ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

STOREFRONT LOGO RECOGNITION AND STEREO VISION BASED DISTANCE ESTIMATION

M.Sc. THESIS

Mehmet BİBERCİ

Department of Computer Engineering

Computer Engineering Programme

SEPTEMBER 2019

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

STOREFRONT LOGO RECOGNITION AND STEREO VISION BASED DISTANCE ESTIMATION

M.Sc. THESIS

Mehmet BİBERCİ
(504121528)

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Prof. Dr. Uluğ BAYAZIT

SEPTEMBER 2019

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

MAĞAZA LOGOSU TANIMA VE STEREO GÖRÜNTÜ TABANLI MESAFE KESTİRİMİ

YÜKSEK LİSANS TEZİ

Mehmet BİBERCİ
(504121528)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Uluğ BAYAZIT

EYLÜL 2019

Mehmet BİBERCİ, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 504121528 successfully defended the thesis entitled "STOREFRONT LOGO RECOGNITION AND STEREO VISION BASED DISTANCE ESTIMATION", which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Uluğ BAYAZIT** Istanbul Technical University	.....
Jury Members :	**Assoc. Prof. Dr. Sanem SARIEL** Istanbul Technical University	.....
	Prof. Dr. Muhittin GÖKMEN MEF University	.....
		.....

Date of Submission : **4 September 2019**
Date of Defense : **4 September 2019**

To my significant other

FOREWORD

I would like to thank my advisor Prof. Dr. Uluğ Bayazıt for his guidance throughout my thesis. I would also like to thank my friends and faculty members at the Computer Engineering Department.

September 2019 Mehmet BİBERCİ

TABLE OF CONTENTS

Page

ABBREVIATIONS

3D	**:** Three Dimensional
6DOF	**:** Six Degrees of Freedom
CIE	**:** Commission Internationale de l'Eclairage
CNN	**:** Convolutional Neural Network
IMU	**:** Inertial Measurement Unit
IR	**:** Infrared
LoG	**:** Laplacian of Gaussian
OCR	**:** Optical Character Recognition
RGB	**:** Red Green Blue
SAD	**:** Sum of Absolute Differences
SIFT	**:** Scale Invariant Feature Transform
SURF	**:** Speeded-Up Robust Features
SVM	**:** Support Vector Machine
YCbCr	**:** Luminance; Chroma: Blue; Chroma: Red
YOLO	**:** You Only Look Once

LIST OF TABLES

Page

LIST OF FIGURES

Page

STOREFRONT LOGO RECOGNITION AND STEREO VISION BASED DISTANCE ESTIMATION

SUMMARY

Since interpretation of images to compute structural properties of the 3D world is an important benefit of computer vision, computer vision applications can help people requiring navigational assistance. In this regard, this work presents a stereo-vision-based perception and navigation approach to assist visually impaired people, which is also transferable to robotics and autonomous vehicles, since the essence of the problem is the same for both of these cases. Frontal view images of stores in a shopping mall are first searched against a database for logo recognition where all the logos are trained beforehand. Distances to the found logos (store signboards) are estimated by stereo matching using geometrical identities and triangulations, mimicking the way human binocular vision perceives depth and 3D structure. Both logo recognition and stereo matching are based on local image features (keypoint descriptors) calculated via Speeded Up Robust Features (SURF) algorithm, which is a widespread approach in computer vision for object recognition due to its robustness to external effects on the object appearance. Final refined distances are calculated using statistical filtering and averaging of the individual keypoint distances found by matching keypoint pairs. For comparison, the distances are also estimated using a convolutional neural network (CNN), where the logos are identified by the CNN and their middle points are stereo-matched. Commonly available stereo datasets and most of the stereo applications are oriented towards dense disparity and depth map calculations where depth values of individual pixels are computed to obtain 3D structure of the whole scene. Since they are not appropriate for the problem at hand, a stereo dataset for logo recognition with depth estimation has been constituted for testing purposes by capturing storefront images with a custom built stereo rig, where the ground truth distance and camera viewpoint angle have been measured for each pair of snapshots. Experimental results on this self-generated stereo dataset of 28 storefront images from various distances and viewpoints demonstrate the performance of the proposed approach. An acceptable estimation accuracy has been obtained, which motivates further enhancements, development in other navigational applications, and integration with handheld devices.

