Paper Title*

*

1st Su,ZhaoLin
dept. National Tsing Hua University,NTHU (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

An increasing number of convenience stores as well as medium-scale and small-scale supermarkets appears in our surroundings. To make a long-term profit, retailers desire to keep their overall cost as low as possible in terms of management, labor, logistics and properties. According to the latest news, The Home Depot yearly labor cost 90,444 million, Costco yearly logistics cost 10,444 million. Labor costs account for a large part of the operating costs of shopping malls. So in recent years, many shopping malls have sought to transform and reduce labor costs with the same profits. In recent year, there are many stores that require little or no manual labor, many stores have changed their business model.. They all use computer identification technology to reduce labor costs, and it has more advantages than traditional shopping malls. They have quick checkout and self checkout technology, make people's shopping faster, make the management of operators more convenient.

Meanwhile, a concept of "cashier-free supermarket" is proposed in nowadays, which draws much public attention. This technology is undoubtedly the best way to solve the labor cost at present, and over time it will gradually replace the traditional business model. The number of unmanned supermarkets is increasing, and it is becoming an emerging business model in the future, for an instance, the novel idea of Take it awayas announced by Amazon Go. Amazon Go, Sam's Club Now and Taocafe are three pioneers that jump into the "cashier-free supermarket" technology and make it come true. Amazon Go already started 5 unmanned stores in U.S. and they are planning to open up to 1000 stores till 2020. Sensors and camera scan detect the categories and quantities of the products that a customer takes, it can make a preliminary settlement the self operation problems. Moreover, with a proper camera layout, more integrated information can be gathered to strengthen our estimation of the quantitative variation of the products [1]. Camera monitoring the purchasing behavior is an important part in a "cashier-free supermarket", and the change from a traditional supermarket to an advanced supermarket can be achieved by upgrading the capabilities and layouts of cameras, to provide an identification feature. This paper will discuss the techniques of the modern cameras, mainly about device layouts, items category identification and quantity counting. And achieve the establishment of Recognition System, the technology used and the problems solved.

The rest of this paper is structured as follows:
Section II xxxx
Section II xxxx
xxxx in Section II.

Identify applicable funding agency here. If none, delete this.

II. Related Works

Since the concept of "cashier-free supermarket" was put forward, it has attracted many people's attention. Not only to save costs, but also to show you another possibility of future shopping malls. In these markets seeking transformation, Amazon Go and Sam's Club Now and Taocafe are popular retailers in unmanned supermarket. All of them use the more advanced identification technology, plus their layout and the choice of camera equipment, realize the idea of "cashier-free supermarket". For one of them, Amazon Go stores use a combination of sensors on the shelves, cameras and computer vision with machine learning. Different from Amazon Go, Sam's Club pays more attention to camera layout aiming to plan shopping routes, and it use RFID [2]. In our experiment, we learnt from both retailers, and we set camera layout first, then use computer vision and sensors to identity the category and quantity of the items.

A. Multi Cameras System

This is the foundation of "cashier-free supermarket". It includes the choice of cameras, the number and placement of cameras in the scene, the angle and depth of field needed for visits. In many shopping malls, the most commonly used camera layout is to place the camera on the top of the scene and on the shelves. The rationality and feasibility of this layout are fully proved by [3]. In order to lower the overall cost of the sellers and economize resources, we aim to use the less cameras to cover whole scene while the clearest image can be obtained. Compare the camera layout of various supermarkets and chose one of the best-functioned layouts of cameras for our experiment, which use multi cameras and each camera is responsible for a specific area. There are two type cameras used in scene. One is one-directional camera placed on the shelf, which requires high resolution in order to obtain high-definition pictures of items for identification. Another is four-directional camera on the top of the scene, it is used to assist the camera on the shelf to recognize and observe the whole scene. Next, we compare the cameras that are commonly used in the common market. **BOSCH** FLEXIDOME IP panoramic 7000 MP, the camera is widely used in many shopping malls. Compared with other cameras, it has the advantages of easy installation, wide viewing angle and fast transmission speed. **AXIS** P3717-PLE is 4-directional cameras, it has reduced bandwidth and storage needs, 360 IR illumination and remote zoom and focus and flexible positioning of four varifocal camera heads. We can get all the information in this area and image obtained for each region are processed separately []. We try to adjust the Field of View (FOV) aiming to ensure both, camera coverage maximizing and image clarity [1].

