Project Report Group#29

Classifying people in a video clip based on their attire

Jugal Gulwani (MT2016067)

Alok Nimrani (MT2016094)

Piyush Singh Bora (MT2016104)

Rahul Sharma (MT2016112)

Rajat Bansal (MT2016113)

Abstract

Classifying apparel or clothing is part of the wider task of classifying scenes. Such a system has many potential applications, ranging from automatic labeling to surveillance. Here we try to classify housekeeping staff, security and students based on their attire. We proceed in three stages – a) human detection where we try to compare detection results of two popular detectors HOG and Haar since accurate detection is foremost step in order to be able to classify people; b) feature extraction – another vital step where we are trying to determine the features best suited for clothing based classification; and c) classification where different machine learning methods such as KNN and SVM are examined on our database. Extensive experiments illustrate that the combination of LBP and SIFT features yields good separation amongst the data points.

Introduction

Classifying apparel or clothing is part of the wider task of classifying scenes. It is also related to detecting and describing persons in images or videos. The objective here is to classify people based on their attire. Such a system has many potential applications, ranging from automatic labeling in private or professional photo collections, over applications in e-commerce, or advertising and even surveillance to automate dress code validation at places which follow strict dress codes be it work places or social events/gatherings or even in cross border surveillance where we can label humans as authorized or un-authorized users

We were motivated by the scenario in which the house keeping staff, security personnel and students can be classified based on their attires since such a classification can be useful particularly during the curfew hours to determine human movement.

Earlier work on attire classification has been done for classifying scenes in a video. It is also related to detecting and describing persons in an image or video. Some work has also been done on segmentation of garments covering the upper body [1]. More recently Wang et al. [2] also investigated segmentation of upper bodies, where the individuals occlude each other. Retrieving similar clothes given a query image was addressed by Liu et al. [3] and Wang et al. [4]. In the latter work, the authors use attribute classifiers for re-ranking the search results. Song et al. [5] predicts people's occupation incorporating information on their clothing. Information extracted from clothing has also been used successfully to improve face recognition results [6]. Very recently, detection and classification of apparel has gained some momentum in the computer vision community. For instance, Yamaguchi et al. [7] show impressive results, relying strongly on state-of-the-art body pose estimation and super pixel segmentation.

In our work, we focus primarily on attire classification, we do not focus on clothing segmentation or similarity search, but on classification, i.e., the problem of describing what type of clothing is worn in an image. To do so, we build on top of existing work for clothing to then fully focus on the classification task.

Building Blocks: Three stages are involved in our classification system

- **Human detection**

  The popular human detectors available are HOG (Histogram of Oriented Gradients), DPM (Deformable Part Model) – a variant of HOG, Haar Cascade.

  HOG as described in [9] detects humans in a frame using a chain of image processing techniques and returns a bounding box around each human. It tries to detect the vertical edges which are form specific contours in case of humans and then, the obtained contours are classified using SVM. But it basically works fine with frontal views of humans as it requires particular contour shapes to work on and so side views or views with occluded arms, legs result in poor detection. Moreover, as it basically tries to find vertical edges, there can be several false positives being detected as humans if they have strong vertical edges. Haar Cascade is trained using a positive set of human images at different views and with all other background objects as part of negative set.

  In our experiments, we have tried to compare the detection accuracy of HOG and Haar Cascade for humans captured at different angles.

- **Feature Extraction**

  Extracting effective features from clothing regions is challenging. Clothing is much more complex and can have a large variety of styles in texture, shapes, and colors. There is little research on extracting clothing features where people have tried classification with extracting features such as SIFT, SURF, LBP and color space.

  Both SIFT and SURF contain detectors that find interest points in an image. The interest point detectors for SIFT and SURF work differently. However, the output is in both cases a representation of the neighborhood around an interest point as a descriptor vector. The descriptors can then be compared, or matched, to descriptors extracted from other images. SIFT uses a descriptor of lengths 64 and 128. Depending on the application, there are different matching strategies. A common method is to compute the nearest neighbor of a feature, and then check if the second closest neighbor is further away than some threshold value. It is found that the SIFT has detected more number of features compared to SURF but it is suffered with speed. The SURF is fast and has performance almost the same as SIFT.

  Color histogram involves color information of clothing and local binary pattern involves texture information. In some cases such as bright clothes and different classes having clear color distinction, color histogram approach tends to give good results. But a color histogram focuses only on the proportion of the number of different types of colors, regardless of the spatial location of the colors. The values of a color histogram are from statistics. They show the statistical distribution of colors and the essential tone of an image. Moreover, color based approach is heavily dependent on luminance and exposure. On the other hand, if a class of clothes is very rich in textures and designs then texture based features need to be extracted and LBP performs well in that case.

- **Classification**

The performance of various classification methods still depends greatly on the general characteristics of the data to be classified. The exact relationship between the data to be classified and the performance of various classification methods still remains to be discovered. Thus far, there has been no classification method that works best on any given problem. There have been various problems to the current classification methods we use today. To determine the best classification method for a certain dataset we still use trial and error to find the best performance.

In our case, we have used KNN and SVM to compare which one performs better. KNN and SVM represent different approaches to learning. Each approach implies different model for the underlying data. SVM assumes there exist a hyper-plane separating the data points (quite a restrictive assumption), while KNN attempts to approximate the underlying distribution of the data in a non-parametric fashion

KNN has some nice properties: it is automatically non-linear, it can detect linear or non-linear distributed data, and it tends to perform very well with a lot of data points. On the minus side, KNN needs to be carefully tuned, the choice of K and the metric (distance) to be used are critical. KNN is also very sensitive to bad features (attributes) so feature selection is also important. KNN is sensitive even to the outliers and removing them before using KNN tends to improve results.

SVM can be used in linear or non-linear ways with the use of a Kernel, when you have a limited set of points in many dimensions SVM tends to be very good because it should be able to find the linear separation that should exist. SVM is good with outliers as it will only use the most relevant points to find a linear separation (support vectors).

