View-Invariant Human Action Detection Using Component-Wise HMM of Body Parts

Bhaskar Chakraborty[1], Marco Pedersoli[1], and Jordi Gonzàlez[2]

[1] Computer Vision Center & Dept. de Ciències de la Computació, Edifici O, Campus UAB, 08193 Bellaterra, Spain
[2] Institut de Robòtica i Informàtica Industrial (UPC – CSIC), Llorens i Artigas 4-6, 08028, Barcelona, Spain

Abstract. This paper presents a framework for view-invariant action recognition in image sequences. Feature-based human detection becomes extremely challenging when the agent is being observed from different viewpoints. Besides, similar actions, such as walking and jogging, are hardly distinguishable by considering the human body as a whole. In this work, we have developed a system which detects human body parts under different views and recognize similar actions by learning temporal changes of detected body part components. Firstly, human body part detection is achieved to find separately three components of the human body, namely the head, legs and arms. We incorporate a number of sub-classifiers, each for a specific range of view-point, to detect those body parts. Subsequently, we have extended this approach to distinguish and recognise actions like walking and jogging based on component-wise HMM learning.

1 Introduction

View-invariant action recognition is a constantly expanding research area due to number of applications for surveillance (behaviour analysis), security (pedestrian detection), control (human-computer interfaces), content-based video retrieval, etc. It is, however, a complex and difficult-to-resolve problem because of the enormous differences that exist between individuals, both in the way they move and their physical appearance, view-point and the environment where the action is carried out. Fig. 1 shows some images from the HumanEva database[1], demonstrating the variation of the human poses w.r.t. different camera views and for different actions.

Toward this end, several approaches can be found in the literature [5]. Some approaches are based on holistic body information where no attempt is made to identify individual body parts. However, there are actions which can be better recognized by only considering body parts, such as the dynamics of the legs for walking, running and jogging [2]. Consequently, action recognition can be based on a prior detection of the human body parts [7].

In this context, human body parts should be first detected in the image: authors like [4,8] describe human detection algorithms by probabilistic body part assembly. However, these approaches either use motion information, explicit models, a static camera,

[1] http://vision.cs.brown.edu/humaneva/

F.J. Perales and R.B. Fisher (Eds.): AMDO 2008, LNCS 5098, pp. 208–217, 2008.

Fig. 1. Images from the HumanEva database demonstrate some of the challenges involved with detecting people in still images where the positions of their body parts changes with great variety while performing some actions like walking, jogging and boxing etc

assume a single person in the image, or implement tracking rather than pure detection. Mohan et al. [6] used Haar-Like features and SVM for component-wise object detection. However, Haar-Like features cannot obtain certain special structural features that can be useful to design a view invariant human detection. Moreover, there is no method to select the best features for the SVM so that the performance can be improved by minimizing the feature vector size. Lastly, these works do not cope well for the detection of profile poses.

Due to these difficulties in the view-invariant detection of the human body parts , action recognition has been restricted by analysing the human body as a whole from multiple views. For example, Mendoza and Pérez de la Blanca [3] detect human actions using Hidden Markov Models (HMM) by using the contour-based histogram of the full body. Also, authors in [1] combine shape information and optical flow based on the silhouette to achieve this goal. Likewise, [11] uses the sum of silhouette pixels.

Our approach solves these issues by introducing a framework for view-invariant human detection and subsequently learning the stochastic changes of the body part components to recognize actions like walking and jogging. On the one hand, we use a hierarchy of multiple example-based classifiers for each of different body part-components and view-invariant human detection is achieved by combining the result of those detectors in a meaningful way. Since human action is viewed as a combination of the motion of different body parts, action detection can be analysed as a stochastic process by learning the changes of such components. A HMM based approach is used to learn those changes. In this way we can only consider features from those body part-components which actually taking part into the action e.g. the legs for walking and jogging. We observe that this component-wise stochastic behaviour is good enough to distinguish between similar displacement actions. Our result has been compared with [9]. Lastly, our method for action recognition is also able to detect the direction of motion from the likelihood map obtained from HMM.

This paper is organised as follows. In Section 2 we have presented view-invariant human detection method in detail. Section 3 describes the component wise HMM method towards the detection of human actions. Section 4 reports on the performance of our system. Finally conclusions are drawn in Section 5.

