Image-based people detection

Tommi Tikkanen
(78748P)

Valvoja: Arto Visala

AS-0.3100 Automaatio- ja systeemiteknikan seminaari
Syksy 2013
Abstract
People detection is a fast developing technology with various applications, including surveillance, crowd management and customer behaviour analysis. We present the short history of image-based people detection algorithms, starting from background subtraction and gradient histograms, ending up in the latest publications where color and depth information are combined. The main goal is to find out what kind of detectors have been implemented using depth cameras, such as Kinect, and what are the reasonable directions for further research.
# Table of Contents

Abstract..................................................................................................................................................2 
1 Introduction.......................................................................................................................................4 
2 Detection from color and intensity images..........................................................................................6 
   2.1 Background subtraction..................................................................................................................6 
   2.2 Pattern matching............................................................................................................................8 
   2.3 Counting people in dense crowds..............................................................................................10 
3 Detection from depth images and point clouds..................................................................................12 
   3.1 Early attempts................................................................................................................................12 
   3.2 Occupancy maps...........................................................................................................................12 
   3.3 Pattern matching............................................................................................................................14 
   3.4 Other notable methods..................................................................................................................16 
4 Summary...........................................................................................................................................21 
References............................................................................................................................................22
1 Introduction

People counting and tracking are technologies that are already widely being used on many fields of research and business. Applications include business intelligence in retail stores (ShopperTrak, 2013; VideoTurnstile, 2013; Leykin & Tuceryan, 2005), surveillance (Valera & Velastin, 2005), crowd management (Axiomatic, 2013), transport (DILAX, 2012), user interfaces (Z. Zhang, 2012) and intelligent environments (Krumm et al, 2000). Motivation for people tracking varies on each field. In retail stores, information about customer behaviour can be used to optimise shop layout to increase the sales, or to evaluate performance of an individual shop. In surveillance, for example activity in forbidden areas (Fuentes & Velastin, 2005), violent behaviour or vandalism (Zaidenberg et al, 2011) can be detected. In crowd management, crowd behaviour can be monitored and predicted to avoid incidents such as the 2010 Love Parade disaster in Germany, where 21 people died and more than 500 were injured (Helbing & Mukerji, 2012). Furthermore, information about crowds and people flow indoors can be utilized to improve energy efficiency by optimising air conditioning, heating and lighting, or to develop emergency evacuation strategies (Wang & Hu, 2013).

The large variety of application environments results in a large variety of different approaches for algorithms. For example, consider three important fields: people counting, surveillance and pedestrian detection for cars. Pedestrian detection is constrained to horizontal camera angles. In people counting, a vertical angle is often preferred to avoid occlusions (Sidla, 2006). Finally, in many surveillance applications it is necessary to both see the faces of persons and to avoid occlusions, so camera is tilted downwards. Crowds tend to look very different from each point of view, therefore it is likely that the best combination of algorithms is different depending on the camera angle.

For the best visibility and less missed detections, the majority of commercial people counting products are cameras that are placed in the ceiling, pointing downwards. However, this is not often the optimal set-up, if detection area needs to be maximized. This is the case especially if the room is not very tall. Also, re-identification from top-view is much more difficult. (Harville, 2002)

On the contrary, a big part of academic research in people detection concentrates on detecting humans in color images from the side. Identification based on e.g. color histogram and silhouette is easy, but occlusions are a problem.

Depth information has value when developing algorithms that are independent of camera angle. The performance of depth cameras, such as Microsoft Kinect, has been already notified in the people detection research. Depth information is not a novel innovation in the field of computer vision: earlier it could be obtained from stereo cameras, but they require more development effort and produce lower quality data. There were also other depth cameras before Microsoft Kinect, but generally with a price of thousands of euros.

During the past 10 years, there have been many breakthroughs in people detection research. First, Dalal and Triggs introduced Histogram of Oriented Gradients (HOG) in 2005, creating a basis for quick development of appearance-based detection. Between 2005 and 2012, many improvements and extensions of HOG were invented, one of the most notable being The Fastest Pedestrian
Detection in the West (FPDW) in 2010. Second, Primesense and Microsoft released the Kinect depth camera in 2010, making dense, high resolution depth imaging available at a low price. Third, the accuracy and computational cost of many decades old background subtraction has been improved by (Barnich & Van Droogenbroeck, 2011) with Visual Background Extractor (ViBe) in 2011.