Attire Dataset

- ## Data collection

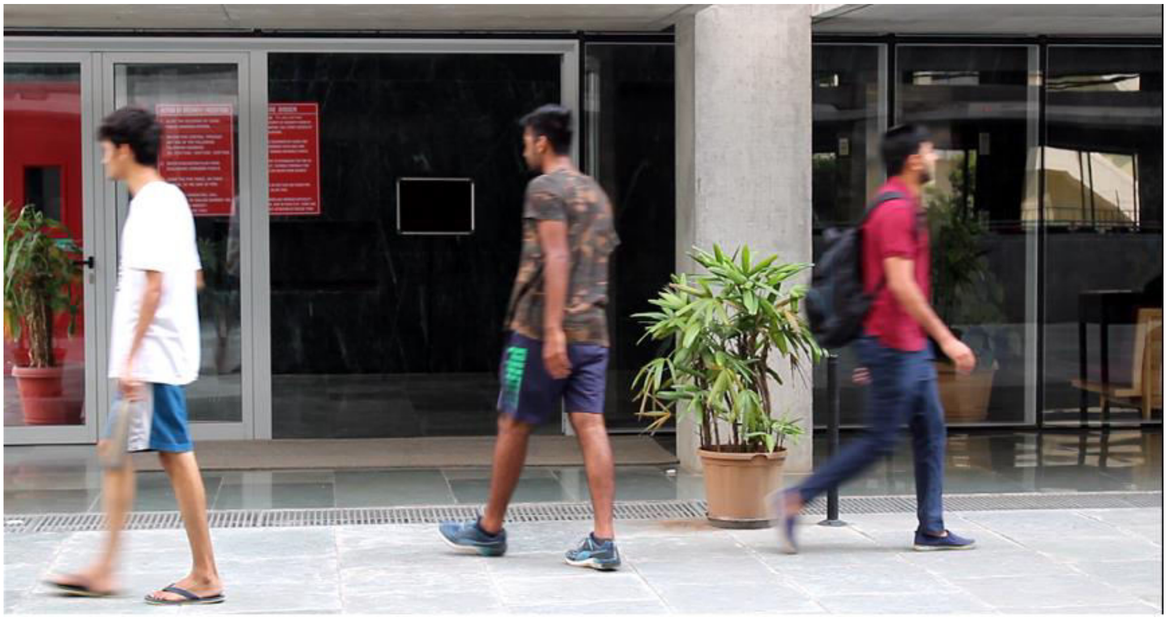
Apparatus used - Tripod, DSLR camera, mobile camera.

We collected video clips of students, housekeeping staff, and security guards in the vicinity of Bhaskara Building main entrance and near the mess area. The videos were collected from 2 angles -

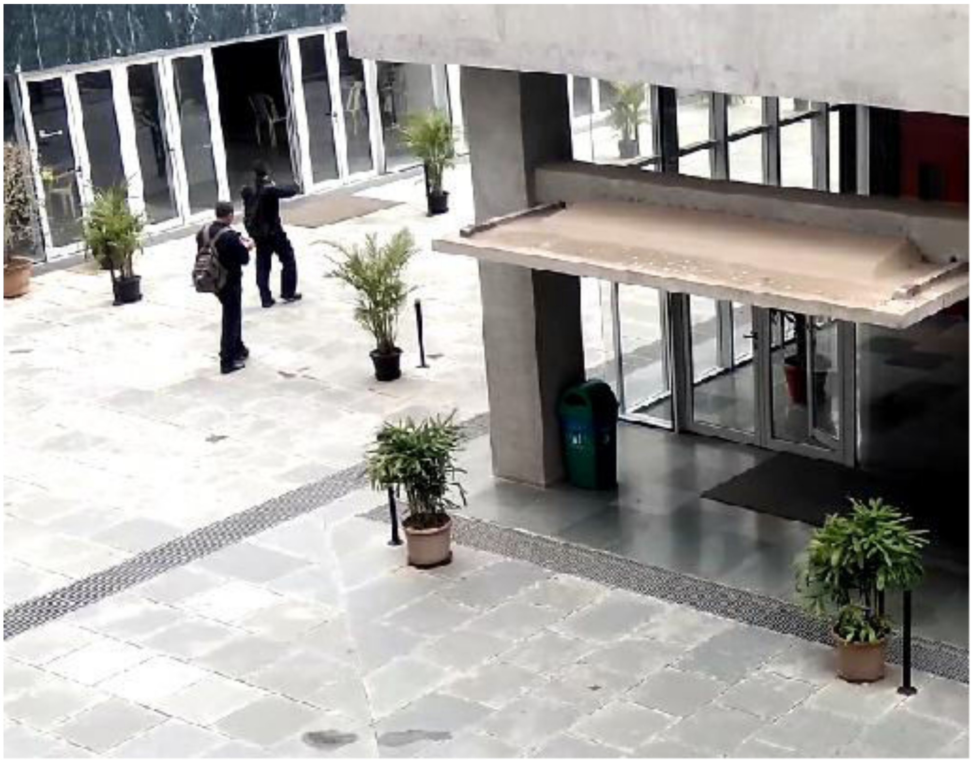
- Straight view i.e. 90 degree view of people
- Slanted view i.e. 45 degree view of people

In both of the cases, the video clips have captured the movements of the people from various angles namely their side, frontal, back views. The video also captures some amount of occlusion between people, however we have tried to minimize that as much as possible for better performance. The brightness in the videos varies and it depends on the time of the video and the position of camera.