B. Computer vision

At present, there are many kinds of architectures for computer vision, their slave processing methods and speed are different. Two popular computer visions are You Only Look Once (YOLO) and Regions with CNN features (RCNN) [4] [5]. Both of them have strong recognition ability. Faster-RCNN is a technology for using CNN to research feature map, and RPN network will complete the full operation of the map before send it to box regression layer and taxonomy [4]. It has the characteristics of fast and accurate identification of items. YOLO just like its name "You Only Look Once", it has faster identification, which has developed to the third generation. It input S*S grid, and every gird is responsible for the objects falling into it, then it can choose the maximal IOU of bounding box [5]. Their comparison is shown in the Fig

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [5]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [8]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [6]	Inception-ResNet-v2 [21]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [20]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [15]	DarkNet-19 [15]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [11, 3]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [3]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [9]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [9]	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608×608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

Fig. 1. Page Size and Optical Resolution

RCNN is often used in scenarios requiring more accurate recognition, a method for face detection [6] [7] [8], to towards Real-Time object detection [4] [9]. YOLO can also be used for accurate identification, it is based on the global information of the image to predict, and the error rate is low. It can be used to interactive environments [10], modeling behavior from visual data [11], anticipating visual representations from unlabeled video [12]. YOLO is more suitable than Faster-RCNN in our system by comparison. YOLOv3 is chosen in our experiment since in processing medium-scale and small-scale supermarkets, it is based on Open Source Computer Vision Library(OpenCV) and Compute Unified Device Architecture(CUDA). In the scene, it will put the images captured by the camera into YOLO for category judgment, and then output the results.

C. Weight sensor

Amazon Go use the weight sensors as an assistant to determine the number of items on the shelves.

III. Technical comparison

A. official tones

Amazon Go officials stressed that they could use computer vision, sensor fusion, and deep learning to make people shopping more convenient. The store concept is considered as a revolutionary model which relies on the prevalence of smartphones and geofencing technology to streamline the customer experience, as well as supply chain and inventory management [**?**] . Electronic shelf label (ESL)

B. end-user comments

According to some end-user comments, the existence of trans-era science and technology changes our consumption pattern. However, some people think it makes shopping a hassle since Amazon Go omitted the need for closing account by human, and it is hard for the aged to use smartphone.

ESL: electronic shelf label small fontsize for elderly people

60yrs age grandma. samsclub , membership fee. self check-out on smartphones. reduce cost, incurring burden onto customers.

younger adults, students no line, quick, smartphone

proofread

IV. Method

Now that the cashier-free supermarket has become a trend in the future development, and more and more businesses will join the ranks in the futures. So in order to conform to the trend of development, we did a series of experiments to achieve "cashier-free supermarket", and we will try to apply it to reality. It include make scene to simulate a real supermarket and use multi cameras to cover whole scene to get images which used to identity items types and number on shelves. Then we combine them into a complete system. We made this model through Unity3D, which is a cross-platform game engine developed, then we can use it to build our scene.

A. Multi-cameras and Build Scene

We will use the identification in daily supermarkets in our life, so we investigated any small and medium-sized supermarkets, such as FamilyMart, JOETEN, 7-ELEVEN. We find that the general pattern of these supermarkets is used strip shelves and parallel emissions. And then we draw lessons from their pattern, we use the Unity3D to achieve the scene. We used multi cameras in our simulated scenario, and try to restore the real scene as much as possible. The Multi cameras scene is based on field of view(FOV), the horizontal field of view(H-FOV) and vertical field of view(V-FOV) [13] determined the [sharpness] and size of the pictures [what we get].

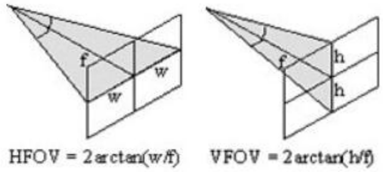

Fig. 2. H-FOV and V-FOV .