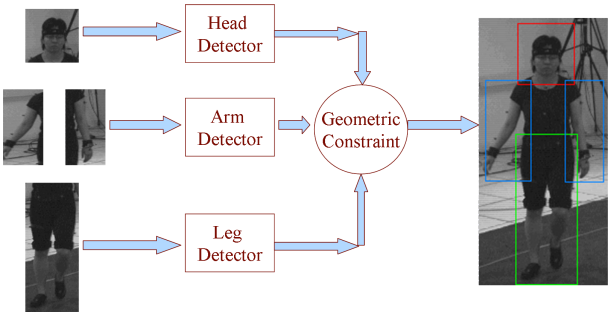

Fig. 2. The overall system architecture of view-invariant human detection. There are three component detectors head, arms and leg. These detectors are combined using geometric constraint for full human detection.

2 View-Invariant Human Detection

The overall structure of the present system is to detect the human action independent of view-point. To achieve this goal our system first detects human body part-components and then combining those body parts to detect the full human. The body parts are combined based on the proper geometric configuration. To ensure the view invariant human detection for each body part more than one detector has been designed and the knowledge of each of those body part detectors are combined finally to increase the robustness of the whole system see Fig. 2.

The component-based human detection has some inherent advantages over existing techniques. A full-body person detector relies solely on visual information and does not take full advantage of the known geometric properties of the human body. The other problem in full human detection is that the system fails to detect the human where body parts are partially occluded. This partial occlusion is accomplished by designing the system, using an appropriate geometric combination algorithm, so that it detects people even if all of their components are not detected.

2.1 Detection of Human Body Parts

The system starts detecting people in images by selecting a 72 x 48 pixels window from the top left corner of the image as an input for head, 184 x 108 pixels window for leg and 124 x 64 for arms. These inputs are then independently classified as either a respective body parts or a non-body part and finally those are combined into a proper geometrical configuration in a 264 x 124 pixels window as a person. All of these candidate regions are processed by the respective component detectors to find the strongest candidate components. Those component-wise window sizes and full human window size comes from HumanEva Database [10], since it is used for training sample creation.

The component detectors process the candidate regions by applying the modified Histogram of Oriented Gradient (HOG) features and then these features become resultant data vector for respective quadratic Support Vector Machine (SVM). Then a standard deviation based feature selection method is applied to take those features where

the standard deviations of oriented gradients are less than one predefined threshold. This threshold has been computed after running the test several times. The strongest candidate component is the one that produces the highest positive raw output, as the component score, when classified by the component classifiers. If the highest component score for a particular component is negative, i.e. the component detector in question did not find a component in the geometrically permissible area, then it is discarded as false positive.

The raw output of an SVM is a rough measure of how well a classified data point fits in with its designated class. The each component where the component score is highest is taken to check whether they are in proper geometrical configuration with the 264 x 124 pixel window. The image itself is processed at several sizes. This allows the system to detect various sizes of people at any location in an image.

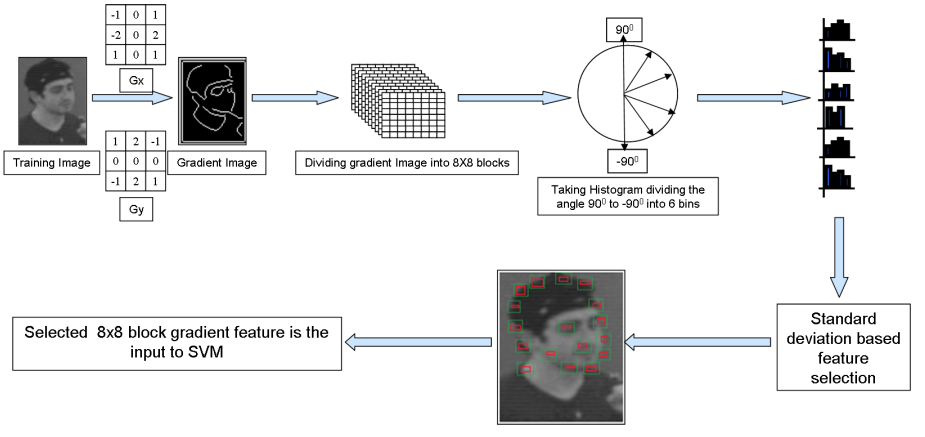

Fig. 3. Feature extraction and selection method from a training image. The training image is divided into several blocks and then HOG features are extracted. Finally standard deviation based feature selection method is applied to obtain feature vector for SVM.