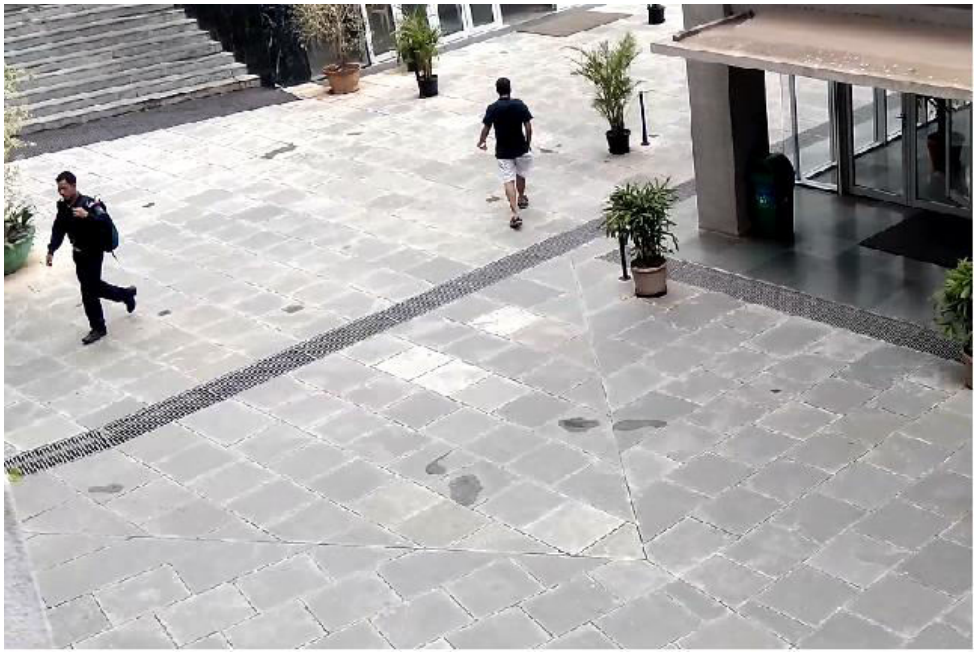
Below figures show sample frames from our dataset**:**



(Figure 1 - Straight view)



(Figure 2 - Slanted view)



(Figure 3 - Slanted view)

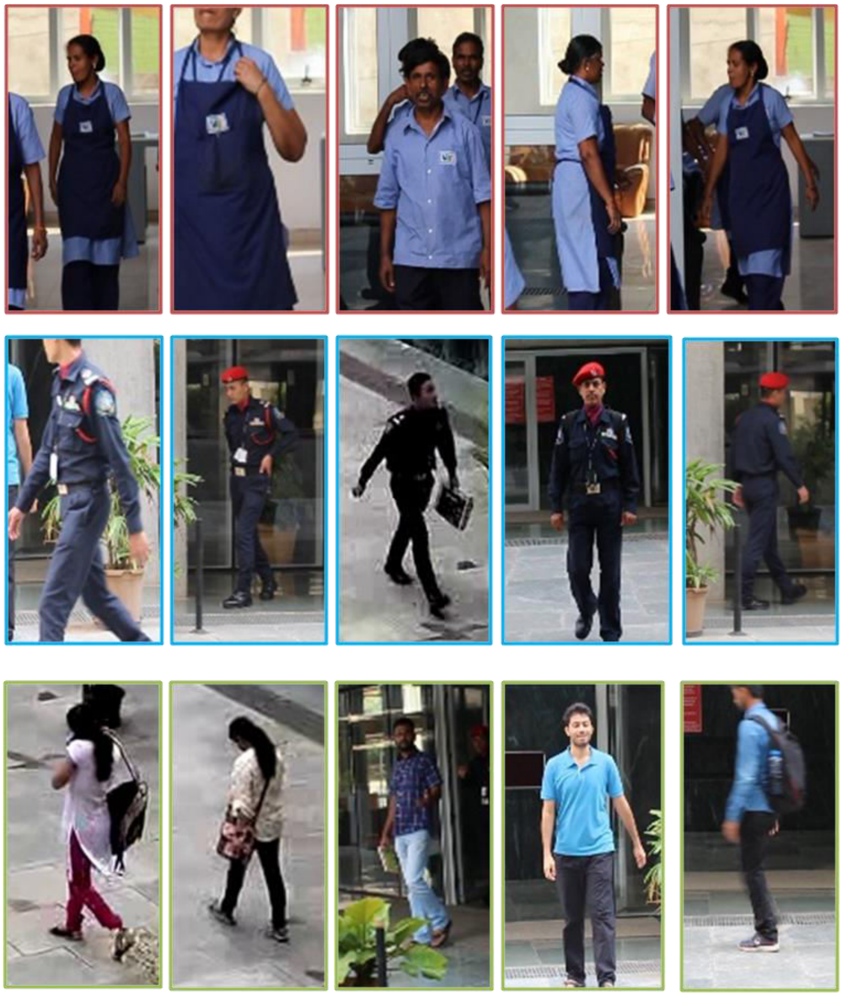
- **Preprocessing**

Since the videos collected were of high quality, we preprocessed the videos to reduce each frame to a maximum width of 600px.

After setting the frame size, we initially performed human detection in order to detect the humans and then we used non-maximum suppression technique to be able to crop out humans with proper, accurate bounding boxes. After the detected humans are cropped, we have resized each of them to the minimum size amongst all cropped objects.

We then manually labeled each cropped human as per its corresponding class and segmented the human data into three folders – housekeeping, security, student – as we are performing supervised classification

Figure 4 shows some sample cropped humans obtained after preprocessing:



(Figure 4 – Top: Housekeeping, Middle: Security, Bottom: Student)

Experimental Setup

For the experiments, we divided the collected video set to get separate videos for training and testing purpose.

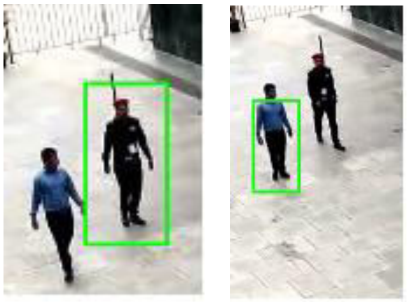
- **Training**
  - **Detecting humans**: To detect humans, we have applied popular human detectors – a) HOG with linear SVM and b) Haar Cascade. This training data of detected humans was generated as part of preprocessing and accordingly the humans were segmented into respective classes.
  - **Features**: For extracting the features from our modified dataset of cropped images, we have used several feature descriptors such as a) SIFT, b) SURF, c) LBP and d) Color space to get either the descriptors which are used to get a histogram representation of image features.
  - **Model**: As part of our model, we have used KNN and SVM classifiers. The histogram features of images along with the image labels are passed on to a KNN/SVM classifier.