Fixed focal length lens, it can get the angel field of view(AFOV), and we can get different size of FOV through adjusting the focal length of the lens through different working distances.

$$AFOV\,(^\circ) = \ 2\,x\,tan^{-1}\left(\frac{h}{2f}\right)$$

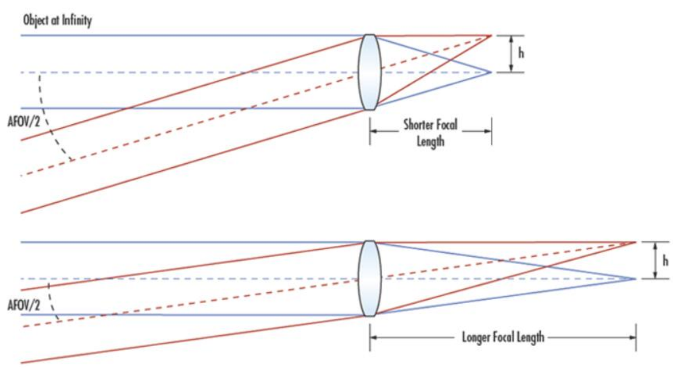

Fig. 3. The field of view of lens is related to focal length, f is the focal length, h is the horizontal dimension of the sensor.

We can follow the specifications of the scene (shelves size, the length and width of walking, and layout of the whole scene) to adjust FOV.

B. Cameras Selected

The terms ”Dots Per Inch” (DPI) and ”Pixels Per Inch” (PPI) are used interchangeably by many. [A 200 DPI print means that for each inch of that printed material, it takes about 200 dots to make the picture.]

In imaging, PPI (pixels per inch) refers to the input resolution of a photograph or image. For example, a 200 PPI technically stands for that 200 dots per inch display in a image.

A pixel is [like] a square dot without gaps. Both of them can describe the quality of a picture, and some cameras save digital images in arbitrary values as 72 dpi. We can calculate the DPI or PPI for what we need page size in Fig.3.

[why u mention this?] FLEXIDOME IP panoramic 7000 MP, which has 12MP / 30 fps sensor for fine details with smooth motion and Intelligent Video Analysis on full panoramic overview. Panoramic surveillance offers full 180 or 360 coverage of the designated area. The 360 version of the camera, when mounted centrally on a ceiling, gives complete wall-to-wall coverage. The 180 version has a higher effective

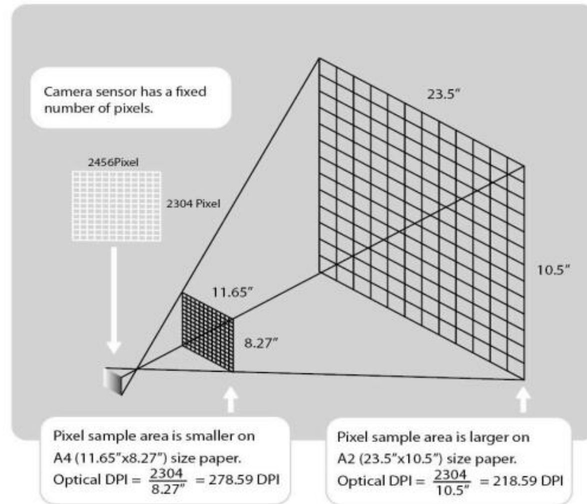

Fig. 4. Page Size and Optical Resolution

resolution and is ideal for wall mounting or for ceiling mounting in corridors.

When mounted at a height of 3.5 m (11.48 ft) the 360° version of the camera has the following coverage radius for the four levels in Fig.4.

DORI	DORI definition	Coverage radius
Detect	25 px/m (8 px/ft)	19 m (62 ft)
Observe	63 px/m (19 px/ft)	8.5 m (28 ft)
Recognize	125 px/m (38 px/ft)	4.5 m (15 ft)
Identify	250 px/m (76 px/ft)	2 m (7 ft)

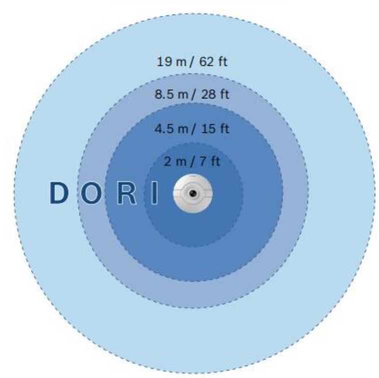

Fig. 5. Page Size and Optical Resolution

When mounted at a height of 3.5 m (11.48 ft) the 180 version of the camera has the following coverage radius for the four levels in Fig.5.