MAĞAZA LOGOSU TANIMA VE STEREO GÖRÜNTÜ TABANLI MESAFE KESTİRİMİ

ÖZET

Görme duyusu, çevreyi algılama ve yön bulmada en önemli faktördür. İnsanlar için günlük bir rutin olan bu algılama, robotik ve otonom araçlarda da temel bir gereksinimdir. İnsanlardaki görme bozuklukları, çevre ile ilgili bilgi akışını sekteye uğratarak yön bulma faaliyetlerini kısıtlar ve insanlara güçlük vererek dezavantajlı duruma düşmelerine sebep olur. Bilgisayarla görü, görme duyusunu ikame edecek araçları temin eder ve böylece eksik görsel bilgilere ulaşma imkanı sağlar. Stereo görüntü algoritmaları, üç boyutlu mekansal bilgileri elde etmek ve insanın iki gözlü görüş sisteminin fonksiyonunu karşılamakta olduğu gibi robotik ve otonom araçlarda kullanılmak için de idealdir.

Tabelalar, sembol ve metin içermeleri sebebiyle, özellikle daha önceden bilinmeyen ortamlarda, yönlendirme ve gezinme konusunda öne çıkan öğelerdir. Dolayısıyla, üç boyutlu mekansal bilgi ile birlikte tabela tanıma, hedefe erişimi kolaylaştırır. Alışveriş merkezi, yol tarifine ihtiyaç duyulan iç mekan ortamına bir örnektir. Belirli mağazalara erişmek istendiğinde, hedefi ve yolu belirlemek için tabelalardan ve logolardan yardım alınmaktadır. Görme bozukluğu bu konuda büyük bir dezavantaja neden olmakla birlikte, bilgisayarla görü yöntemleriyle ve bu çalışma özelinde logo tanıma ile telafi edilebilir.

Son yıllarda, bilgisayarla görü algoritmaları ve dijital kameralar, yardımcı/destekleyici teknolojiler için kullanışlı uygulamalara elverecek şekilde önemli bir gelişme kaydetmiştir. Bu teknolojiler günümüzde akıllı telefonlarda da kullanılabilmektedir. Son araştırmalar, engellerden kaçınma amaçlı üç boyutlu görüntü algoritmaları ile nesne, işaret ve metin algılama/tanıma algoritmalarına odaklanmıştır. Bununla birlikte, görme engellilerin yön bulması için stereo görüntü ve nesne tanımayı birleştiren algoritmalar henüz erken bir aşamadadır. Ayrıca, mağaza logolarının/tabelaların tanınması ve görme engellilerin mağazalara erişiminin kolaylaştırılması ile ilgili herhangi bir çalışmaya rastlanmamaktadır.

Bu çalışma, logo tanımanın yanı sıra bir alışveriş merkezindeki mağazalara olan mesafeleri kestirmede stereo görüntü tabanlı bir yaklaşım sunmaktadır. Buradaki asıl amaç, belirli hedefleri tanıyarak yol tarif etmektir, ki buradaki hedefler de mağazalardır. Birincil katkı olarak, SURF nokta eşleştirme yöntemi, hem logo tanıma hem de stereo görüntü eşleştirme için kullanılmıştır. Mesafe kestirimi ise geometrik hesaplamalar ve istatistiksel iyileştirmelerle yapılmıştır. Bu amaç doğrultusunda, gerçek mesafeler ölçülerek mağaza vitrin ve tabela resimlerinden oluşan bir stereo veri seti oluşturulmuştur. Bu iş için hususi olarak bir stereo teçhizat yapılmış ve bununla mağaza önü resimleri çekilmiştir. Her seferinde gerçek mesafe ve kamera bakış açısı ölçülmüştür. Bu veri seti, logo tanımayla birlikte derinlik kestirimi için oluşturulması sebebiyle bir ilktir. Mevcut stereo veri setleri ve stereo uygulamaların çoğu, tüm

sahnenin 3B yapısını elde etmek için her pikselin derinlik değerinin hesaplandığı derinlik haritası hesaplamalarına yöneliktir.

Önerilen yöntem, kalibre edilmiş kamera parametrelerini ve sisteme önceden tanıtılmış logo görüntülerini kullanır. Stereo görüntüler önce gri tonlamaya dönüştürülür ve kamera parametreleri kullanılarak düzgünleştirilir. Sol görüntü SURF anahtar nokta eşleştirmesi kullanılarak sistemdeki logolarla eşleştirilir. Görüntü homografisi yoluyla, logonun sınırlayıcı kutusu hesaplanır ve bu, stereo eşleme için sağ görüntüdeki arama bölgesini tanımlar; burada tekrar SURF anahtar nokta eşleştirmesi kullanılır. Kamera parametreleri ve eşleşen noktaların yatay fark/sapma değerleri, üç boyutlu koordinatların hesaplanmasını, dolayısıyla mesafelerin hesaplanmasını sağlar. Mesafelerin ortalama değeri ve standart sapması, olası aykırı değerleri filtrelemek için kullanılır, bu da daha kesin bir mesafe kestirimi sağlar.