We will introduce three relevant areas of people detection. First, different background models for background subtraction are presented. Background subtraction is widely used in surveillance and people counting applications. Second, pattern matching methods, popular for pedestrian detection, are reviewed. Although not as common in people counting, pattern matching methods are relevant, as they can be used to improve accuracy of people counting in combination with depth-based methods (Salas & Tomasi, 2011; Spinello & Arras, 2011). Finally, we take a look on the most relevant previous work: people detection from depth images. Emphasis is on recent studies that employ Kinect-style sensors, but also earlier research with stereo- and time-of-flight cameras is presented.
2 Detection from color and intensity images

2.1 Background subtraction

The majority of currently deployed people detection devices are based on extracting movement in the image by background subtraction. The reason for popularity is probably that it makes it fast to find objects of interest from the image (Teixeira et al, 2010). It is suitable especially when the background scene is either static or slowly changing.

However, there are major flaws in simple background subtraction (Teixeira & Savvides, 2007):

1. Natural oscillations in pixel intensity
2. Changes in lighting
3. Presence of repetitive background motion, such as waving tree leaves
4. Changes in position of static objects, such as furniture

Many improvements for background modelling have been suggested, most popular being Gaussian Mixture Model (GMM) (Stauffer & Grimson, 1999), also known as Mixture of Gaussians (MoG) (Moeslund et al, 2006). The downside of GMM is in the assumptions it makes: that the background is more frequently visible than the foreground, and that foreground's variance is significantly lower. (Barnich & Van Droogenbroeck, 2011)

Lately, a simple algorithm called Visual Background Extractor (ViBe) (Barnich & Van Droogenbroeck, 2011) has surpassed GMM's performance (Figure 1) both in terms of processing time and accuracy. ViBe has been further improved by (Van Droogenbroeck & Paquot, 2012), now called ViBe+.
Figure 1. ViBe surpasses the performance of GMM and various other background subtraction algorithms. Dataset: PETS2001. Image: Barnich and Van Droogenbroeck, 2011.
With depth cameras background subtraction is a much more robust technique. The distortion from changing lighting and shadows are eliminated (Han et al, 2013; Zhao & Thorpe, 2000). Also, noise from the Kinect depth sensor, PS1080, is relatively small, with standard deviation of less than 5 cm at 5 meter distance (Khoshelham & Elberink, 2012). In indoor environments, the changes in furniture are perhaps the only challenge that depth-based background subtraction has to solve. Background models such as GMM or ViBE typically update themselves over time, adapting to changes in background.

2.2 Pattern matching

Pattern matching is a very actively researched appearance-based technique for people detection. The advantage of appearance-based methods is that no background model is needed, so they are stable even in quickly evolving environments, such as on-board a driving car (Teixeira et al, 2010). In the field of pedestrian detection, all the top performing methods are based on pattern matching (Dollár et al, 2010). Matching is usually done in one or many feature spaces, due to distortions caused by pose and illumination changes in image space matching (Mori et al, 2005). Features are derived from the color image, for example using edge detection. Pattern matching utilizes machine learning: the object’s typical appearance taught to the classifier by giving a large image database as a reference (Teixeira et al, 2010).

Various features for pattern matching have been proposed during the past years (Table 1).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haar-wavelets</td>
<td>(Papageorgiou et al, 1998)</td>
</tr>
<tr>
<td>Haar-like features</td>
<td>(Viola &amp; Jones, 2001)</td>
</tr>
<tr>
<td>Global Chamfer distance</td>
<td>(Barrow et al, 1977)</td>
</tr>
<tr>
<td>Scale-Invariant Feature Transform (SIFT)</td>
<td>(Lowe, 1999; Lowe, 2004)</td>
</tr>
<tr>
<td>Histogram of Oriented Gradients (HOG)</td>
<td>(Dalal &amp; Triggs, 2005)</td>
</tr>
<tr>
<td>Shape Context</td>
<td>(Mori et al, 2005)</td>
</tr>
<tr>
<td>Local Chamfer Distance</td>
<td>(Mori et al, 2005)</td>
</tr>
<tr>
<td>Edgelets</td>
<td>(Wu &amp; Nevatia, 2007)</td>
</tr>
<tr>
<td>Shapelets</td>
<td>(Sabzmeydani &amp; Mori, 2007)</td>
</tr>
</tbody>
</table>

Table 1. Commonly features used in people detectors. Some of the features have been shown to be obsolete. Global chamfer distance is outperformed by local chamfer distance (Seemann, 2005). Also, Wojek & Schiele (2008) conclude that HOG and Shape Context perform better than Haar-wavelet and Haar-like features.
HOG and other versions of gradient histograms have quickly become the standard appearance-based people detector (Dollár et al, 2010). Therefore we will pay little attention to other features and concentrate on gradient histograms.