The camera provides the full resolution circular image for recording even if we are viewing only a portion of the scene. Unity3D is a game engine, the engine can be used to create both three-dimensional and two-dimensional games as well as simulations for its many platforms. In our experiment, we use

DORI	DORI definition	Coverage radius
Detect	25 px/m (8 px/ft)	28 m (92 ft)
Observe	63 px/m (19 px/ft)	12 m (40 ft)
Recognize	125 px/m (38 px/ft)	6.5 m (22 ft)
Identify	250 px/m (76 px/ft)	3.5 m (12 ft)

Fig. 6. Page Size and Optical Resolution

Unity3D to build a scene like a truth supermarket, and camera placement experiments are performed in this simulated scene.

C. Item Recognition

The main method current approaches to object recognition make essential use of machine learning methods and deep learning methods. To achieve the item recognition, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. In order to achieve item recognition, we use the YOLOv3 method to train our datasets. YOLO means "You only look once, real time object detection explained", it is a new item recognition method, now it has developed to the 3rd generation, it has more accurate recognition rate and ability to process larger amounts of data. We will use the YOLOv3 to train image what we get, it is based on OpenCV 3.40 and CUDA 8.0.

A single GTX 1060 GPU has 6GB of memory, it is enough for us to train examples, therefore we spread the net across GPU. We can use the Python Reptile to get image from Internet, then use visual GUI-software for marking bounded boxes of objects and generating annotation files. It will create text-file for each jpeg-image-file in the same directory and with the same name, but with .txt-extension, and put to file: object number and object coordinates on this image, for each object in new line: object-class, x, y, width, height.

D. Sensor Fusion

V. Camera Displacement

camera category camera type: resolution specifications. camera design case: walmart,cvs, 7-11 item, customer cashier [recommendation] auto-cashier footage

VI. Simulator

In the previous section, it introduces how we design scenarios and how to arrange them. In this section, we're going to build up the scenarios described earlier. It is included achieve the cashier-free supermarket in Unity3D, and the item recognition method what we use, then we combine them into a whole and application of simulation in real scene.

A. Scene simulation

Achieving our scene in Unity3D which can simulate scene like a truth supermarket, the model is basic on the layout of Supermarket after our investigation. It is length and width and height of the scene are 40m, 25m, 6m, and aisle width is 2.7m. In the scene we set 5 bins, and each of bin width is 2.35m and height is 1.7m. When have built the scene, we can design add the multi cameras to our scene, then put the cameras on the bins. In the scene we put the each camera on the bins which height is 2.22m, it is through our calculation, the most conducive to obtaining the height of the image, and set cameras 6.8m apart from the previous calculations. Then build four-direction cameras on the aisle, which is placed every 3 meters above the corridor, the cameras coverage needs to cover the entire scene. A four-direction camera has 4 cameras for 4 directions, and we put them on the 3.5m. height of aisle. In the scene we built, we totaly use 152 cameras in this scene. The overall layout is shown in the Fig.6.

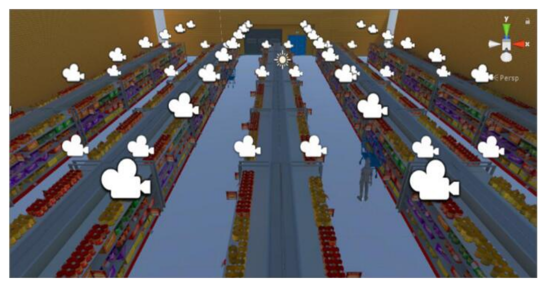

Fig. 7. Page Size and Optical Resolution

After setting up the scene, we also need to adjust the visual field and angle of each camera to get the best shooting pictures. Through calculation and field investigation, we draw the h-fov is 1.2 bin is best for catching the maximum range and clearer picture, and the each camera can cover 2 bin. In Fig.7, we can see the cameras which on shelves can clearly obtain the type and quantity of items, this provides a great help for our later identification.

