Human action recognition using an ensemble of body-part detectors

Bhaskar Chakraborty, Andrew D. Bagdanov, Jordi Gonzàlez and Xavier Roca
{bhaskar, bagdanov, poal, xavir}@cvc.uab.es
Universitat Autònoma de Barcelona, Computer Vision Center
Campus UAB Edifici O, 08193 Bellaterra, Spain

May 3, 2010

Abstract

This paper describes an approach to human action recognition based on the probabilistic optimization model of body parts using Hidden Markov Model (HMM). Our proposed method is able to distinguish between similar actions by only considering the body parts having major contribution to the actions, for example, legs for walking, jogging and running; hands for boxing, waving and clapping. We apply HMMs to model the stochastic movement of the body-parts for action recognition. The HMM construction requires an ensemble of body-part detectors, followed by grouping of part detections to perform human identification. Three example-based body part detectors are trained to detect three components of the human body: the head, the legs and the arms. These detectors cope with viewpoint changes and self-occlusions through the use of *ten* sub-classifiers that detect body parts under a specific range of viewpoints. Each sub-classifier is a Support Vector Machine (SVM) trained on features selected for the discriminative power for each particular part/viewpoint combination. Grouping of these detections is then performed using a simple geometric constraint model which yields a viewpoint invariant human detector. We test our approach on the most commonly used action dataset, the KTH

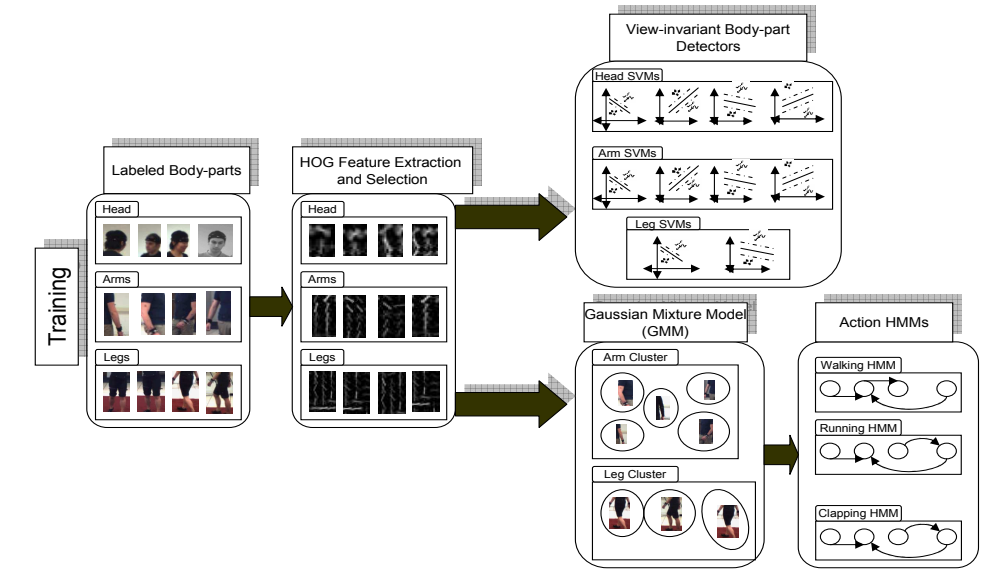
dataset, and obtain promising result, which proves that with a simple and compact representation we can achieve robust recognition of human actions, compared to complex representation.

1 Introduction

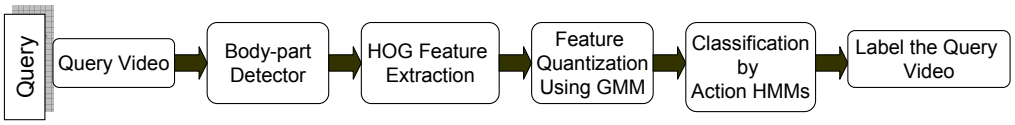
Visual recognition and understanding of human actions are among the most important research areas in Computer Vision (Moeslund, et al. 2006, Shipley & Zacks 2008, Wang, et al. 2003, Turaga, et al. 2008). Good solutions to these problems would yield huge potential for many applications such as the search and structuring of large video archives, video surveillance, human-computer interaction, gesture recognition and video editing. Human detection and action recognition are extremely challenging due to the non-rigid nature of humans in video sequences caused by changes in pose, changing illumination conditions and erratic non-linear motion.

When viewed as a stochastic estimation problem, the critical issue in action recognition becomes the definition and computation of the likelihood, $\Pr(a|H, I)$, of action a given the human H in the image sequence I, since the human H is the main agent performing the action a. The difficulty in working with this likelihood is directly related to the complexity of the joint distribution $\Pr(H, I)$ over all possible human figures H in the image sequence I. Holistic approaches which attempt to model the entire human figure, generally, must resort to very sophisticated and complex models of this joint distribution, resulting in very demanding model estimation and optimization problems.