A histogram of gradients is built in the following way. First, an image is divided into a grid of cells. Then, a histogram of the orientations of luminance gradients is computed in each cell. The histograms are normalized and concatenated into a single vector for the whole image. Different sizes of detection windows are slid across the histogram image. A linear Support Vector Machine (SVM) classifies the resulting vectors into person or non-person. (Dalal & Triggs, 2005).

After its introduction in 2005, many improvements for HOG have been developed. Dalal et al (2006) extended HOG to include the use of motion, decreasing the amount of false positives dramatically. Schwartz et al further incorporated texture information. Zhu et al (2006) improved performance by an order of magnitude using a cascade of rejectors, while keeping accuracy almost similar.

While no single feature has been shown to outperform HOG, additional features can provide complementary information. Wojek and Schiele (2008) showed how a combination of Haar-like features, shapelets, shape context and HOG features outperforms any individual feature.

The accuracy of pattern matching have developed fast during the last decade. The Viola and Jones detector (2001) had approximately 10 false positives per image (FPPI) on the INRIA dataset (80 % detection rate). Dalal and Triggs decreased FPPI to 1, while the latest methods with more advanced learning methods and combinations of features have reached 0.1 FPPI on 80 % detection rate. (Dollár et al, 2010)

Although the detection results for the latest methods might seem impressive, there are still many limitations that need to be overcome (Dollár, 2012). First, computational cost of the most recent algorithms is notably high, processing less than 1 frame per second on a desktop PC (Dollár et al, 2010). The latest pedestrian detectors have partly solved this problem by selectively decreasing the number of iterations in sliding window matching. Such methods include the Fastest Pedestrian Detector in the West (FPDW) (Dollár et al, 2010) and (Benenson et al, 2012); these two run real-time on PC hardware, with similar or better recognition ability compared to their predecessors.

Second, HOG and other methods based on gradient histograms do not perform very well in case of occlusions (Dollár et al, 2012). Various ways to improve HOG in occluded scenes have been tried. Wang et al (2009) have combined HOG with Local Binary Patterns (LBP) to handle partial occlusions. Salas has improved detection results by combining HOG with depth data from Kinect, further discussed in section 3.4. Other solutions to the occlusion problem are discussed in the next chapter.

Dollár et al (2012) mention that low resolution, e.g. humans less than 50 pixels tall, is another serious challenge for even the best gradient histogram detectors. Besides solving problems with occlusions and small scales, they have listed five other areas of research that should be looked into, if these detectors want to be improved.
1. Motion features. The detector with the highest accuracy in Dollár's comparison (Walk et al, 2010) is the only one that takes advantage of motion. Dollár et al conclude that motion is a very effective method for human perception and thus very promising research direction. However, the method by Walk et al (2010) is computationally very complex.

2. Temporal two-way integration of tracker and detector. Studies show that the probabilistic prediction of human location improves detection results.

3. Ground-plane assumption. Again, knowing where to look from improves results.

4. Novel features. The best detectors use multiple other features in combination with gradient histograms. It is expected that new independent features result in additional gains.

5. Better datasets for testing. Commonly, INRIA dataset is used to evaluate gradient histogram detectors (Dollár et al, 2012). The problem is that it contains very few occlusions, and therefore it is too forgiving to algorithms that cannot cope with them.

2.3 Counting people in dense crowds
Since a major part of people detection research concentrates on pedestrians, the previously presented algorithms also favor these applications. It has to be kept in mind that pedestrian detection and people counting often have different priorities. In pedestrian detection, it is important to keep the frequency of missed detections very low, since a failure to detect a person can lead to injury or death (Chan, 2006). In people flow monitoring applications, accurate people count and maintaining track of an individual are more important, therefore occlusion handling is crucial. As camera is positioned higher to avoid occlusions, more people become visible in a dense crowd. However, only the upper part of the body is visible for many individuals, e.g. head and shoulders. While in sparse crowds, foreground segmentation combined with connected component analysis or point clustering may be enough to locate individual objects, in dense crowd the problem of overlapping people becomes very significant (Figure 2).

![Figure 2](image.png)

*Figure 2. An example of a dense group where it is hard to count or locate the individuals.*

In dense crowds, foreground segmentation is still useful, but needs to be combined with additional methods. Kilambi (2008) has shown one way to handle big groups of e.g. 10 people after
foreground extraction. The approach includes calculating the area of person's projection on the ground and after this, ellipse fitting on the group. Zhao et al (2008) have a similar method where they use a three-ellipse human shape model and additionally utilize color histograms to improve tracking. Zhao and Nevatia (2003) have combined four-ellipse human model segmentation with head detections from foreground and intensity image edges.

Rodriguez et al (2011) have applied crowd density estimation to minimize the error of people count in a scene. The principle is to minimize the difference between density by density estimation and density by people detection. In other words, detections are added to places where they are too few, and removed from places with too many of them. This method implements one of the most accurate object detectors available (Felzenszwalb et al, 2010), yet is able to significantly reduce false positives and increase true positives (Figure 3).

Figure 3. Improving head detections with density error minimization. On lower row green means true positives, red means false positives and yellow means new true positives received with crowd density estimate (top-right image).
3 Detection from depth images and point clouds

According to Harville (2002), depth data has great potential for improving people tracking performance for many reasons:

1. Depth is a powerful cue for foreground segmentation
2. Three-dimensional shape and metric size information improve foreground object classification, i.e. humans are better distinguished from other objects.
3. Occlusions can be detected and handled more explicitly.
4. New types of features for matching person descriptions across time become available, allowing better data association and re-identification.
5. A third dimension for prediction in tracking is provided.

Third dimension enables new algorithms for people detection, e.g. occupancy maps. On the other hand, big part of the depth-based methods are just variations of color or intensity image algorithms. HOG, point clustering, connected-component analysis and background subtraction have been re-implemented using depth images or point clouds.

Interest in people detection from depth images has increased with the development of processing and imaging hardware. Earliest depth-based people detectors were published in the 1990's, mostly working on stereo disparity images. In the late 2000's, novel time-of-flight (TOF) cameras gained popularity. Nowadays research is widely done with Kinect-style structured light sensors, but stereo- and TOF-cameras are still actively used in research.

3.1 Early attempts

The first real-time stereo vision systems were published approximately in 1996, one of them being (Kanade et al, 1996). At that time, most stereo systems took advantage of DSP (Digital Signal Processor) or FPGA (Field-Programmable Gate Array) hardware (Konolige, 1998), however, Eveland et al (1998) were able to run stereo computation and people tracking on a normal desktop PC (233 MHz Pentium II). Nowadays stereo systems are able to run on low-cost general purpose processors, e.g. ARM Cortex A9 (Point Grey, 2012).

Early depth-based people detectors, such as (Eveland et al, 1997) and (Darrell et al, 1998) and concentrate on extracting foreground from the depth image, while methods for finding occluded individuals in a group are rather simple. Darrell et al (1998) find gradients with a magnitude of more than 20 centimeters to separate different people in a connected foreground area.

3.2 Occupancy maps

Depth measurements provided by stereo cameras have a much greater noise compared to color values. Also, much of the depth image is unusable due to low visual texture. (Harville, 2002)

To cope with inaccuracies, Beymer (2000) has presented the idea of occupancy maps for people detection. Occupancy map divides the ground plane X-Y into a set of vertical bins. Each 3D point is then accumulated in one of these bins by its X-Y coordinates (Figure 4). Finally, humans can be
found in the occupancy map in various ways, e.g. matching a 2D Gaussian model to the map (Beymer, 2000).

However, Harville (2002) notes that shape information in the vertical dimension is lost when using an occupancy map. Furthermore, partly occluded persons may be left unnoticed. A height map, on the other hand, preserves object shapes in vertical dimension and handles occlusions better. However, height map alone would misinterpret small objects in the human head level. Harville (2002) has solved the disadvantages of both methods by using occupancy statistics to refine the height map.

Nedevschi (2009) and (Hordern & Kirchner, 2010) have both used the occupancy map to extract regions of interest (ROI) from the depth image. Nedevschi (2009) further filter the ROI by pattern matching side-view edge images. Finally, moving objects are counted as pedestrians if their motion signature, caused by moving limbs, is human-like. Hordern and Kirchner (2010) projected 3D data onto 2D planes and used Fourier descriptors to classify the shapes (Figure 5).

Figure 4. Principles of transferring a camera to a virtual “overhead” pose and occupancy map construction. Image: (Harville, 2002)
3.3 Pattern matching

Satake and Miura (2012) improved silhouette matching by overlapping silhouette templates, meaning that templates included multiple persons, possibly giving more accurate results in crowded scenes (Figure 6).
As noted in chapter 2.2, pattern matching such as HOG suffers from bad performance; various sizes of human feature templates have to be slid over the image, because distance at different parts of image is unknown. However, this is not the case with depth-based techniques, i.e. it is known what size of template needs to be applied at each point, decreasing iterations and making pattern matching a much lighter algorithm (Ikemura & Fujiyoshi, 2010; Salas & Tomasi, 2011; Spinello & Arras, 2011). Another advantage of depth images is that foreground can be extracted much more reliably. Therefore, background is removed before applying pattern matching in almost all the papers evaluated for this study that employed Kinect.

Lately implemented algorithms have been strongly influenced by HOG. First, Ikemura & Fujiyoshi (2010) presented Relational Depth Similarity Features (RDSF). It is a normalized depth value histogram of a small image patch (Figure 7). Authors claim RDSF is a better feature than HOG, however this has not been verified by other researchers.

**Figure 6.** Compared to single person templates, overlapping silhouette templates are better suited to analyze crowds. *Image: Satake and Miura (2012).*

**Figure 7:** Relational Depth Similarity Features describe the normalized depth value distribution of a certain cell in an image patch. *Image: Ikemura & Fujiyoshi (2010).*
One of the most popular depth image features has been Histogram of Depth Difference (HDD) (Wu et al, 2011), which is essentially similar to HOG, but with depth images. Simultaneously, Spinello and Arras (2011) came up with the same solution and called it Histogram of Oriented Depths (HOD). Furthermore, they presented Combo-HOD – a probabilistic method to combine results of HOG and HOD. Obviously, Combo-HOD is more accurate than either of the features independently. Also, if depth data is unavailable in case of sunlight or reflections, results from color image are still usable.

Usually, HOG is applied in camera angles where human silhouette can be clearly seen. (Tian et al, 2013) has shown that HOD is effective also from top view.

While pattern matching is still mainly done in image space, some experiments on 3D shape matching have been conducted. First, Bajracharya et al (2009) measured various 3D properties of the foreground objects found with a polar-perspective occupancy map. These features include properties of the total point cloud, such as variance, and point counts of certain pre-set volumes, similar to a low resolution 3D histogram. Xia et al (2012) have tried to find human heads by matching a 3D model of a hemisphere to a point cloud obtained by Kinect. The reason for selecting a hemisphere was view-invariance, as the human head will appear almost similar whether it is observed from the front, side, back or above. 3D model matching is used to remove false positives from 2D chamfer distance matching (Figure 8).

![Figure 8](image.jpg)

*Figure 8: Xia et al (2012) have used chamfer distance image with head-shoulder contour matching to locate potential heads. Furthermore, head candidates are filtered by matching a hemisphere model on their locations. Image: Xia et al (2012).*

### 3.4 Other notable methods

Almost always, when pattern matching is not used, foreground objects are extracted with background subtraction. In the evaluated 21 publications, only one did not subtract the whole background (Munaro et al, 2012). However, the floor was still detected and removed. Therefore, only two steps for people detection remain. First, it must be found which parts of the foreground correspond to which individual objects. Second, objects need to be classified as human or non-human. However, the second step is not even needed in many environments, where there are very few moving objects besides people.

One of the simplest way to separate individuals from the foreground is to see which pixels of a height thresholded foreground image are connected to each other, also called connected component analysis. Hernandez-Lopez et al (2011) have done this using an overhead Kinect, so that people are rarely overlapping in the image. Another approach, implemented for Kinect by Hsieh et al (2012), is
to obtain a point cloud by a tilted camera and then rotate it to an overhead point of view, similarly as presented in Figure 4 by Harville (2002).

The previously described method will obviously fail if people get too close to each other. For example limbs in the same planar coordinates erroneously connect two individuals as one detection, even if they are on whole different height. This is a strong reason to use connected components analysis in 3 dimensions if high quality 3D data is available. After using GMM occupancy grid background subtraction, Salas and Tomasi (2011) have labelled foreground points as blobs using 26-connectivity for 3D points. Even connected components analysis in 3 dimensions is not very robust in case people are actually touching. Therefore, Salas and Tomasi (2011) used the point cloud connectivity to create initial track candidates called tracklets, which are further refined using HOG for corresponding color images.

Seer et al (2012) use a more sophisticated method called complete-linkage clustering (Duda et al, 2001), which is more tolerant to weak links between individuals. This way, persons may be touching, e.g. holding hands, and still be recognized as two.

Fu et al (2012) have taken advantage of convex hull segmentation to handle dense groups of people (Figure 9). After finding best candidates with head-shoulder template matching, they fill the pedestrian in depth image in a way that adjacent points belong to the same segment if depth difference is between them is below a certain threshold. Finally, unintentionally merged areas are split if the depth map inside the convex polygon has a strongly concave feature, i.e. a low enough valley between two heads.

X. Zhang et al (2012) claim that pattern matching has two unsolved problems: multiple detections per person and missing detections by occlusion. As the head is normally the uppermost part of the human body, finding local maxima in a height map will give the positions of the heads. However, this method is not very robust if noise is present or if adjacent humans have very different heights.
For example if a child is next to an adult, the adult's shoulder is often taller than the child's head. X. Zhang et al (2012) solve these problems with a robust method called water filling. An inverse height map is used as a surface where raindrops are randomly dropped on. Upon reaching the surface, the raindrop will move towards a lower point if one is found in the nearby environment. This is repeated until the raindrop has cannot move any lower and amount of water in the resting position is incremented. After dropping a large amount of water, humans are found by locating the bodies of water that exceed a certain deepness threshold (Figure 10). The cleverness of water filling is that local maxima are not found using a constant size search window, but adapting to find the relevant maxima regardless if they are the actual maxima of a small or a larger area. X. Zhang et al (2012) have implemented the algorithm for an overhead camera, which is probably the optimal set-up. However, it is unclear if the same approach could be used for tilted angles to reach a wider scene.

Additionally, to emphasize the most relevant research, publications where Kinect or similar sensor was used, are more listed with more details in Table 2. As a conclusion detection results are impressive, but there are some limitations. Silhouette matching methods are heavily dependant on the viewing angle and inevitably suffer from occlusions. Gradient histograms are able to function from any camera position, but require wide learning material from multiple angles. Overhead setups may cover a smaller area than side view or tilted cameras, if the ceiling is low. It is unclear how well clustering or water filling works for other camera angles. Perhaps the most significant flaw in the evaluated publications is that practical usability is not considered enough: all detection systems run on PC hardware, which is power-consuming and also increases the cost of processing hardware. On a real use case where cameras need to be deployed around a building, processing is rather performed on the camera node, to avoid network load by not sending image data through network.
Table 2. Previous applications of Kinect-like depth cameras in people detection and tracking

Finally, to give a good overview about previous research on depth based people detection, we present a graph that describes the combinations of algorithms used in all the previously mentioned publications (Figure 11).
Figure 11: Selection of algorithms for people detection with depth sensors. Colors of the boxes correspond to the type of sensor used: stereo camera as gray, TOF camera as black and Kinect as white. Names of publications are shortened by taking the first 3 letters of the first author and the last two numbers of the publication year, except Muñoz Salinas et al, who are shortened MS05 to avoid confusion with (Munaro et al, 2012).

* For a clearer visualization, not all algorithms are shown. [Sal12] also employ connected component analysis, [Xia12] use local chamfer distance and [Zha12] water filling.
4 Summary

Regardless of the type of sensor, pattern matching is an accurate and popular method to find people in images. Gradient histograms, such as HOG and HOD, have been a major direction for the research. Other recently used features include head-shoulder contour matching, local chamfer distance and Histogram of Depth Difference. Pattern matching has mainly two disadvantages: it is computationally intensive and it does not perform very well in crowded scenes. Both of these challenges have been addressed. First, computational cost has been lowered by intelligently decreasing the number of comparison operations. Second, accuracy in a crowded scene has been improved by estimating crowd density in each pixel, then fusing the results of people detector and density calculator.

Another, lighter approach is to first extract foreground pixels, i.e. define which pixels in an image contain humans. After this, groups of multiple people need to be analyzed to count and locate individuals.

Taking advantage of highly accurate depth sensors makes foreground extraction an easy task. If the camera is stationary, background subtraction will give most of the foreground pixels with practically no false positives. If the camera is moving, or background scene changing fast, occupancy maps will give a good estimate on the locations of tall objects such as humans.

Group segmentation, on the other hand, is an unsolved problem. In simple cases, where all individuals form their own uniform area in the foreground, connected components analysis is satisfactory. However, when an area in the foreground consists of multiple persons, more advanced methods are required. Depth images enable much better possibilities to analyze groups, compared to color images. Examples of well-performing group analysis methods include 3D point clustering, depth image segmentation and water filling. It is still unclear which would be the best set of algorithms in different situations, as they have not been compared using a common dataset.
References