2.2 Feature Extraction and Selection

In our approach for the body part detectors modified HOG feature is used. HOG features are extracted from a 8x8 pixel window from top left hand corner of the training image dividing gradients into 6 bins from -90^o to $+90^o$. In this way that 8 x 8 pixels window slides over the total area reserved for head (72 x 48), leg (184 x 108) or arms (124 x 64). From each of this 8x8 pixel window 6 feature vectors has obtained.

Next step is to select the best 6 feature packs obtained from the method described above. This feature selection method is based on standard deviation (σ). For each position of that 8x8 pixel window the σ is calculated for each of the gradient of that 6 bin taking into account the total training image. Now the σ value has been sorted and those 6 feature packets are taken where the σ is less than a predefined threshold value. In this way the feature size is minimized and those features are fed into the corresponding detector. Fig. 3 shows the general scheme for feature extraction and selection.

Fig. 4. Training samples for each body part detectors e.g. head, arm and leg

Fig. 5. Results of head, arm and leg detector as a validation process. Images with detection of heads with different views, detection of arms and legs are shown.

2.3 Training Body-Part Detectors

To identify human into one particular scale of image each of the individual body part detector has been applied simultaneously. In the present system there are four head detector one leg detector and four arm detector. The four head detectors are for the view angle 45^o to 135^o, 135^o to 225^o, 225^o to 315^o and 315^o to 45^o. For arm, there are four classifiers corresponding different position of arms. Detecting arms is a difficult task since arms have more degree of freedom than the other components. We have used major four poses of the arm with the pose symmetry the detection of other pose

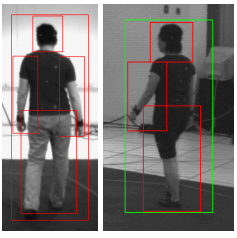

Fig. 6. Validation process of full human detection. Detection of human with all the visible body parts and with profile pose are shown.

possibilities can be achieved. To detect the legs two sub-classifiers have been added to the leg detector, one for open legs and other for profile legs. In the training of component wise detectors 10,000 true positive and 20,000 false positive samples are used.

Fig. 4 shows few training samples of our body part component training database.

The results of those component detectors have combined based on geometrical configuration. Since the head is the most stable part of the body, the geometric combination has been done by first considering the head. Subsequently, the leg component is taken into account and, after that, both arms are combined. We include the Leg Bounding Box (LBB) (of size 184 x 108) after the Head Bounding Box (HBB) is computed (of size 72 x 48) from the head detector. This is done by checking that the x component of the center of the LBB must be within the width of HBB and the y component must be greater than the y component of the centre of the Full Human Bounding Box (FHBB) (of size 264 x 124). We then include arms in a similar way.

We have chosen the result from that sub classifiers which gives the best score result. When a person is moving in a circular path in some cases we can have best score for two sub classifier for arms and other cases provide us only one best score since the arms can be occluded behind the body. We have used some sequences of HumanEva Database [10] to train each body part component and after training we use those detector on the other sequences of such a Database. Fig. 5 shows the result of this validation process. Full detection has been in Fig.6.

3 View-Invariant Human Action Recognition

The aforementioned component wise view invariant human detection is next extended to human action recognition. Toward this end, we learn the stochastic changes of detected body parts using HMM to recognize human actions like walking and jogging. In our system we use HMM for each body part component which has major contribution on the action. From the HMM likely-hood we can recognise and distinguish very similar actions. From the HMM likelihood map we can detect the direction of motion by which we can infer the view point of the person with respect to camera.