Basic human actions can be represented using the local motion of individual body parts. Actions like walking, jogging, running, boxing and waving are systematic combinations of the motion of different human body components. From this perspective, it can also be observed that not all body parts contribute equally to all action classes. For example, actions like walking, running and jogging are characterized mostly by the movement of the legs. Boxing, waving and hand clapping, on the other hand, mostly depend on the arms. Our approach is based on these observations and we define the action likelihood $\Pr(a|H, I)$ instead as $\Pr(a|B, I)$, where B is an ensemble of body parts



(a)



(b)

Figure 1: The proposed framework for action recognition based on probabilistic optimization model of body parts using Hidden Markov Models (a) Construction of body-part detectors and Action HMMs (b) Action recognition.

which the human H is composed of. Moreover, the likelihood is further simplified by conditioning actions only on those body parts which contribute most to a particular action. Features from body parts B are used to model the action likelihood $\Pr(a|B, I)$, and optimizing this likelihood over all known actions yields a maximum likelihood estimate of the action in the image sequence.

The ensemble of body part detectors is an important component of our method and we use SVM-based body part detectors over a range of viewpoints to build viewpoint-invariant body part detectors. We model human actions as a repeating chain of body-

part poses (Chakraborty, et al. 2008). Similar poses are grouped together by applying a Gaussian Mixture Model (GMM) on the features of the body parts in order to identify key-poses. These key-poses then serve as a vocabulary of hidden action-states for Hidden Markov Models (HMMs) that model the temporal-stochastic evolution of each action class. Figure 1 shows the overview of our proposed method.

In the next section we review relevant work in the literature. Section 3 describes our probabilistic model for action recognition. In section 4 we describe the viewpoint-invariant part detectors used to define the part ensembles in an image sequence and the HMM-based approach to model the likelihood. Section 4.2 describes the HMM learning process. Experimental results are reported in section 5 and section 6 concludes the paper with some discussions and indications of future research directions.

2 Related work

Several methods for learning and recognizing human actions directly from image measurements have been proposed in the literature (Black & Jepson 1996, Davis & Bobick 1997, Zelnik-Manor & Irani 2001, Chomat & Crowley 1999). These global feature based action recognition techniques mainly employ flow, interest points and silhouette features. These methods may work well under fixed, well-defined conditions, but are not generally applicable, especially for varying viewpoints. For example, accurately estimating the positions of joints in different view-points is an extremely difficult problem and is computationally expensive. Optical flow images are the least affected by appearance but are often too noisy due to inaccuracies in flow calculation. A common issue with interest point detectors is that the detected points are sometimes too few to sufficiently characterize human actions, and hence reduce recognition performance. This issue has been avoided in (Niebles, et al. 2008) by employing a separable linear filter (Dollár, et al. 2005), rather than space-time interest point detectors, to obtain motion features using a quadrature pair of $1D$ temporal Gabor filters. View-invariance in action recognition is addressed in (Yilmaz & Shah 2005, Rao, et al. 2002, Parameswaran & Chellappa 2006). In our

approach we overcome these problems by identifying human body parts under different view-points and explicitly modelling their influences on action interpretation.

For action modelling and classification, many researchers use HMMs due to the sequential-stochastic nature of human actions. The HMMs and AdaBoost are used to recognize $3D$ human action, considering the joint position and pose angles (Fengjun & Ramkant 2006). In (Ahmad & Lee 2006) action recognition is performed using silhouette features and the Cartesian component of optical flow velocity. HMMs are then used to classify actions. Also (Mendoza & de la Blanca 2007), detect human actions using HMMs based on the contour histogram of full bodies. Again, all these methods use features from the whole body to model the HMMs. Human silhouette changes intensely under different view-points and the extracted features from the whole body do not represent the action for all views. In our approach we reduce the complexity of HMM learning by identifying the contributing body parts for an action. In this way action modelling becomes more explicit and can be generalized in different view points.

Although conceptually appealing and promising, the merits of part-based models have not yet been widely recognized in action recognition. One goal of this work is to explore this area. Recent works on part-based event detection use hidden conditional random fields (Wang & Mori 2008), flow based shape feature from body parts (Ke, et al. 2005) and Histograms of Oriented Rectangles (HORs) (Ikizler & Duygulu 2009). These part-based approaches to action recognition employ complex learning techniques and the general concept of view invariance is missing. Although very promising, these approaches have several problems. Learning the parameters of a hidden conditional random field is a very complex procedure and requires a lot of training examples. Methods like HOR need to extract contours and filled silhouettes and the success of the method depends strongly on the quality of those silhouettes. Our approach, in contrast, is built upon example-based body part detectors which are simple to design and yet robust in functionality.

In our method we combine the advantages of full body and part-based action recognition approaches. We use simple SVM-based body part detectors inspired by (Mohan, et al. 2001), but instead of using Haar-like features we use HOGs (Dalal & Triggs 2005).

The disadvantage of Haar-like features is that they are affected by human appearance, while HOG features are designed to extract shape information from the human body contour. Furthermore, we apply a feature selection technique to reduce the dimensionality of the resulting SVMs. Those selected features from the body parts are then used to learn an HMM for each action class. We use Gaussian Mixture Models (GMM) to identify the group of similar body-parts, that have major contribution in an action, to obtain the *action-states* or *key-poses*. View-invariance is addressed by designing multiple sub-classifiers for each body-part, responsible for detecting the body-limbs in different view-points.