- **Testing**
  - **Detecting humans**: For testing, we are using a video clip as our input. We then apply human detection algorithm (HOG or Haar Cascade) and crop the detected humans.
  - **Features**: Then we use feature descriptors to extract the features and apply K-Means to get the code book. Using the code book, we create a histogram of features for the cropped image.
  - **Classification**: Next to classify the detected humans we use our trained KNN/SVM models.

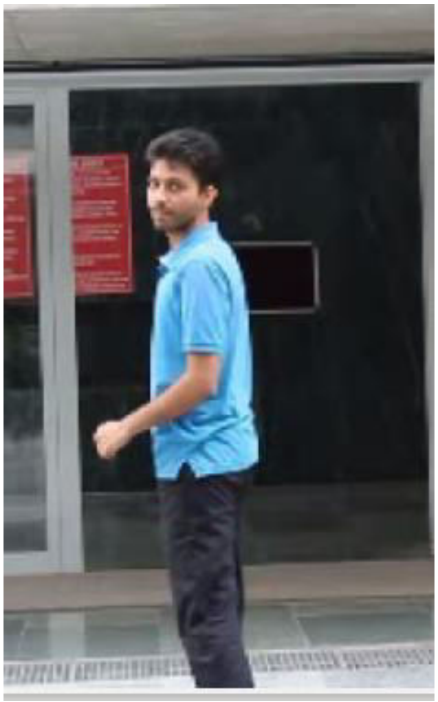
Attire based Classification - Results and Analysis

- **Human detection**

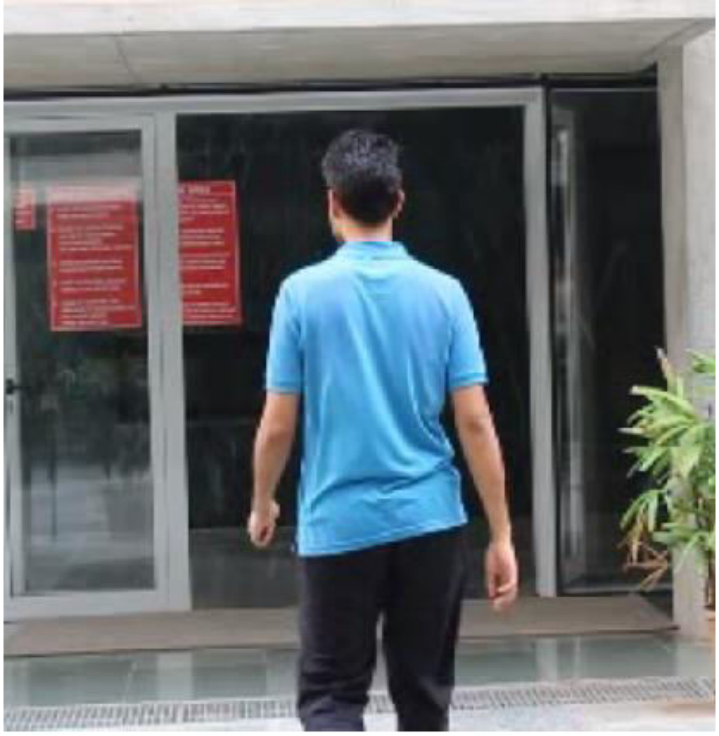
  We performed experiments with HOG and Haar on our videos captured at different angles. A general observation is that both methods work well when human features are very clearly visible or at least one arm, one leg; one side of shoulder is visible. When humans are captured with a straight view not all parts gets captured – generally the lower body is cropped out and in that case if only one side is visible then the detector fails to detect. Thus, when we performed detection on humans captured at a slanting – 45 degree angle, there was huge improvement in detection since in majority cases both arms and legs are visible.



(Figure 5 – humans detected in slant view)

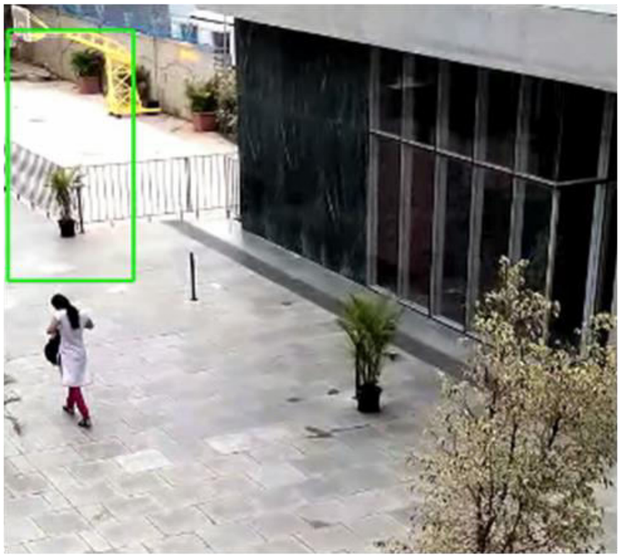


(Figure 6 - Not detected with one side visible in straight view)

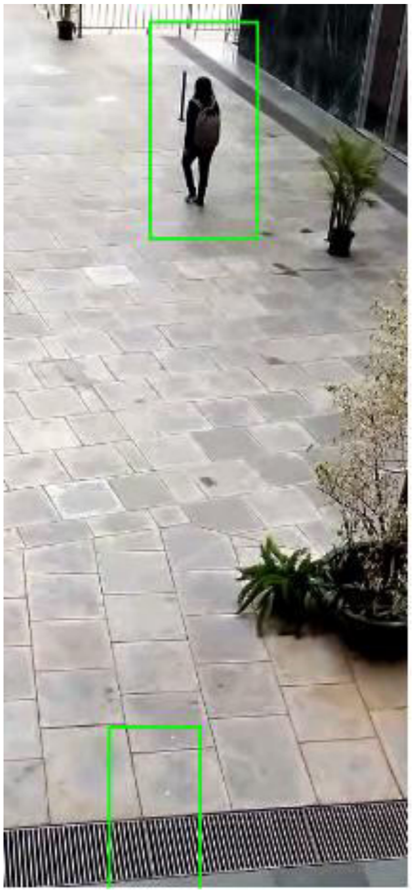


(Figure 7 - Not detected as the face/legs are not visible)

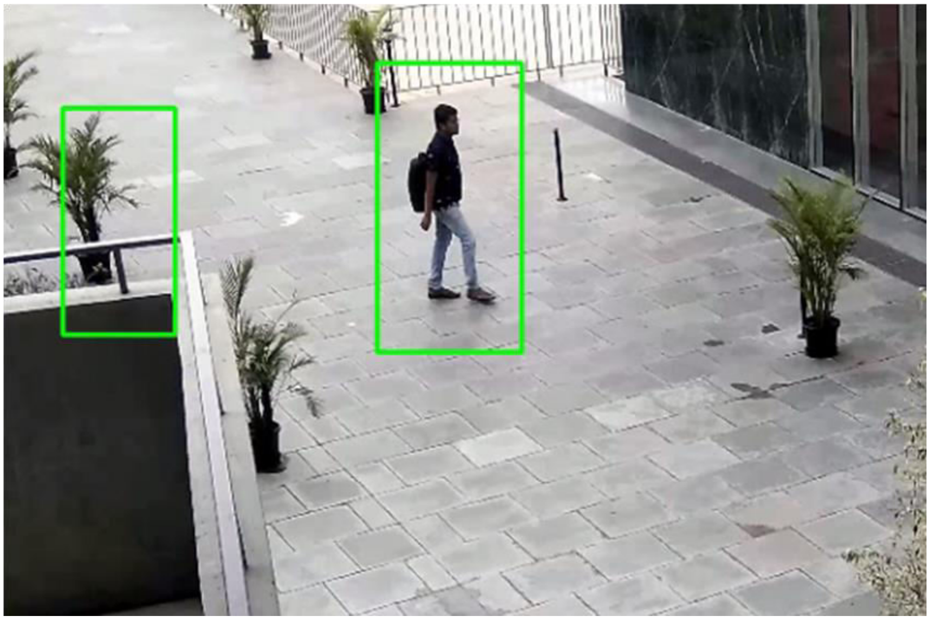
Other observation is about false positives – since HOG tries to determine the vertical edges, it generally produces more false positives compared to Haar cascade. Proper tuning of HOG parameters – window size and scale – is required to neglect the false positives.



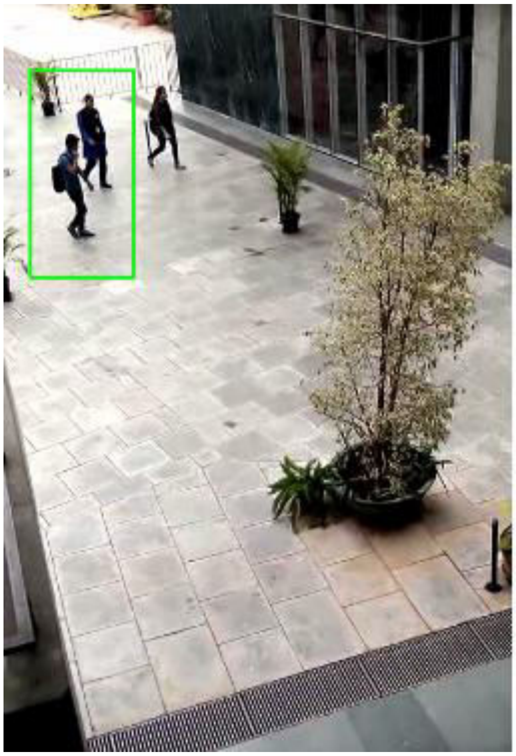
(Figure 8 – False positive with HOG: plant detected as human)



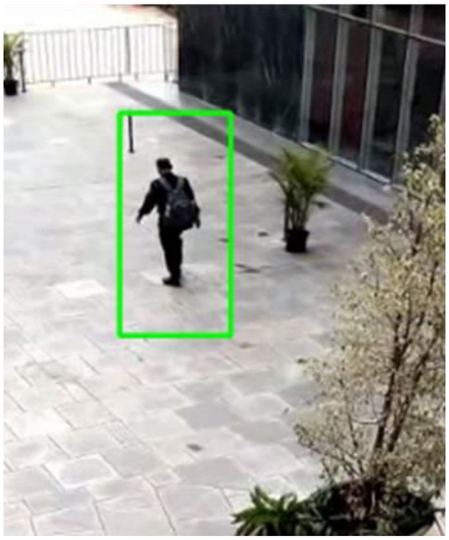
(Figure 9 - False Positive with HOG)



(Figure 10 – Plant detected with HOG)



(Figure 11 - No false positive with haar)

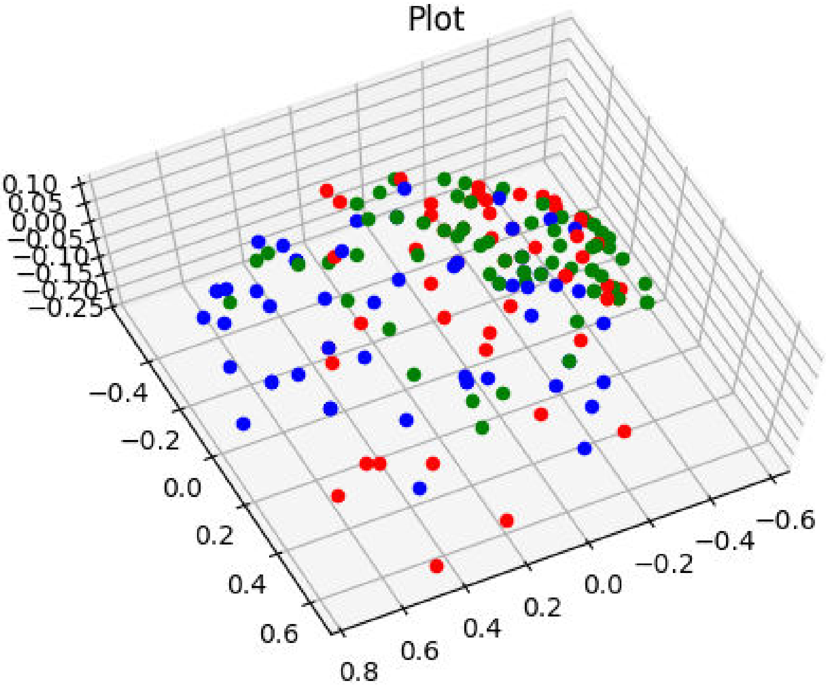


(Figure 12 – No plant detection as false positive)