3.1 HMM Learning

The feature set used to learn HMM is almost same as that used for SVM of each body part detector. Instead of selecting features here we take mean of each 6 bin angle

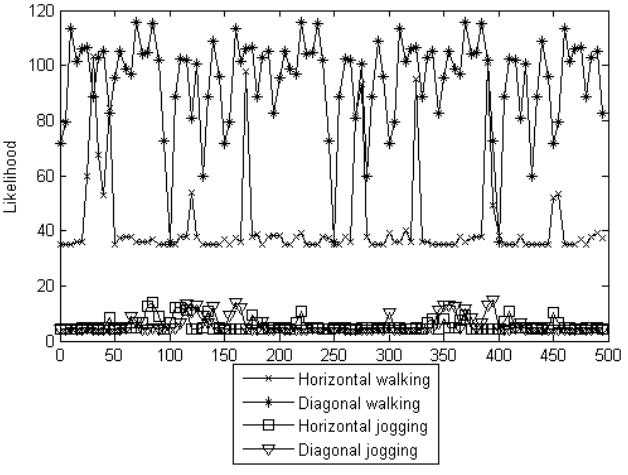

Fig. 7. Likelihood map of walking and jogging actions tested on a walking HMM

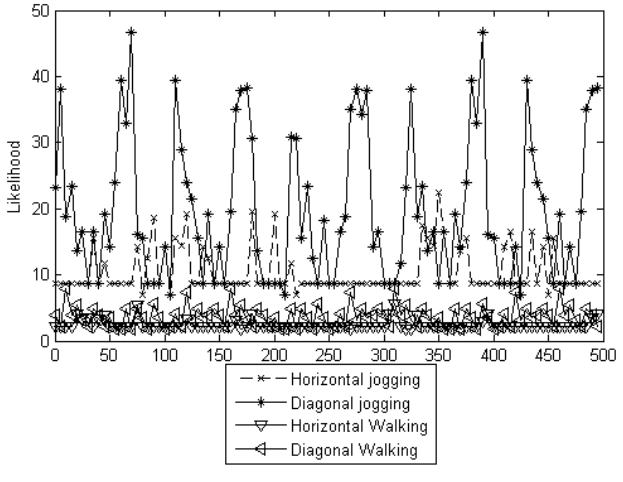

Fig. 8. Likeliihood map of jogging and walking actions tested on a jogging HMM

histogram vector. The significance of taking mean is to get general orientation of body parts which intern signify one pose or series of similar poses in a action e.g. walking. We fit Gaussian Mixture model in to those feature value to obtain different states and key pose of particular action, the pose alphabet of our HMM. To detect actions like walking and jogging we only use legs for HMM and we have found that this HMM is quite good enough to distinguish the similar action like walking and jogging.

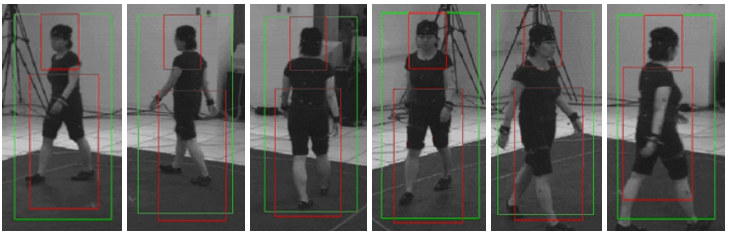

Fig. 9. Frames corresponding to maximum likelihood values of walking

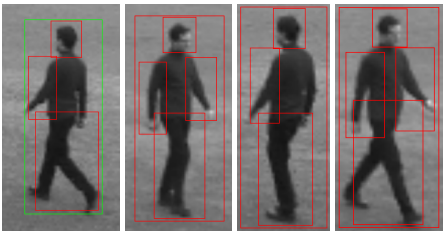

Fig. 10. Performance of full human detection in KTH's database

We have used same training set as of the body part detectors to train component wise HMM. We have chosen a sequence of frames to define one cycle for walking and jogging. After that we map those frame sequence into pose alphabet to obtain one state sequence for HMM learning. We use HumanEval Database as our validation database. This database contains several action sequences like walking, jogging and boxing which are performed by 4 different agents of both sex. We have trained our component HMM using some sequence of that database and tested the same with other similar sequences to validate the HMM learning.

The likelihood map has been computed using the probability values obtained from the HMM for each frame of the test sequence. The maximum values of likelihood map actually describe the end of each walking cycle. We here use two steps starting with right leg as one walking and jogging action cycle. Fig. 9 demonstrates frames corresponding to first 6 maximum likelihood values which clearly shows the end of each walking cycle and also detects the direction of motion.

4 Experimental Results

We have used KTH's [9] Database[2] to test the performance of our system architecture. In that database we have found different types of actions: walking to the right, walking to the left, jogging to the right, jogging to the left etc. These actions were performed outdoors by 25 different people of both sexes with different lighting conditions and shadows. Fig. 10 shows some examples of the performance of our component wise

[2] http://www.nada.kth.se/cvap/actions/

Table 1. Comparison of detection action detection with other two approaches. Column (a) is Local feature and SVM based detection [9], and (b) is our approach.

	Approach (a)	Approach (b)
Walking	83.8	**100.0**
Jogging	60.4	**60.0**

human detection. In this figure we have shown one example of profile pose detection and one example of frontal pose detection. We have tested our component HMMs, which have been learned for the actions performed in circular path, on KTH's database where agents are not in circular path and our approach can detect those actions robustly which justify the view-invariance action detection. We have taken all sequences of the walking and jogging for testing of HMM.

Fig. 7 shows the likelihood of the walking sequence and jogging sequence when applied to walking HMM. Fig. 8 is just the same but for jogging HMM. In both the figures we can observe that a difference in likelihood values. When the walking action sequence is tested in the walking HMM we have found the higher likelihood values and when jogging action sequence is tested on the same the likelihood value decreases and this is same for jogging HMM. The two actions presented in the figure are horizontal and diagonal walking and jogging for two different agents. Table 1 shows detection rates of action recognition using local feature and SVM from [9] and using our approach. We have achieved higher detection rate in walking and similar result in jogging. Since our method considers the stochastic change of body parts having major contribution in action so if there is some similarity in movement of those components then recognition using HMM becomes difficult. We have found some of jogging sequences are misclassified as walking.

5 Conclusions

This paper presents one approach to detect view point invariant human action detection. Human detection in a view-invariant framework is achieved by example-based classifiers for each body part components. Our work performs really well in profile poses.The performance can be improved by adding more training samples and introducing more angular views. One difficultly is that there is not good database for different body part components so building component database is an important task.

In the action detection phase focus is given to distinguish similar actions by considering only the major contributing body parts. This approach is computationally efficient since we have used the same kind of features for the body part detectors and results show that it is enough to consider leg to distinguish those similar actions. We can use this approach to learn HMM for hands and to apply that to detect and distinguish actions like boxing, different gestures e.g. waving.

Acknowledgements

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, and by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER-INGENIO 2010 MIPRCV CSD2007-00018. Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

1. Ahmad, M., Lee, S.: Human action recognition using multi-view image sequence features. In: FGR, pp. 10–12 (2006)
2. Davis, J., Taylor, S.: Analysis and recognition of walking movements. In: Analysis and recognition of walking movements, Quebec, Canada, pp. 11–15 (2002)
3. Mendoza, M., Pérez de la Blanca, N.: Hmm-based action recognition using contour histograms. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (eds.) IbPRIA 2007. LNCS, vol. 4477, pp. 394–401. Springer, Heidelberg (2007)
4. Micilotta, A., Ong, E., Bowden, R.: Detection and tracking of humans by probabilistic body part assembly. In: British Machine Vision Conference, pp. 429–438 (2005)
5. Moeslund, T., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. In: CVIU 104, pp. 90–126 (2006)
6. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. IEEE Transaction on Pattern Analysis and Machine Intelligence 23(4), 349–361 (2001)
7. Park, S., Aggarwal, J.: Semantic-level understanding of human actions and interactions using event hierarchy, 2004. In: CVPR Workshop on Articulated and Non-Rigid Motion, Washington DC, USA (2004)
8. Ramanan, D., Forsyth, D., Zisserman, A.: Tracking people by learning their appearance. IEEE Transaction on PAMI 29(1), 65–81 (2007)
9. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR III, pp. 32–36 (2004)
10. Sigal, L., Black, M.: Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University (2006)
11. Sundaresan, A., RoyChowdhury, A., Chellappa, R.: A hidden markov model based framework for recognition of humans from gait sequences. In: ICIP, pp. 93–96 (2003)