3 A probabilistic model for action recognition

The task of human action recognition can be formulated as an optimization problem. Let $A = \{a_1, a_2, \ldots, a_n\}$ denote a set of possible actions, where each a_i denotes a specific action such as walking, running or hand clapping. We write the likelihood of a specific action a_i in a given image sequence I with human H as:

$$\Pr(a_i|H, I) \text{ for } a_i \in A.$$

Given an instance of I with detected human H, a maximum likelihood estimation of the action being performed is:

$$a^* = \underset{a_i \in A}{\operatorname{argmax}} \Pr(a_i|H, I). \tag{1}$$

Rather than holistically modelling the entire human H, we consider it to be an ensemble of detectable body parts:

$$B = \{B_1, B_2, \ldots, B_m\},$$

where each B_i represents one of m-body parts such as the head, legs, arms or torso. The likelihood can now be expressed as:

$$\Pr(a_i|H, I) = \Pr(a_i|\{B_1, B_2, ..., B_m\}, I) \tag{2}$$

Our model strengthen the fact that not all actions depend equally on all body parts. Actions like walking, jogging and running, for example, depend primarily on the legs. Actions like boxing, waving and clapping, on the other hand, depend primarily on the arms. To simplify the likelihood in (Equation 2), we define a dependence function over the set of all subsets of body parts:

$$d(a_i) : A \longrightarrow \mathcal{P}(B),$$

where $\mathcal{P}(B)$ is the power set of B, or $\{c|c \subseteq B\}$.

The set $d(a_i)$ determines the subset of body parts on which action a_i most strongly depends. The likelihood is now approximated as:

$$\Pr(a_i|H, I) \approx \Pr(a_i|d(a_i), I), \tag{3}$$

and the maximum likelihood estimate of the action given an ensemble of body parts and an image sequence is:

$$a^* = \underset{a_i \in A}{\operatorname{argmax}} \Pr(a_i|d(a_i), I). \tag{4}$$

The approximate likelihood in Equation 3 makes explicit the dependence of an action class on a *subset* of body parts. This approximation assumes that the action a_i is independent of the body parts excluded from $d(a_i)$ and thus the likelihood can be computed by simply excluding the irrelevant parts rather than optimizing (or integrating) over them. In the following sections we describe how we model the approximate likelihood in Equation 3 using viewpoint invariant body-part detectors and HMMs over detected features of these body parts, and then from these arrive at the estimate of Equation 4 for action classification.

4 Action modelling by HMM

Human action is a stochastic movement of an ensemble body parts. Moreover, for many actions it is possible to find the body parts that have a major contribution on it. We choose Hidden Markov Models for modelling $\Pr(a_i|d(a_i), I)$ [Equation 3] where $d(a_i)$ is

either B_{legs} or B_{arms}. That is, we use $d(a_i)$ to indicate whether action a_i depends mostly on the arms or on the legs.

An HMM is a collection of finite states connected by transitions. Each state is characterized by two sets of probabilities: a transition probability, and either a discrete output probability distribution or a continuous output probability density function. These functions define the conditional probability of emitting each output symbol from a finite alphabet, conditioned on an unknown state. More formally, it is defined by: (1) A set of states S, with an initial state S_I and a final state S_F; (2) The transition probability matrix, $T = \{t_{ij}\}$, where t_{ij} is the probability of taking the transition from state i to state j and (3) The output probability matrix R. For a discrete HMM, $R = \{r_j(O_k)\}$, where O_k represents a discrete observation symbol. The initial state distribution is $\pi = \{\pi_i\}$, and the complete parameter set of the HMM can be expressed as:

$$\lambda = (T, R, \pi). \tag{5}$$

Here, for each action a_i one discrete HMM is constructed using features from the contributing body parts: B_{legs} or B_{arms}. We obtain the set of hidden states, S, using Gaussian Mixture Model (GMM) on the features from the detected body-parts. The transition probability matrix, T, is learnt and action classification is done after computing the output probability matrix, R, accordingly. Following sections describe the body-part detection and HMM learning.

4.1 Body-part detection

Here, we use three body part detectors for the head, leg and arms respectively. Body part detection is based on *sliding-window* technique. Specific sized rectangular bounding boxes are used for each body part. These bounding boxes are slided over the image-frame, taking the section of the image-frame as an input to the head, leg and arm detectors. These inputs are then independently classified as either a respective body part or a non-body part. In order to prevent possible false positives, those detected components are combined into a proper geometrical configuration into another specific sized bounding

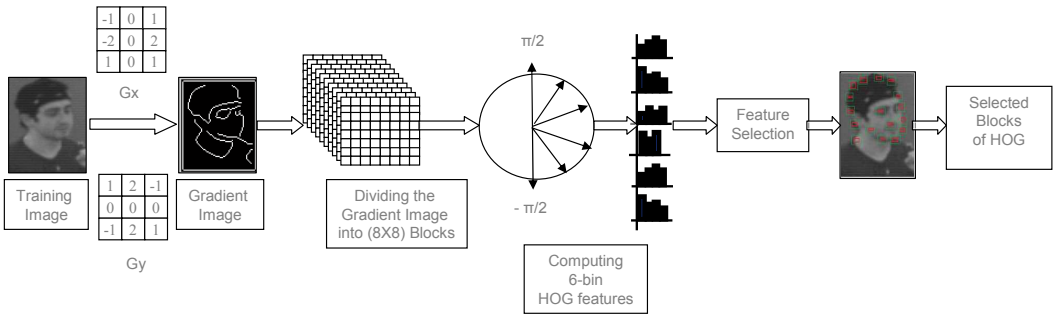


Figure 2: Feature extraction and selection method from a training image. The training image is divided into several blocks and then HOG features are extracted. Finally standard deviation based feature selection method is applied to obtain feature vector for SVM.

box as a full human. All these bounding boxes are obtained from the training samples. Furthermore, the image itself is processed at several scales. This allows the system to be scale invariant.

4.1.1 Features extraction and selection

In our approach (Algorithm 1), labeled training images for each body part detector are divided into (8×8) blocks after applying Sobel mask on them. HOG features are extracted from those blocks and a 6 bin histogram is computed over the angle range $(-\frac{\pi}{2}, +\frac{\pi}{2})$. So, this gives one 6-*dim* feature vector for each of those (8×8) pixel blocks. Next, we select the best feature vector group among all of them. This feature selection method is based on the minimization of the standard deviation (σ) of different dimensions of those feature vectors.

Algorithm 1 Feature extraction for body component SVM

Require: Training images of body component.

Ensure: Features for SVM.

1: **for** every training image **do**
2: Apply Sobel operator
3: Divide Sobel image into (8×8) blocks
4: **for** each blocks **do**
5: Compute 6 bin histogram of HOG features within the range $(-\frac{\pi}{2}, +\frac{\pi}{2})$
6: Keep it in feature array.
7: **end for**
8: Apply *feature selection* algorithm over the feature array.
9: **end for**
10: Selected features are used to learn SVM.

Let there be N number of training image samples of size $W \times H$, divided into n number of (8×8) blocks. For each of these (8×8) blocks we have 6-bin HOG features as, $G = \langle g_1, g_2, \ldots, g_6 \rangle$. We define the standard deviation, σ_{ij}, of ith block and jth bin of HOG feature as:

$$\sigma_{ij} = \left(\frac{1}{N}\right) \sum_{t=1}^{N} \left(g_{ij}^{(t)} - \mu_{ij}\right)^2 \tag{6}$$

where $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, 6$; $t = 1, 2, \ldots, N$; $g_{ij}^{(t)}$ is defined as jth bin HOG feature of ith block of tth training image and μ_{ij} is defined as the *mean* of jth gradient of ith block over all the training images. The values of σ_{ij}, in Equation 6, are sorted and those $6dim$ feature vector packets are taken for which the σ_{ij} is smaller than a predefined threshold. In our case we ran the experiment several times to obtain this threshold empirically. These selected features are used to train SVMs for the body part detectors. Figure 2 shows the feature extraction technique.

4.1.2 Geometric constraints on body-part detection

The obtained outcome of those component detectors are combined based on the geometric constraint of full human body (Algorithm 2) to avoid the false positives obtained from the part detectors. We define the detected body-part component bounding boxes as, Head Bounding Box (R_H), Leg Bounding Box (R_L) and Arm Bounding Box (R_A) obtained from body-part detectors and the full human bounding box (R_F).

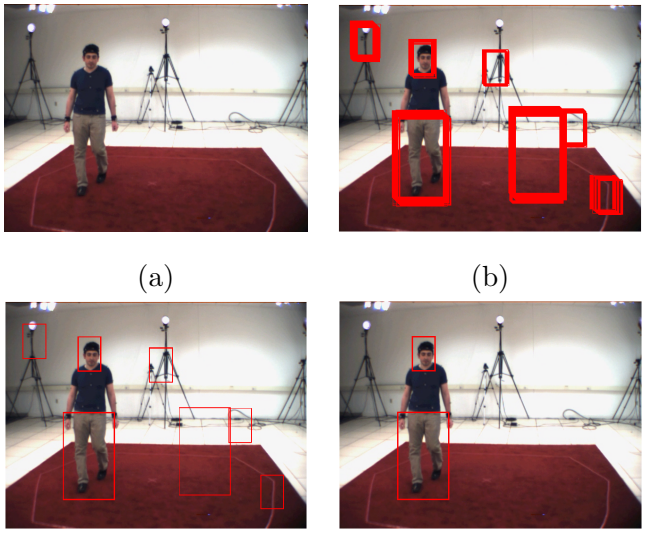
Algorithm 2 Geometric constraints on body-part detection

Require: Detected body-part bounding boxes: R_H, R_L and R_A;
Width and height of full human bounding box: W_{R_F} and H_{R_F}.

Ensure: Full human bounding box: R_F, if geometric constraints are satisfied;
NULL, otherwise.

1: $(X_{C_H}, Y_{C_H}) \leftarrow$ CENTROID(R_H).
2: $(X_{C_L}, Y_{C_L}) \leftarrow$ CENTROID(R_L).
3: **if** $(X_{C_H} - \frac{W_{R_H}}{2}) < X_{C_L} < (X_{C_H} + \frac{W_{R_H}}{2})$ **then**
4: **if** $Y_{C_L} < \frac{H_{R_F}}{2}$ **then**
5: Obtain R_F using $\{(X_{C_H} - \frac{W_{R_F}}{2}), (Y_{C_H} + \frac{H_{R_H}}{2}), W_F, H_F\}$.
6: $(X_{C_F}, Y_{C_F}) \leftarrow$ CENTROID(R_F).
7: $(X_{C_A}, Y_{C_A}) \leftarrow$ CENTROID(R_A).
8: **if** $(X_{C_F} - \frac{W_{R_F}}{2}) < X_{C_A} < (X_{C_F} + \frac{W_{R_F}}{2})$ **then**
9: **if** $Y_{C_H} > Y_{C_A} > Y_{C_F}$ **then**
10: Return(R_F).
11: **end if**
12: **end if**
13: **end if**
14: **else**
15: Return(NULL).
16: **end if**

Removal of false positives are depicted in Figure 3. When the detectors are applied



(a) (b)

(c) (d)

Figure 3: Removal of false positives using geometric constraints of different component detectors. (a) original image (b) detection of head and legs with overlapping bounding boxes (c) after getting single detection window for head and leg including false positives (d) detection of head and leg after removal of false positives using geometric constraint.

in the image there are usually many overlapping detected windows for a particular body component. We obtain a single bounding box from those overlapping detected windows in the following way. If two bounding boxes share more than 70% overlapping area they are merged to obtain one single box. In this way overlapping detection windows are converted into a single one. After that, the above geometric constraint is applied to remove possible false positives and we obtain an ensemble of body-part detectors resulting in a full human detection.

4.2 Construction of action HMMs

For learning an HMM for a particular action, several sequences of frames are chosen and every such sequence is called *action-cycle*/*cycle* for that action. The frames that define one *cycle* depends on the training sequences and the action itself. For example, the number of frames in an action cycle varies when it is performed in a circular path. Let

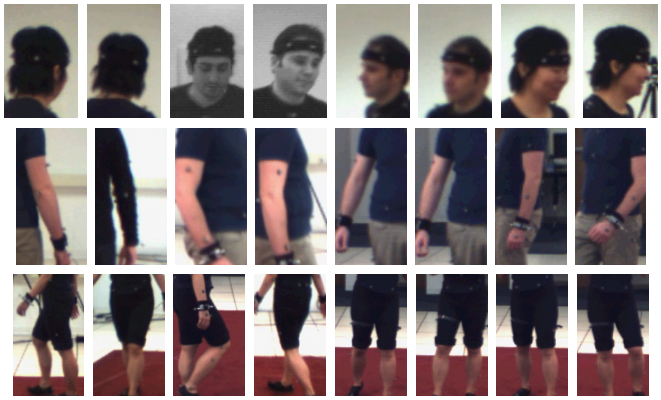


Figure 4: Training dataset for each body-part detectors. It shows training samples for each sub-classifier.

there be M frames inside one action cycle and in each of these frames the body parts are detected using component detectors. Let assume that in the kth frame we detect body part bounding boxes where there are n best $6dim$ feature vectors from the Section: 4.1.1 as $\{G_1^k,\ G_2^k,\ ...,\ G_n^k\}$ where each $G_i^k = \langle g_1^i, g_2^i, \ldots, g_6^i \rangle^k$. Then we compute *mean* over all these G_i^k's to get the features from the kth frame. So, we have $\langle \mu_1, \mu_2, \ldots, \mu_M \rangle$ as a feature to construct the HMM where each of these μ_i's can be expressed as:

$$\mu_i = \left(\frac{1}{M}\right) \sum_{i=1}^{M} \langle g_1^i, g_2^i, \ldots, g_6^i \rangle \,. \tag{7}$$

The significance of taking the *mean* is to get the general orientation of the body part which in turn signifies one pose or a series of similar poses in an action. We fit a Gaussian Mixture Model (GMM) using Expectation Maximization (EM) on these *mean* features to obtain *key-pose* alphabets for a particular action. These *key-poses* are the centre of each cluster obtained from the GMM.

Once we obtain these *key-pose* alphabets, we assign every detected body pose in the frames of each of the action cycles using the *nearest-neighbour* approach to get a sequence of HMM states for them. These state sequences are used to learn the parameters of the HMM in Equation 5. In this way, we obtain the model to compute probability value in Equation 3.

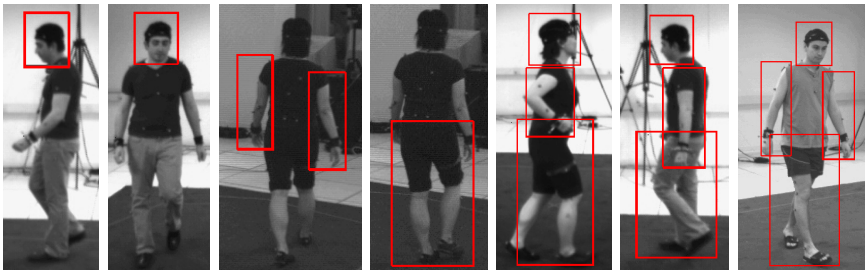


Figure 5: Results of head, arm and leg detection as a validation process. Here detection of profile head and leg is shown. Arm detections are shown where both arms are visible and one is occluded.

For an unknown action, the detected body poses in all frames are mapped into the state sequences in the similar way, but without dividing them into *cycles*, since in such cases the cycle information is not known. $P(O_t|s_t = s)$, the probability of an observation sequence O_t given the state s at time t, is computed using both **Forward** and **Baum-Welch** (Ghahramani 2002, Rabiner 1989) algorithms for every action class HMM. The action class that gives the maximum of these probability values (Equation 4), is the class label of that unknown class.

5 Experiments

We use three publicly available datasets for our experiments on body-part detection and action recognition.

HumanEva dataset[1]**:** This dataset is introduced in (Sigal & Black 2006). We use this to create our human body-part component dataset. HumanEva dataset contains 7 calibrated video sequences (4 grayscale and 3 colour) that are synchronized with $3D$ body poses obtained from a motion capture system. The dataset contains 4 subjects performing 6 common actions (e.g. walking, jogging and gesturing etc.).

KTH dataset[2]**:** This dataset is the most common dataset for the action recognition

[1]http://vision.cs.brown.edu/humaneva/
[2]http://www.nada.kth.se/cvap/actions/

introduced in (Schuldt, et al. 2004). It covers 25 subjects and *four* different recording conditions of the videos. There are *six* actions in this dataset: walking, running, jogging, boxing, clapping and waving. The resolution, (120×160), of this dataset is quite low.

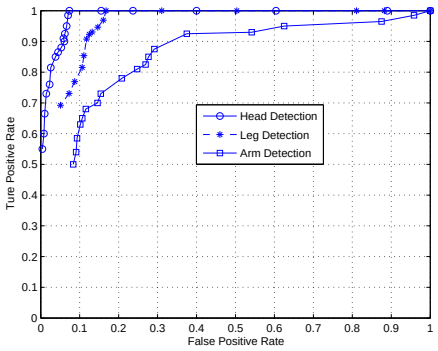
HERMES indoor dataset[3]: This dataset is described in (González, et al. 2009). In this HERMES sequence (2003 frames @ 15 fps, 1392 × 1040 pixels) there are three people in a room. They act in a discussion sequence sitting around a table, where bags are carried, left and picked from the floor, and bottles are carried, left and picked from the vending machine and from the table. In this discussion sequence several agents are involved in different simultaneous grouping, grouped and splitting events, while they are partially or completely occluded.

5.1 Training of the body part detectors

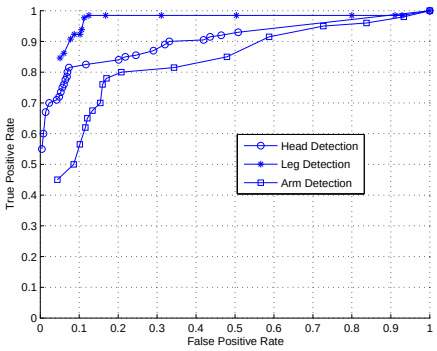
Our system uses four head detectors, two leg detectors and four arm detectors. The four head detectors correspond to view angle ranges $(\frac{\pi}{4}, \frac{3\pi}{4})$, $(\frac{3\pi}{4}, \frac{5\pi}{4})$, $(\frac{5\pi}{4}, \frac{7\pi}{4})$ and $(\frac{7\pi}{4}, \frac{\pi}{4})$. We chose this division so that the consecutive ranges have smooth transition of head poses. For arms, there are four classifiers corresponding to different arm positions, grouped in the same angular views of the head. Detecting arms is a difficult task since arms have more degrees of freedom compared to the other body parts. For each action we use four major arm poses and considering the pose symmetry the detection of other pose variation is achieved. To detect the legs, two sub-classifiers have been added: one representing front(rear) view and the other one for profile views of the legs.

Figure 4 shows some training samples from our body-part dataset. In our method for head, leg and arms detection the bounding box sizes are fixed to (72×48), (184×108) and (124×64) pixels respectively. A (264×124) pixel bounding box is applied for full human. Since the test image is zoomed to various sizes, and in each zoomed image the components of those sizes are searched, the fixed component sizes do not affect scale invariance of human detection. To train each component detector, $10,000$ true positives and $20,000$ false positives selected from the HumanEva dataset are used. The

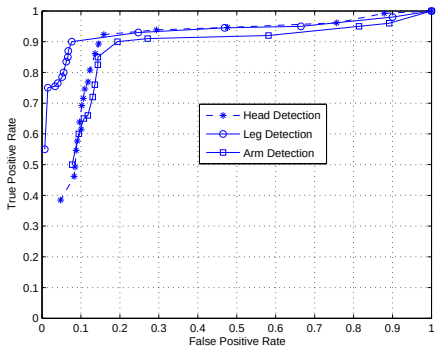
[3]http://iselab.cvc.uab.es/indoor-database



(a) HumanEva dataset



(b) KTH dataset



(c) HERMES indoor sequence

Figure 6: ROC curves for different body part detectors on various datasets. The false alarm rate is the number of false positive detections per window inspected.

sizes of different body-part component bounding boxes are determined after learning the statistics of height and width of each component from all the sequences of HumanEva dataset. A tolerance of 10 pixels is also used along the height and width of the each bounding box.

5.2 Performance Evaluation of Body-part Detection

We performed extensive experiments and quantitative evaluation of the proposed approach of body part detection. We validate our part detector using HumanEva dataset. For a particular body-part detector we usually obtain different detection scores from the

Figure 7: Performance of body part detectors on the KTH dataset. Detection of head, arms and leg shown for profile poses.

sub-classifiers that the detector is composed of. We chose the detection result based on the best detection score among themselves. Figure 5 shows the results of the component detectors in HumanEva dataset. These detections are the strong examples of the view invariant detection since in the HumanEva dataset the agents perform actions in a circular path and our component detector show robust performance in detecting them. To test the performance of body part detectors, KTH Dataset (Schuldt et al. 2004) and HERMES indoor sequence dataset (González et al. 2009) are used.

The Receiver Operating Characteristic (ROC) curves are shown in the Figure 6 for three component detectors: head, legs and arms of different datasets and also for the detection of full human. These ROC curves are generated considering the overall performance of all the sub-classifiers of each group of the three classifiers. ROC analysis reveals that head detection and leg detection are quite accurate. Although there are false positives, but the geometric constraint eliminates them. For the arms, however, the detection rate is not very high since arm pose varies dramatically while performing actions. In the HumanEva and Hermes indoor datasets we obtain high detection rates for full human compared to the KTH dataset.

In the KTH dataset, there are several image sequences where it is impossible to detect arms due to different clothes than the other two datasets. In such cases we are able to detect head and legs, and when they are in proper geometrical arrangement, the full human bounding box is constructed. There are some sequences where only legs are

detected due to low resolution. Figure 7 shows some examples of human detection on the KTH dataset. In those images the detected bounding boxes are found at different scales and they are drawn after rescaling them to 1:1 scale. For low resolution image sequences much information has been lost due to scaling and the Sobel mask hardly found important edges from the particular body parts. Our system works well in high resolution datasets like, the HumanEva (resolution 644×484) and the HERMES indoor sequences (resolution 1392×1040). It gives an average 97% recognition rate for walking, jogging and boxing actions. However, on low resolution datasets like the KTH (resolution 160×120) we achieve lower performance on the actions like jogging and hand clapping. In the KTH dataset there are many cases where the human is visible at the original resolution but when it is zoomed to 640×480 to detect the body-parts, the objects are blurred and detection suffers.

Table 1: Comparison of mean classification accuracy on the KTH dataset.

(Ke et al. 2005)	62.9%
(Wong & Cipolla 2007)	71.16%
(Schuldt et al. 2004)	71.72%
Our Approach	**79.2%**
(Dollár et al. 2005)	81.17%
(Niebles et al. 2008)	81.5%

5.3 Action Recognition

We train the walking, jogging and boxing HMMs on the HumanEva dataset and tested on KTH dataset. But for the actions running, boxing, hand waving and hand clapping we use KTH dataset for both training and testing. In these cases we take random sample of 50% of the video sequences as training and the rest as test. Table 1 shows a comparison of our recognition rate with other methods. The average action recognition rate obtained from our method is quite promising but fails to surpass (Dollár et al. 2005) and (Niebles et al. 2008). But, in both of these cases the training and test are done on

the KTH dataset. We, instead, use HumanEva as a training dataset for the actions, walking, boxing and jogging and test on the KTH dataset.

Table 2 shows the confusion matrix of our approach on the KTH dataset. In the confusion matrix we can see that our approach, clearly, can able to distinguish the leg and arm related actions. We obtain 100% recognition rate for walking and boxing actions which is quite impressive. But the jogging and hand clapping have the poor recognition rate. For jogging, the major confusion occurs with walking (40%) and 23.1% running actions get confused as jogging. In the case of hand clapping we get the confusions with other two actions, the boxing and hand waving. These confusions occur due to the low resolution of the KTH dataset which affects the performance of the body-part detectors.

Table 2: Confusion matrix of action recognition using our component-wise HMM approach. KTH dataset are used.

	Walk	Jog	Run	Box	Wave	Clap
Walk	**100.0**	0.0	0.0	0.0	0.0	0.0
Jog	40.0	**60.0**	0.0	0.0	0.0	0.0
Run	0.0	23.1	**76.9**	0.0	0.0	0.0
Box	0.0	0.0	0.0	**100.0**	0.0	0.0
Wave	0.0	0.0	0.0	20.0	**73.4**	6.6
Clap	0.0	0.0	0.0	13.5	19.8	**66.7**

6 Discussion and Conclusion

This work presents a novel approach for recognizing actions based on a probabilistic optimization model of body limbs. It also includes a view-point invariant human detection is achieved using example-based body-part detectors.

Stochastic changes of body components are modelled using HMM. This method is also able to distinguish very similar actions like walking, jogging and running (considering features from legs); boxing, hand waving and hand clapping (considering features from

hands). The leg movements, sometimes, are very similar in the actions like jogging and running. Also, in some cases there are problems of resolution and contrast which makes it is difficult to distinguish those actions. Actions, involving hand motion also suffer from similar problems, some part of hand waving action look similar to hand clapping, which in tern causes ambiguity.

We observe that there are major confusions occurred in actions jogging and clapping. In the KTH dataset some sequences of jogging action resemblance with walking and running actions. Moreover, other state-of-the the art methods also suffer in the same way. The confusion between clapping and waving could be because of the arm detectors are difficult to design for every possible degree of freedom specially in those two actions there are a variety of arm pose changes. On the hand, arm detectors perform well for boxing since in this case the pose changes do not vary a lot. The advantage of our approach is two fold; first, the body part detection is robust except for resolution limited images. Second, the HMM-based action model is capable of recognizing actions even when the body part detectors fail on some frames of an action sequence.

The performance can be improved in human detection by adding more training samples and introducing more angular views. There is no good dataset for different body-part components, so building a component dataset is an important task. For action recognition, higher order HMM can be applied. Information of other body parts which have minor contribution in the action like, arms for the walking can be included in order to minimize the misclassification rate.

References

M. Ahmad & S.-W. Lee (2006). 'HMM-based Human Action Recognition Using Multi-view Image Sequences'. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, vol. 1, pp. 263–266, Washington, DC, USA. IEEE Computer Society.

M. Black & A. Jepson (1996). 'Eigentracking: Robust matching and tracking of articulated objects using a view-based representation'.

B. Chakraborty, et al. (2008). 'View-Invariant Human Action Detection Using Component-Wise HMM of Body Parts'. In *V Conference on Articulated Motion and Deformable Objects*, pp. 208–217, Andratx, Mallorca, Spain.

O. Chomat & J. L. Crowley (1999). 'Probabilistic recognition of activity using local appearance'. In *In Conference on Computer Vision and Pattern Recognition (CVPR '99), Fort Collins*, pp. 104–109.

N. Dalal & B. Triggs (2005). 'Histograms of Oriented Gradients for Human Detection'. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 01, pp. 886–893, Washington, DC, USA. IEEE Computer Society.

J. W. Davis & A. F. Bobick (1997). 'The Representation and Recognition of Action Using Temporal Templates'. In *CVPR '97: Proceedings of the IEEE Computer Vision and Pattern Recognition*, pp. 928–934.

P. Dollár, et al. (2005). 'Behavior recognition via sparse spatio-temporal features'. pp. 65–72.

L. Fengjun & N. Ramkant (2006). 'Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost'. In *9th European Conference on Computer Vision*, pp. 359–372, Graz, Austria.

Z. Ghahramani (2002). 'An introduction to hidden Markov models and Bayesian networks' pp. 9–42.

J. González, et al. (2009). 'Understanding dynamic scenes based on human sequence evaluation'. *Image and Vision Computing* **27**(10):1433–1444.

N. Ikizler & P. Duygulu (2009). 'Histogram of oriented rectangles: A new pose descriptor for human action recognition'. *Image and Vision Computing* **27**(10):1515–1526.

Y. Ke, et al. (2005). 'Efficient Visual Event Detection using Volumetric Features'. In *IEEE International Conference on Computer Vision*, vol. 1, pp. 166–173.

M. Mendoza & N. P. de la Blanca (2007). 'HMM-Based Action Recognition Using Contour Histograms'. In *Iberian Conference on Pattern Recognition and Image Analysis, Part I*, pp. 394–401, Girona, Spain.

T. Moeslund, et al. (2006). 'A Survey of Advances in Vision-Based Human Motion Capture and Analysis'. In *Computer Vision and Image Understanding*, vol. 8, pp. 231–268.

A. Mohan, et al. (2001). 'Example-based Object Detection in Images by Components'. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **23**(4):349–361.

J. C. Niebles, et al. (2008). 'Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words'. *International Journal of Computer Vision* **79**(3):299–318.

V. Parameswaran & R. Chellappa (2006). 'View Invariance for Human Action Recognition'. *International Journal of Computer Vision* **66**(1):83–101.

L. R. Rabiner (1989). 'A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.'. *Institute of Electrical and Electronics Engineers* **2**:257–286.

C. Rao, et al. (2002). 'View-Invariant Representation and Recognition of Actions'. *International Journal of Computer Vision* **50**(2):203–226.

C. Schuldt, et al. (2004). 'Recognizing Human Actions: a Local SVM Approach.'. In *International Conference on Pattern Recognition*, pp. 32–36, Cambridge, UK.

T. F. Shipley & J. M. Zacks (2008). In *"Understanding Events From Perception to Action"*. Oxford University Press.

L. Sigal & M. J. Black (2006). 'HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion'. In *Technical Report CS-06-08, Brown University.*

P. Turaga, et al. (2008). 'Machine Recognition of Human Activities: A Survey'. *Circuits and Systems for Video Technology, IEEE Transactions on* **18**(11):1473–1488.

L. Wang, et al. (2003). 'Recent developments in human motion analysis'. *Pattern Recognition* **36**(3):585–601.

Y. Wang & G. Mori (2008). 'Learning a discriminative hidden part model for human action recognition'. In *Advances in Neural Information Processing Systems*, vol. 21, pp. 1721–1728. MIT Press.

S. Wong & R. Cipolla (2007). 'Extracting Spatiotemporal Interest Points using Global Information'. In *11th IEEE International Conference on Computer Vision*, pp. 1–8.

A. Yilmaz & M. Shah (2005). 'Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras'. *IEEE International Conference on Computer Vision* **1**:150–157.

L. Zelnik-Manor & M. Irani (2001). 'Event-based analysis in video'. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 123–130.