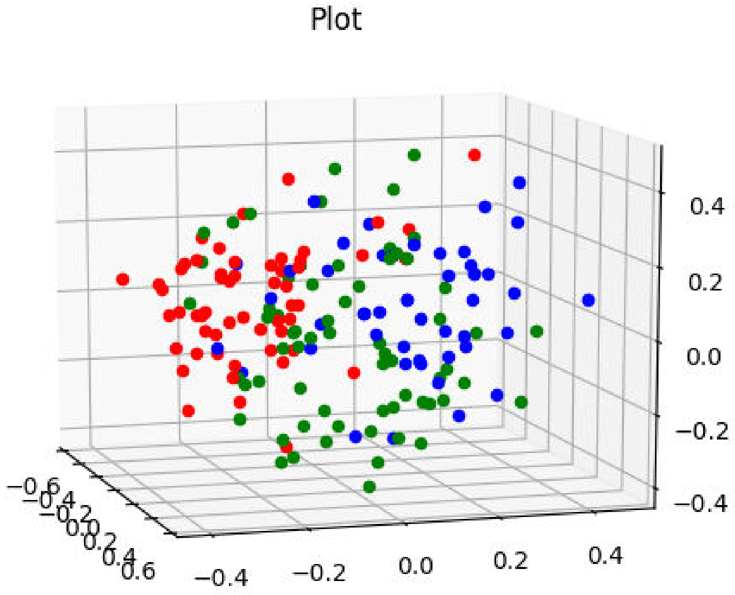
- **Feature Extraction**

  We performed experiments with SIFT, SURF, LBP and Color histogram. Since our videos were taken in the morning with less illumination, color histogram features are not turning out to be good features for classification.

To compare the feature descriptors, we reduced the histogram features to 3 dimensions using dimensionality reduction technique and plotted them to visualize their performance (it gives a rough view about data distribution as histograms are reduced from higher dimensions to only 3 dimensions). In between SIFT and SURF, SIFT is giving better separable data.

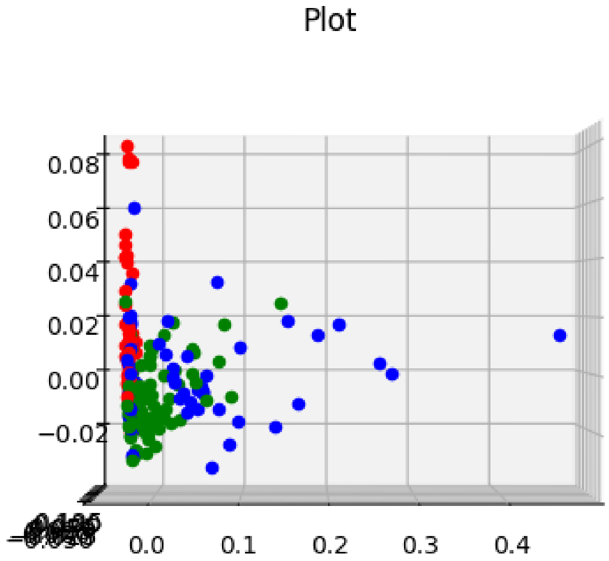


(Figure 13 - SURF Output: not well separated data)
(Red – Housekeeping, Blue – Security, Green – Student)



(Figure 14 - SIFT Output)
(Red – Housekeeping, Blue – Security, Green – Student)

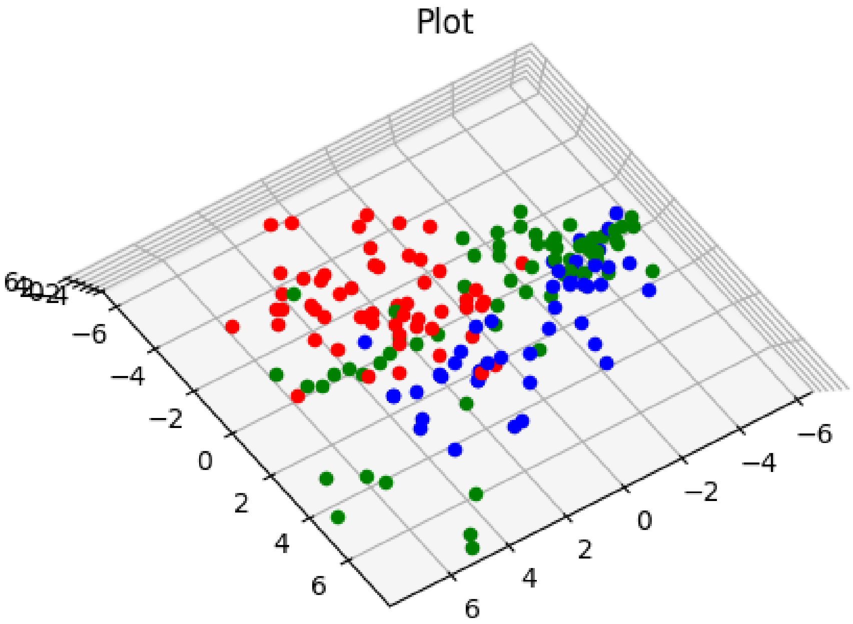
SIFT is able to separate security better but green student points are scattered all over. On the other hand, LBP is separating housekeeping staff in a better way.



(Figure 15 - LBP output)
(Red – Housekeeping, Blue – Security, Green – Student)

As can be seen in figure 15, red points lie majorly in the region above 0.01 y-axis and < -0.02 x-axis with only few blue and green dots having y>0.01 and x<-0.02.

And so we even tried with combining SIFT + LBP by concatenating corresponding histogram features which gave a much better separation amongst the classes as compared to that obtained from individual features

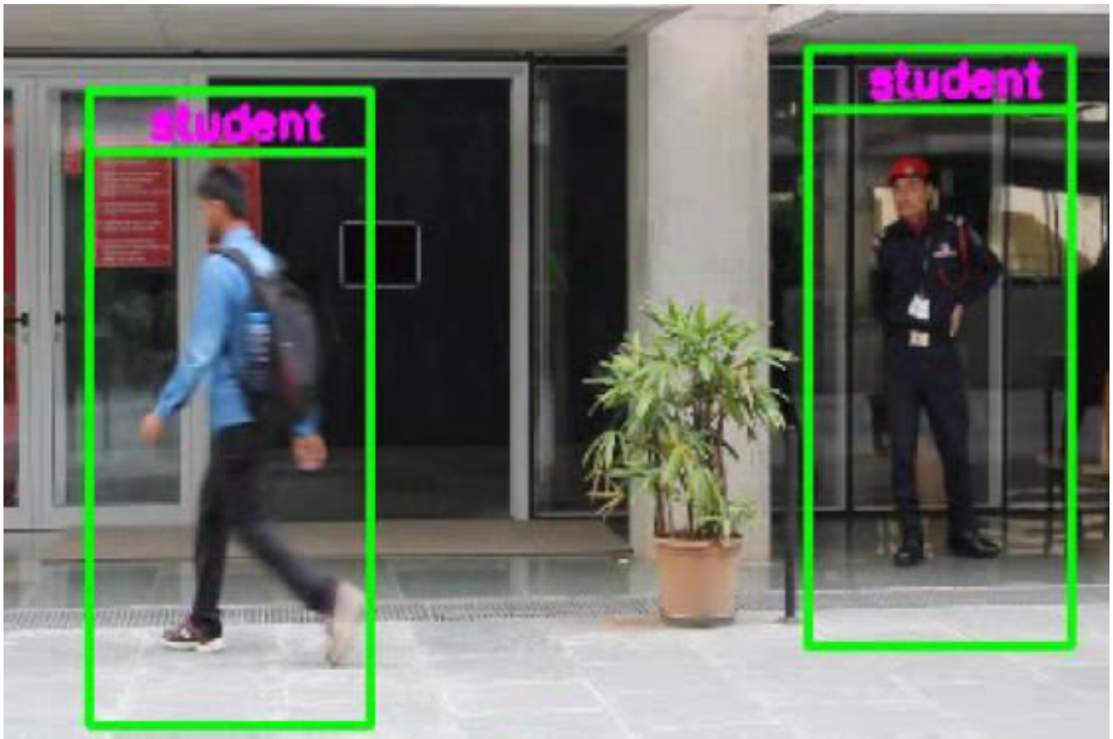


(Figure 16 – SIFT+LBP output)
(Red – Housekeeping, Blue – Security, Green – Students)

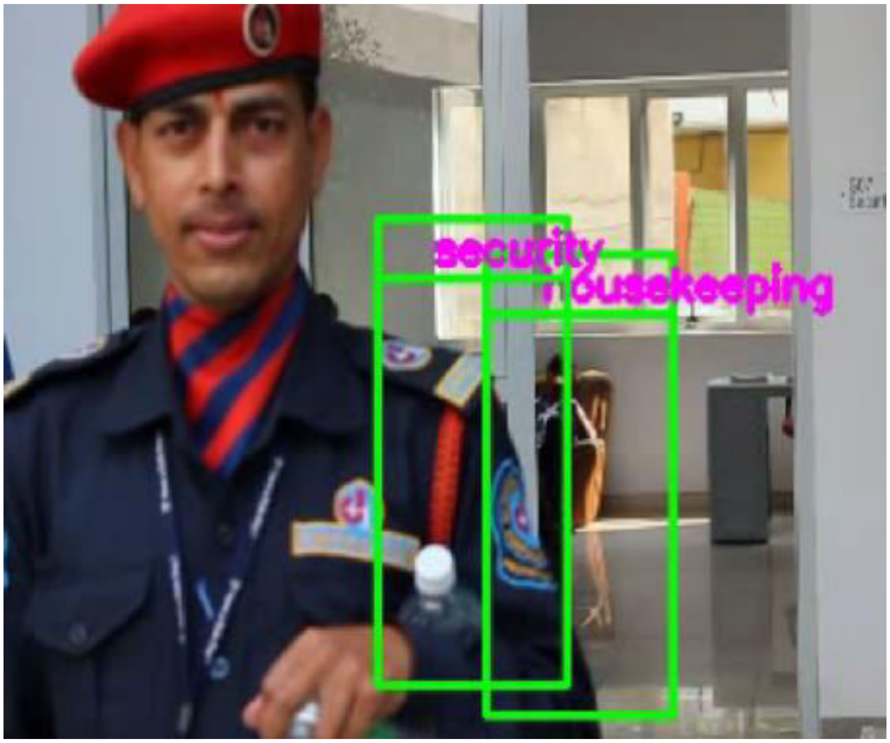
We can clearly see a separation between red and others while majority of blue and green points are also separable.

- **Classification**

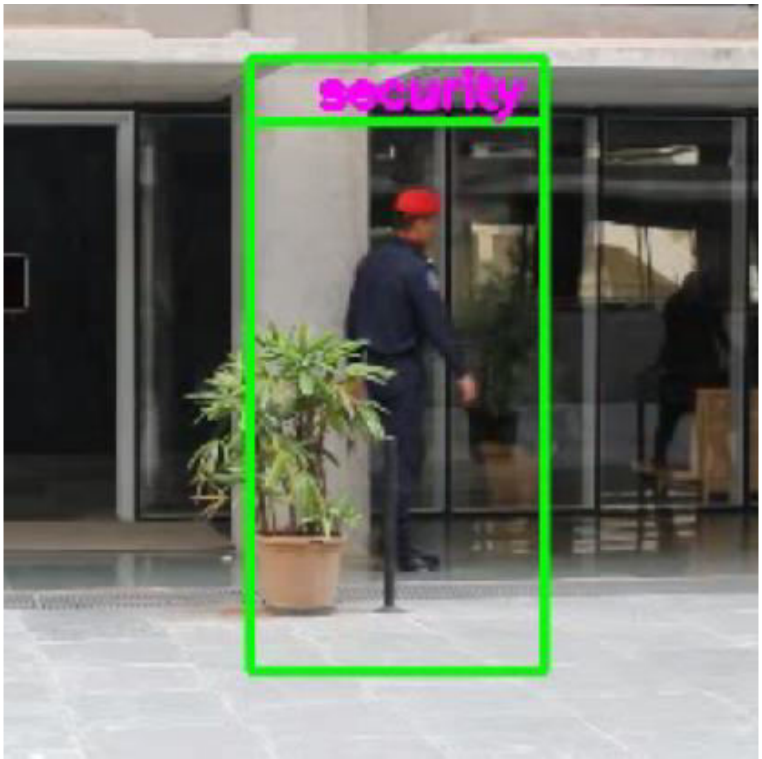
  For most of the cases, both KNN and SVM are giving almost the same performance. A general observation is that when there are multiple humans to classify in a frame, SVM performs better classification than KNN. Both KNN and SVM are performing well for only single human in a frame but with multiple humans, classification results are not so good.



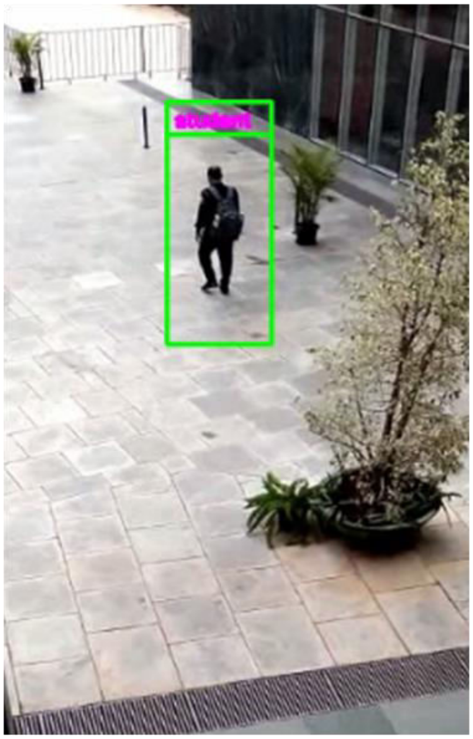
(Figure 17 - Misclassification for multiple humans with KNN)



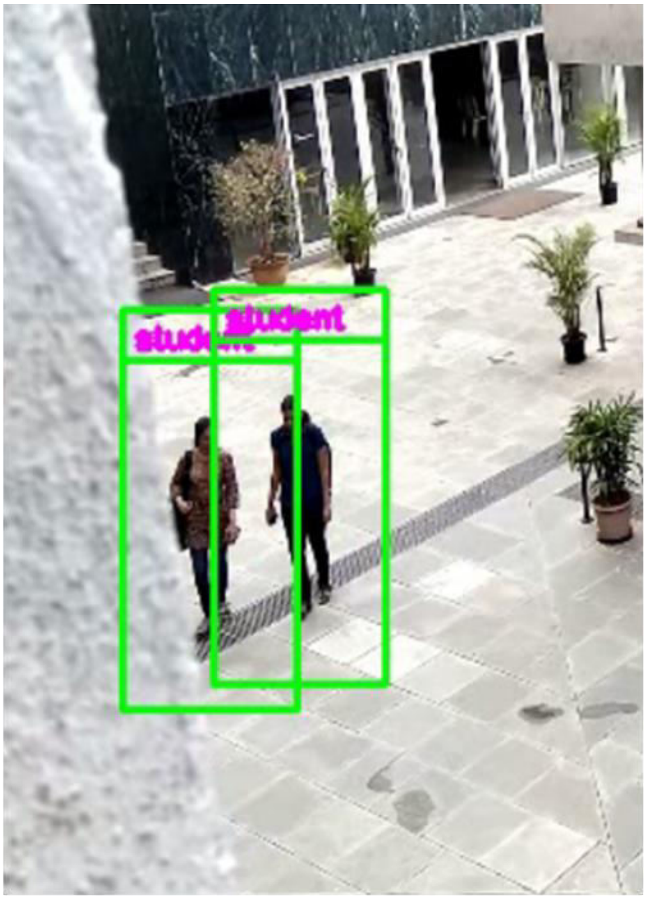
(Figure 18 - Proper classification for multiple humans)



(Figure 19 - Proper classification)



(Figure 20 – Misclassification of security as student due to presence of bag features)



(Figure 21 – Correct classification with SVM)

Conclusion & Future work

In summary,

- Haar cascade is giving less false positives and able to detect humans more accurately while HOG requires proper parameter tuning.
- Detection accuracy improves when humans are captured from a height
- LBP+SIFT feature combination is able to give a better separation amongst the classes.
- Both KNN and SVM have almost the same performance

Future Work – just as detection results improved with better dataset, we can try to compare the classification results when data is collected at different point of times in a day to determine how well the feature extraction can work.

Readings/References

[1] Hu, Z., Yan, H., Lin, X.: Clothing segmentation using foreground and background estimation based on the constrained Delaunay triangulation. Pattern Recognition 41 (2008)

[2] Wang, N., Ai, H.: Who Blocks Who: Simultaneous clothing segmentation for grouping images. ICCV (2011)

[3] Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., Yan, S.: Street-to-Shop: Cross-Scenario Clothing Retrieval via Parts Alignment and Auxiliary Set. CVPR (2012)

[4] Wang, X., Zhang, T.: Clothes search in consumer photos via color matching and attribute learning. MM, ACM Press (2011)

[5] Song, Z., Wang, M., Hua, X.s., Yan, S.: Predicting occupation via human clothing and contexts. ICCV (2011)

[6] Gallagher, A.C.: Clothing cosegmentation for recognizing people. CVPR (2008)

[7] Yamaguchi, K., Kiapour, H., Ortiz, L., Berg, T.L.: Parsing Clothing in Fashion Photographs. CVPR (2012)

[8] Bossard L., Dantone M., Leistner C., Wengert C., Quack T., Gool L.: Apparel classification with Style. ACCV (2012)

[9] Dalal N., Triggs B.: Histograms of Oriented Gradients for Human Detection. CVPR (2005)