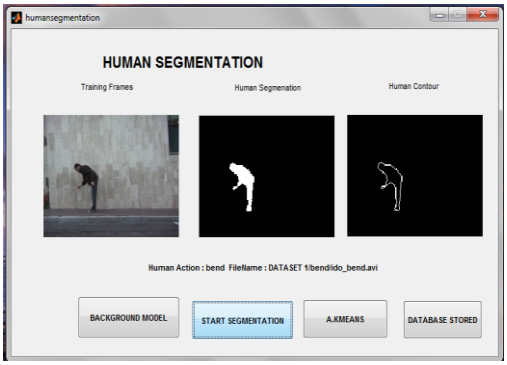

Fig. 4: Human silhouette extraction and contour based feature.

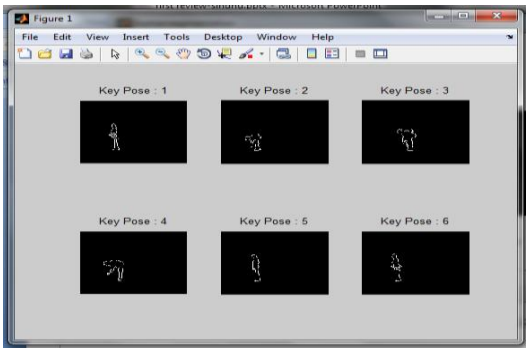

Fig. 5: Key Poses for bend action.

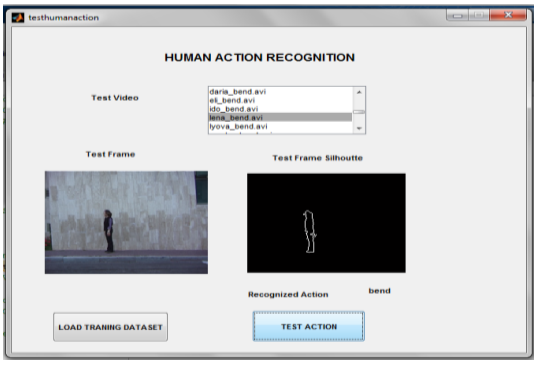

Fig. 6: Recognized action from test video.

An action recognition of unknown key poses is matched with (trained) known classes of human action cluster center. The distance metric (sum squared distance) computed between unknown key poses of action and known action cluster center and then summed distance between trained classes and cluster center. The summed distance process is repeated for each class of action in database which stores cluster center of known class's action. The unknown action recognized from selecting minimum distance action among distances belongs to different human action as matched action. Thus Figure 6 shows the load training dataset in the test video slider window where one video data is selected and so the action is recognized while clicking the test action button.

Journal of Industrial Engineering Research, 1(3) May 2015, Pages: 10-15

Table 1: Confusion matrix of the Weizmann dataset.

	Run	Walk	Jump	Bend	Wave1	Wave2
Run	5/6	1/6				
Walk	1/7	6/7				
Jump			6/6			
Bend				6/6		
Wave1					7/7	
Wave 2						9/9

The above table1 shows while taking a closer look to the misclassifications of sequences from the run action class, it can be seen that the running or walking speed of the actors varied significantly. In addition, some of the actors do not move their arms along when running, which increases even more the similarity between running and walking. It is analysed that a specific misclassification of a run sequence and walk sequence. The ten closest sequences include seven sequences of the right class, which means that, for instance, a Minimum Distance Classifier approach could have worked better in this case.

Conclusion:

In this project, the human silhouette obtained with background subtraction is used as initial input. The silhouette's contour leads to the used pose representation by means of a distance signal feature which in conjunction with the model learning approach and the action classification shows to be a highly efficient technique. Accurate recognition results are obtained using minimum distance classifier in which an action recognition of unknown key poses is matched with known classes of cluster center achieves an average success rate of 93% .The future work include evaluating the method using other available datasets and also with real time videos. Various other actions can also be trained using this recognition method.

REFERENCES

[1] Aggarwal, J., M. Ryoo, 2011. Human activity analysis: A review. ACM Comput. Surv., 43: 16:1-16:43.

[2] Angeles Mendoza, M., N. Pérez de la Blanca, 2007. Hmm-based action recognition using contour histograms. In: Martí, J., Benedí, J.Mendonça, A. Serrat, J. (Eds.), Pattern Recognition and Image Analysis, Lecture Notes in Computer Science, 4477. Springer, Berlin/Heidelberg, pp: 394-401.

[3] Chaaraoui, A.A., P. Climent-Pérez, F. Flórez-Revuelta, 2012. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Exp. Systems Appl., 39: 10873-10888.

[4] Cheema, S., A. Eweiwi, C. Thurau, C. Bauckhage, 2011. Action recognition by learning discriminative key poses. In: IEEE Internat. Conf. on Computer Vision Workshops (ICCV Workshops), pp: 1302-1309.

[5] Cherla, S., K. Kulkarni, A. Kale, V. Ramasubramanian, 2008. Towards fast, view-invariant human action recognition. In: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops CVPRW '08, pp: 1-8.

[6] Hernández, J., A. Montemayor, J. Pantrigo, A. Sánchez, 2011. Human action recognition based on tracking features. In: Ferrández, J., Álvarez Sánchez, J., de la Paz, F., Toledo, F. (Eds.), Foundations on Natural and Artificial Computation, Lecture Notes in Computer Science, 6686: 471-480.

[7] Saghafi, B., D. Rajan, 2012. Human action recognition using pose-based discriminant embedding. Signal Process. Image Commun., 27: 96-111.

[8] Schuldt, C., I. Laptev, B. Caputo, 2004. Recognizing human actions: A local svm approach. In: Proc. of the 17th Internat. Conf. on Pattern Recognition, ICPR 2004, 3: 32-36.

[9] Singh, S., S. Velastin, H. Ragheb, 2010. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: Seventh IEEE Internat. Conf. on Advanced Video and Signal Based Surveillance (AVSS), pp: 8-55.

[10] Wang, J., M. She, S. Nahavandi, A. Kouzani, 2010. A review of vision-based gait recognition methods for human identification. In: Internat. Conf. on Digital Image Computing: Techniques and Applications (DICTA) 2010, pp: 320-327.