Fig. 8. Page Size and Optical Resolution

The simulated scene is based on the real scene, it can serve as a reference. In this scene, we can design the placement of cameras, and calculate the location of the camera in the real scene and the number of cameras needed. Unity3D provide method of capturing each camera image, use the clear screenshot F9 to get the picture which is describe the items type and quantity on shelves, and it is resolution is 1600*1200. Then use pictures what we get to do item recognition and identify their type and number, ultimately get the data for the all items, calculate what kind and quantity customers take away.

B. Item Recognition

After setting up the scene, we need to add an image recognition system to each camera in the scene. As mentioned above, YOLOv3 is suitable for items image recognition, we will use this recognition method for items image processing. The YOLOv3 depends on OpenCV3.40 and CUDA8.0, which are our tools for training and recognition. There are other ways to train and recognize our data. TensorFlow [14] is a free and open-source software library for dataflow and differentiable programming across a range of tasks, it can as well be used in computer vision processing. In our simulation, OpenCV still is our main train and recognize, because it's easier to operate in our limited equipment.

In order to recognize the picture what we get for the scene's cameras, we have to pre-process the types of items and train them. There are lots of goods pictures in the internet, then we selected 1000 representative types for us to train. In order to make the simulation results closer to the real scene, the goods we selected were placed on the shelves and they were placed according to the way of the shopping mall. After obtaining samples pictures, these pictures should be marked so that the system can remember their characteristics and what exactly item is.

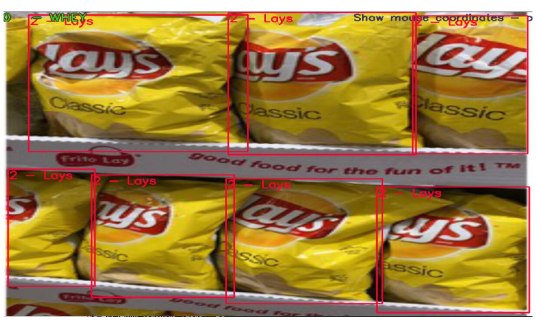

Fig. 9. Page Size and Optical Resolution

After marking all items, we will put all the pictures and data of the pictures into our pre-designed YOLO architecture for training. The purpose of training is to get the weight of each item so that our system can distinguish their type and accurately judge when inputting new pictures. In our simulator, we had 5,000 iterations of the data to get a lowest loss. When all the data has been trained, new data can be allowed to enter. We selected a picture that was not included in our data set to try the training results. The category of this item is the one that has been trained by our system. We searched the pictures of this item randomly on the Internet and put them into our system for identification. The recognition result as shown in Fig.

Fig. 10. .

Our system can accurately identify the object and the number of objects in this picture, its recognition rate can reach 98%. This means that when the camera in the scene intercepts the picture and puts it into our system, it can accurately identify the type and number of them. It shows that the system we built is feasible to be used in real scenes. Ultimately we'll have it with multi cameras to achieve cashier-free supermarket.

References

[1] U. M. Erdem and S. Sclaroff, "Automated camera layout to satisfy task-specific and floor plan-specific coverage requirements," *Computer Vision and Image Understanding*, vol. 103, no. 3, pp. 156–169, 2006, special issue on Omnidirectional Vision and Camera Networks.

[2] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, "Landmarc: indoor location sensing using active rfid," in *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003. (PerCom 2003).*, Mar. 2003, pp. 407–415.

[3] V. Vaish, B. Wilburn, N. Joshi, and M. Levoy, "Using plane + parallax for calibrating dense camera arrays," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, Jun. 2004, pp. I–I.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[6] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 650–657.

[7] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015.

[8] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[9] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 391–405.

[10] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[11] K. Ehsani, H. Bagherinezhad, J. Redmon, R. Mottaghi, and A. Farhadi, "Who let the dogs out? modeling dog behavior from visual data," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[12] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[13] K. K. Ball, B. L. Beard, D. L. Roenker, R. L. Miller, and D. S. Griggs, "Age and visual search: expanding the useful field of view," *J. Opt. Soc. Am. A*, vol. 5, no. 12, pp. 2210–2219, Dec. 1988.

[14] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. Savannah, GA: USENIX Association, 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi