On Supervised Human Activity Analysis for Structured Environments

Banafshe Arbab-Zavar, Imed Bouchrika, John N. Carter, and Mark S. Nixon

School of Electronics and Computer Science, University of Southampton,
Southampton, SO17 1BJ, UK

Abstract. We consider the problem of developing an automated visual solution for detecting human activities within industrial environments. This has been performed using an overhead view. This view was chosen over more conventional oblique views as it does not suffer from occlusion, but still retains powerful cues about the activity of individuals. A simple blob tracker has been used to track the most significant moving parts i.e. human beings. The output of the tracking stage was manually labelled into 4 distinct categories: walking; carrying; handling and standing still which are taken together from the basic building blocks of a higher work flow description. These were used to train a decision tree using one subset of the data. A separate training set is used to learn the patterns in the activity sequences by Hidden Markov Models (HMM). On independent testing, the HMM models are applied to analyse and modify the sequence of activities predicted by the decision tree.

1 Introduction

Automated detection and tracking human activities within video sequences is a challenging problem which finds application in monitoring and surveillance systems as well as human-machine interactions. Recently, parallel to advances in video camera technologies as well as storage and computation capabilities, there has been an increase of research interest in the area of human action recognition in the computer vision community.

Various types of features have been proposed for this task. Parameswaran et al. [1] detects a number of body joints and analyses their trajectories in 2D invariance space. Detecting and tracking body parts have also been used to infer the higher level activities [2,3]. In this, state space methods have been employed to analyse a sequence of lower level events. Rather than tracking various body parts or joints, other methods have used holistic features [4], and local spatio-temporal interest points [5,6]. Sun et al. [7] experimented with both holistic features and local interest points and showed that the effectiveness of these features depends on the characteristics of the dataset. Apart from the approach to recognize the actions, various proposed methods differ significantly in terms of: the activities which they aim to recognize; camera angle; background properties and image quality.

G. Bebis et al. (Eds.): ISVC 2010, Part III, LNCS 6455, pp. 625–634, 2010.

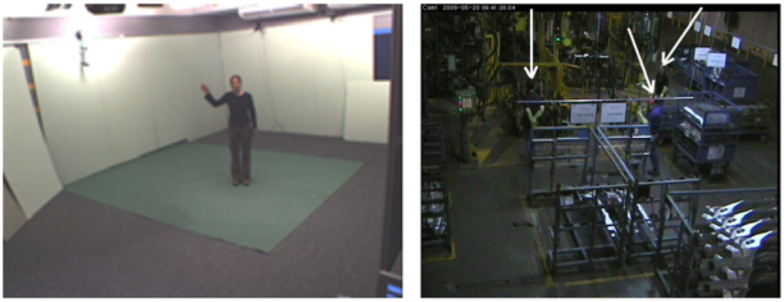
Despite various approaches to human action recognition, the datasets which are used are mainly well-constrained and occlusion-free, which are far from what may be observed by a surveillance camera. The side and frontal views appear to be the dominant view angles for these analyses. In this paper, we will consider the problem of human action recognition from a continuous feed of video capturing from a top view panoramic camera monitoring an industrial plant. In this, the conventional view angles are subject to unworkable levels of occlusion. Multiple subjects may appear on each frame while the background is also changing. We analyse four action categories: walking; carrying; handling and standing still which are taken together from the basic building blocks of a higher level work flow analysis. We use a simple blob tracker to detect the main moving parts i.e. human beings. Various shape-based and motion-based features are then extracted for the action recognition. These features are extracted from a 10 frames long window. A binary decision tree which uses the features selected via the ASFFS feature selection algorithm provides initial prediction for the activity which is being performed. Exploiting our continuous video data, we can then analyse the validity of the predicted sequence of activities and their stability over time. Note that given the nature of the data, which captures a stage in an industrial work flow, there are patterns in the sequences of activities, and these activities are also spatially constrained. The sequence of predicted activities is analysed by HMM models which have been trained on a separate training data.

2 Human Activity Analysis

2.1 On Viewpoint Selection

There has been very little work in recognition of human activities for the top view. Parameswaran et al. [1] model actions in terms of view-invariant canonical body poses and trajectories in 2D invariance space. On a small dataset they obtained 12/18 true classifications for top view, which is similar to what they achieve for frontal view, while side view obtains a better classification rate. It has been repeatedly mentioned that the top view obtains the lowest classification rates as compared to the other views. The recognition rates of 33.6% [8] and 66.1% [9] have been reported on the IXMAS dataset [8], while the recognition rates from the other views average around 63.9% and 74.1% respectively. These methods are mainly concerned with achieving a viewpoint invariance, which could handle images from the top view as well as the frontal and side views. Lv et al. [10] offer better results for single camera recognition, with a 78.4% recognition rate for top view and an average rate of 81.3% for the other views. In this, they search for the best match to the input sequence among synthetic 2D human pose models for different actions rendered from a wide range of viewpoints. For comparison purposes, note that the IXMAS dataset is a well-constrained dataset with a single moving subject at each frame. Figure 1 shows the front/side view images from IXMAS and our dataset.

Our data is from video cameras monitoring an industrial plant. Note the severely cluttered scene and the level of occlusion for the side/front view



(a) IXMAS with one person

(b) Our data with three people

Fig. 1. Compare the frontal view from the IXMAS data with a side/front view of our data

camera. In fact, our dataset is characterized by the severe levels of occlusion which affects all the camera views except for the overhead camera (see Figure 2(a)). We propose that this scenario is likely to arise in many surveillance systems specially within similar industrial environments. Thus, for detecting human activities, we have chosen to use the overhead view, which is not affected by occlusion. Unlike the methods mentioned above, we propose to design methods which primarily capture the information from the top view.

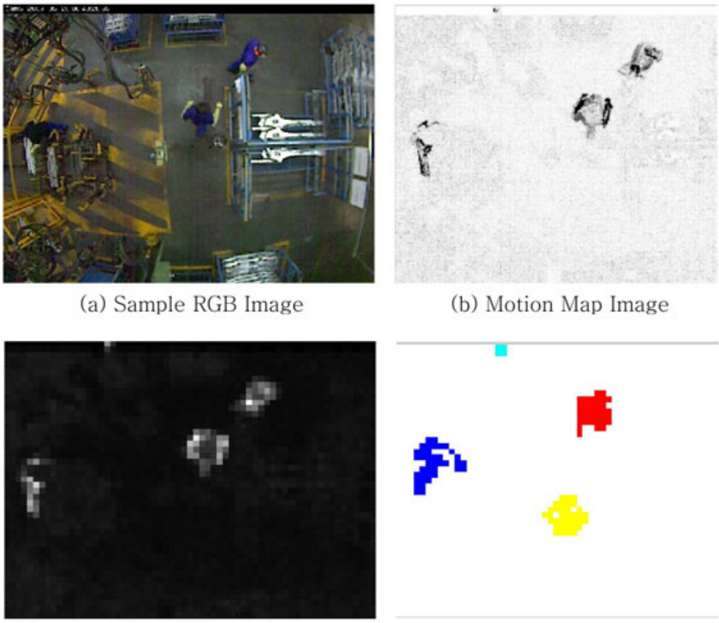
2.2 Human Detection and Tracking

In order to derive a set of features for the classification of human behaviour, first we need to determine the number of individual workers and their bounding boxes at each frame. Considering that the humans are the main moving objects in these videos, we apply frame differencing to compute the motion map image based on the change detection for the inter-frame difference. The motion map M_t at frame t is computed as the absolute difference of two consecutive frames I_t and I_{t+1} as:

$$M_t = ||I_t - I_{t+1}||. \tag{1}$$

An accumulation process is thereafter applied on the motion map by dividing the map into a grid with a bin of size 10×10 pixels. Summing the values in each bin, a threshold is then applied to the accumulated image. Finally, Connected Component Analysis is applied to derive the larger blobs which correspond to the human workers. Figure 2 shows the various stages of detection.

In order to track multiple objects across consecutive frames, we propose to model the moving objects as temporal templates characterized by a combination of three basic features: the size, the centroid position, and the aspect ratio of height to width of the bounding box. Shape-based features are considered because they involve low-complexity computation and yet they enjoy robust characteristics. A number of constraints are imposed on these features to handle complex cases of split and merge of moving regions as well as exit and entry into the scene.



(a) Sample RGB Image

(b) Motion Map Image

(c) Accumulated data

(d) Segmented Image

Fig. 2. Four stages of the human detection

2.3 Feature Extraction

A label — walking; carrying; handling and standing — will be assigned to each detected blob at each frame determining its activity. However, to arrive at this label, we consider a period of ten consecutive frames in which the individual is detected. Schindler et al. [11] have also asked the question: how many frames is required for human action recognition? They showed that for the set of actions which they were aiming to recognize a short sequence of 5-7 frames can obtain a performance similar to the analysis of the entire sequence. However, an analysis of recognition from top view has not been considered in this work.

Since both the temporal features and the shape of the moving blob include cues as to the activity which is being performed, we extract both shape-based and motion-based features for the detected blobs. These features are:

- Hu Invariant Moments [12], which are seven moments providing a global description of the shape. These are translation, scale and rotation invariant.
- Region-based properties: area, diameter, etc.
- Motion-based: speed and the direction of speed.

The mean value, within the 10-frame window, for each of these features is considered. However, as well as the mean, the changes in the value of these parameters can provide discriminant cues. Therefore, the sequence of values for each feature is analysed for the frequency of changes via discrete Fourier transform. Magnitude and phase in different frequencies are then added to the feature vector. Let ϕ be the set of all shape and motion based features which have been listed above. Let $f_i(n)$ be the feature f_i, where $f_i \in \phi$, detected on the n^{th} frame of the 10-frame period analysed for each sample. F_i is the set of features f_i across the 10 frames interval;

$$F_i = \{f_i(n)\} \ , \ n = 1..10 \ . \tag{2}$$

Let $\mathcal{F}$ denote discrete Fourier transform.

$$X_i = \mathcal{F}(F_i)$$

$$A_i(n) = |X_i(n)| \ , \ \varphi_i(n) = arg(X_i(n)) \tag{3}$$

where A_i and φ_i denote the magnitude and phase in different frequencies. Thereby the feature vector V is generated for each sample as:

$$V = \{A_i(n), \ \varphi_i(n), \ \mu_i, \ \sigma_i \ \} \ , \ i = 1..|\phi| \ , \ n = 1..10 \tag{4}$$

where μ and σ denote the mean and the standard deviation of the feature values. Thereby, a large and variant feature vector with 345 features is created.

As discussed in section 2.1, our industrial framework introduces extra complications in terms of limitations in quality and control over the acquired samples. The occlusion in the conventional oblique views have been discussed and a solution was offered through the use of the top view. However, other difficulties include poor image quality, noisy environment, camera shakes, changes in lighting and, in the case of our dataset, a practical issue with random phases of temporal inconsistency resulted from dropped frames. Thereby robustness to noise and outliers appears a desirable feature. Due to the composite nature of our 345-dimensional feature space and that various feature types are susceptible to different levels of corruption in noise, a feature subset selection method is employed to derive the discriminative cues whilst removing the corrupted and irrelevant features. This is explained in more detail in the next section.

2.4 Supervised Binary Tree Classification

A binary decision tree approach has been adopted for the classification. The taxonomy is being structured for the different types of activities as shown in Figure 3. The output of the tracking stage was manually labelled into four distinct categories: walking; carrying; handling and standing still. Note that the two categories: object and noise (see Figure 3) have not yet been considered and only the detected humans are considered for activity recognition. A feature subset selection is being applied at each node of the tree to derive the best features at the selected node. For this, we use the Adaptive Sequential Forward Floating Selection (ASFFS) [13] algorithm. This is an improved version of the SFFS method which was shown by Jain et al. [14] to outperform the other tested suboptimal methods. Using the gallery of manually labelled activities and the selected subset of features, a k-nearest neighbour is applied at each node to obtain a classification.

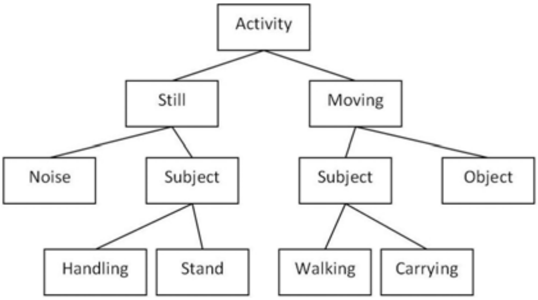


Fig. 3. The binary tree structure for initial classification of activities

2.5 Spatially Specific HMMs for Sequence Analysis

The classification of activities based on visual characteristics and motion features has limitations. For example, carrying might appear as walking if the part being carried is too small. However, there are logical and structural patterns within a sequence of activities, which can be exploited to evaluate the validity of a sequence of predictions. Figure 4 shows some correctly classified activities in individual frames and how they relate to form a work flow within our dataset. The main pattern being displayed here is picking up a part from a rack and placing it on the welding cell. About half of the activities detected fall within this pattern while the rest of activities include walking and standing at arbitrary directions and locations as well as occasional handling of objects.

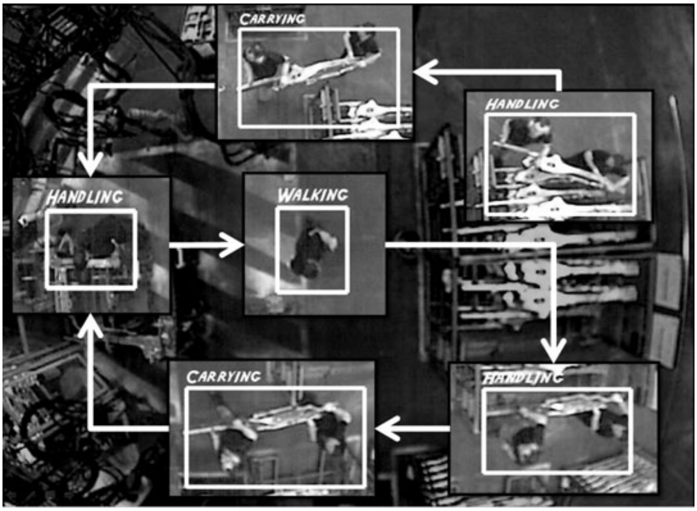


Fig. 4. Human activities detected on individual frames and superimposed on a still image of the plant showing the patterns in the work flow

Hidden Markov Models (HMM) can model underling dependencies within a sequence of unobserved states. As such, they appear an attractive method to analyse the patterns of activities within our data. A HMM with the structure shown in Figure 5 is used to learn the probabilities. In this, the hidden states are the activities — walking; carrying; handling and standing — and the visible states or observations are the predictions obtained by our binary tree classifier. Let the set of predictions, A, by the decision tree be:

$$A = \{a_t\} \ , \ t = 1..T \tag{5}$$

where a_t is the predicted action at time t, and T is the duration of the sequence. Let H be the set of hidden states for our HMM models. Given the set of predictions, the probability of being in state α at time t, denoted by $S_t = \alpha$, is recursively calculated by:

$$\begin{aligned} &P(S_t = \alpha|A) = \\ &\qquad P(A_t = a_t|S_t = \alpha) \cdot \max_{\beta \in H}[P(S_t = \alpha|S_{t-1} = \beta) \cdot P(S_{t-1} = \beta|A)] \end{aligned} \tag{6}$$

In this, the probabilities $P(A_t = a_t|S_t = \alpha)$ and $P(S_t = \alpha|S_{t-1} = \beta)$ are given by the HMM model.

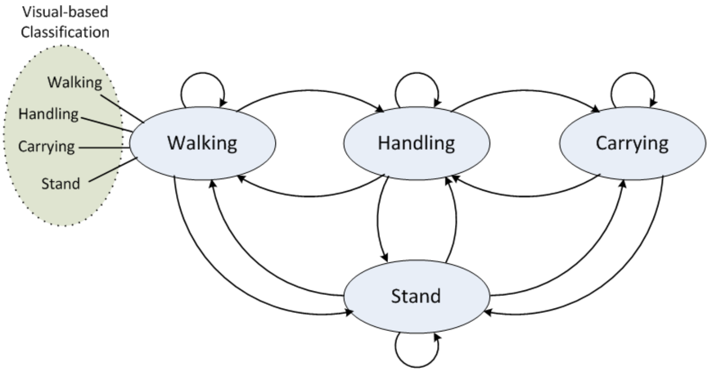


Fig. 5. HMM structure; the hidden states are the activities and the observations at each state are the initial classifications obtained by the binary tree classification

Our data also imposes that there is a spatial dependency regarding the expectation of various activities. Given a low-level knowledge of the work flows, we have identified three main areas wherein the expectation of occurrence and the sequential order of activities differs significantly: i) the racks (pick up area); ii) the welding cell (put down area); iii) walk ways. Figure 6 highlights these three areas. A hysteresis thresholding improves the stability in determining the area of each sample at each frame. An HMM model has been trained for each of these areas. A separate, manually labelled training set is used for training the HMMs.

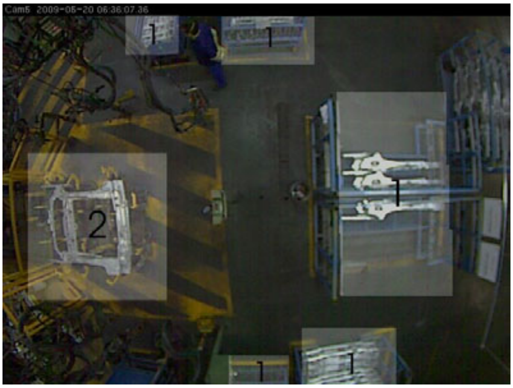


Fig. 6. The three areas for which different HMM models are generated are highlighted. Area 1 is the racks; area 2 is the welding cell; and the remaining are the walk ways.

3 Experimental Analysis

A total of 170,000 frames have been used in our experiments. The frames are of form shown in Figure 2(a), which is the overhead view of the industrial plant. Multiple moving blobs might be detected at each frame. In average, there are 1,613 samples in each 10,000 frames; a sample being a detected blob in a frame which has been also detected in five frames prior to and in five frames after the current frame. From this, 50,000 frames have been used for feature subset selection. These frames also constitute the gallery to which a sample is compared. 60,000 frames are used in training of the HMM models. The remaining 60,000 frames are used for testing. The output from the tracking is manually labelled into: walking; carrying; handling and standing for all the test and training data.

Figure 7 shows the correct classification rates (CCR) on six separate test sets. Each test set consists of 10,000 consecutive frames. The CCRs for three approaches are shown:

- Binary tree classification: as described in section 2.4
- Binary tree classification with smoothing: In this, each activity which does not persist for more than 5 frames is set to the previous stable activity.
- Binary tree classification with HMM : The sequence of predictions from the binary tree is examined and is set to the most probable underling sequence using the HMMs.

Clearly, HMM improves the performance in all the test sets. Table 1 gives the details of the recognition performance. Note that these CCRs, which are determined by comparing the auto-classifications to the manual labels at each frame and counting the miss-matches, are the lower-bounds for classification, since there is an ambiguity in labelling the activities in a frame by frame basis. Also, there is an uncertainty in determining when one activity ends and the next one starts. For example, we have manually evaluated the classification labels obtained by the binary tree classifier on testset 4. This manual evaluation shows

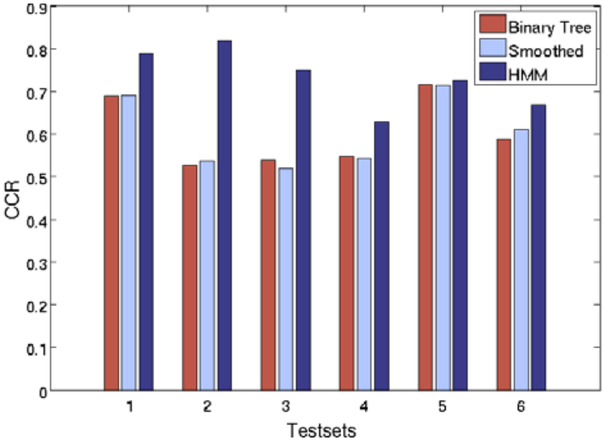


Fig. 7. The CCRs of activity detection on six testsets, each including 10,000 frames

that the assigned class for each sample is correct in 67% of the times, while the auto-evaluation shows a 55% CCR. A more credible evaluation of performance would be via evaluating the accuracy in detecting the higher level work patterns using these activities. The higher level work flows are deterministic in nature and are easier to label manually. Detecting the work flow patterns is the main avenue for our future research.

Table 1. Correct classification rates (CCR) on various testsets

	Testset 1	Testset 2	Testset 3	Testset 4	Testset 5	Testset 6
Binary tree	1617/2345 68.96%	350/665 52.63%	694/1286 53.97%	1316/2403 54.76%	138/193 71.50%	1638/2784 58.84%
Smoothed	1538/2226 69.09%	321/597 53.77%	652/1252 52.08%	1242/2288 54.28%	132/185 71.35%	1625/2658 61.14%
HMM	1851/2345 78.93%	544/665 81.80%	965/1286 75.04%	1508/2403 62.75%	140/193 72.54%	1863/2784 66.92%

4 Conclusions

In this paper we have considered the problem of automatically detecting human activities in industrial environments. The top panoramic view have been chosen for the analysis since this view is less likely to be affected by occlusion. At present there is a dearth of analysis of imagery derived from overhead views. This is well suited to industrial environments, and might extend to indoor surveillance scenarios. Shape-based and motion-based features have been used to derive a classification based on a binary-tree structure of activities which are taken from a higher level work flow. Classifying the activities based on the visual cues has limitations were the activities appear similar. A large improvement is observed when we employ Hidden Markov Models to analyse the sequence of detected activities. Having learned the patterns in activity sequences, these models offer a more viable and stable sequence of predictions based on the initial classification and their spatial properties. Considering the origin of our data which shows a period in a manufacturing cycle, the main avenue for our future research is detecting these higher level work flows.

References

1. Parameswaran, V., Chellappa, R.: View invariance for human action recognition. IJCV 66, 83–101 (2006)
2. Ryoo, M.S., Aggarwal, J.K.: Semantic representation and recognition of continued and recursive human activities. IJCV 82, 1–24 (2009)
3. İkizler, N., Forsyth, D.A.: Searching for complex human activities with no visual examples. IJCV 80, 337–357 (2008)
4. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. TPAMI 23, 257–267 (2001)
5. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV 79, 299–318 (2008)
6. Laptev, I., Caputo, B., Schüldt, C., Lindeberg, T.: Local velocity-adapted motion events for spatio-temporal recognition. CVIU 108, 207–229 (2007)
7. Sun, X., Chen, M., Hauptmann, A.: Action recognition via local descriptors and holistic features. In: CVPR, Miami, USA (2009)
8. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3d exemplars. In: ICCV, Rio de Janeiro, Brazil (2007)
9. Junejo, I.N., Dexter, E., Laptev, I., Pérez, P.: Cross-view action recognition from temporal self-similarities. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 293–306. Springer, Heidelberg (2008)
10. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: CVPR, Minneapolis, MN, USA (2007)
11. Schindler, K., van Gool, L.: Action snippets: How many frames does human action recognition require? In: CVPR, Anchorage, AK (2008)
12. Hu, M.: Visual pattern recognition by moment invariants. IEEE Transactions on Information Theory 8, 179–187 (1962)
13. Somol, P., Pudil, P., Novovičová, J., Paclík, P.: Adaptive floating search methods in feature selection. Pattern Recognition Letters 20, 1157–1163 (1999)
14. Jain, A., Zongker, D.: Feature selection: Evaluation, application, and small sample performance. TPAMI 19, 153–158 (1997)