Yöntemin etkili bir şekilde uygulanması için aşağıdaki varsayımlar yapılmıştır. İlk olarak, kamera doğrudan karşıya bakmaktadır ve mağaza tabelaları her zaman yukarıdadır. Böylece görüntülerin sadece üst yarısı logolar için aranır. İkincisi, tabelalar dikdörtgen / düzlemsel bir şekle ve dik yönelime sahiptir, bu nedenle tanıma sırasında logonun döndürülmüş olup olmadığına bakmaya gerek yoktur. Tabelalar genellikle mağazaların girişlerinin üstünde bulunduğundan, mesafe ve yön hesaplarken hedef olarak kullanılır.

Logo tanıma, nesne tanımada yaygın bir yaklaşım olan SURF algoritmasını kullanır. Logo görüntülerinin SURF anahtar noktaları ve tanımlayıcı vektörleri önceden hesaplanır ve logo veritabanında saklanır. Logo tanıma için bir sorgu görüntüsü geldiğinde, üzerinde aynı şekilde anahtar nokta ve tanımlayıcı vektör hesaplaması yapılır. Sorgu görüntüsünün anahtar noktaları, tanımlayıcı vektörlerin Öklid uzaklığına bağlı olarak logo anahtar noktaları ile eşleştirilir. Eşleştirme işlemi, her bir anahtar nokta için en yakın iki eşleşme döndürür. En yakın mesafenin ikinci yakın mesafeye oranı belirli bir eşiğin üstünde ise, eşleşme reddedilir. Bu filtrelemeden sonra, kalan eşleşmeler görüntü homografisini hesaplamak ve logonun sınırlayıcı kutusunu bulmak için kullanılır.

Uzaklık kestirimi, stereo eşleştirme ve üçgenleme yoluyla 3B projeksiyondan sonra gelir. Bir noktanın sol ve sağ görüntülerdeki koordinat farkı/sapması ve kamera parametreleri biliniyorsa, o noktanın uzaklığı geometrik özdeşlikler kullanılarak hesaplanabilir.

SURF algoritması uygulanmadan önce sol ve sağ görüntülerin her ikisi de mercek kaynaklı deformasyonları gidermek için düzgünleştirilir ve ideal bir stereo sistemi elde etmek için hizalanır. Sağ görüntünün anahtar noktaları ve SURF tanımlayıcı vektörleri, tanıma adımında belirlenen logo sınırlayıcı kutusu tarafından tanımlanan sınırlı bir alanda hesaplanır. Sol ve sağ görüntü arasındaki anahtar nokta eşleşmesi aynı SURF metodolojisi kullanılarak gerçekleştirilir. Sonuç olarak ortaya çıkan eşleşmeler, koordinat sapma/fark değerlerinin hesaplanmasını sağlar. Kamera parametreleri ve sapmalar kullanılarak, anahtar noktaların üç boyutlu koordinatları hesaplanır. Yanlış eşleşmeler ve yanlış hesaplamalar olabileceğinden, aykırı değerlere sahip noktalar istatistiksel olarak filtrelenir. Kalan noktaların ortalaması, kestirilen mesafeyi verir.

Önerilen yöntemin logo tanıma performansı ve mesafe kestirim doğruluğu, mağaza mesafelerinin 5,48 ila 14,63 metre arasında; kamera bakış açısının ise yatayda -45 ila 45 derece arasında değiştiği veri seti üzerinde test edilmiştir. Önerilen yöntem

logoları direkt karşıdan ve 30 dereceye kadar olan kamera açılarında tanıyabilmektedir. Logo belirgin bir üç boyutlu yapıya sahipse ve kamera açısı 30 derecenin üzerindeyse tanıma başarısız olmaktadır. Mesafe kestirimindeki ortalama hata, %12'dir. Gerçek uzaklıkların ortalaması 10 metre olduğundan, bu 1,2 metrelik bir hataya karşılık gelmektedir.

Karşılaştırma amacıyla, derin öğrenme yaklaşımı ile bir mesafe kestirim sistemi daha gerçeklenmiştir. Burada evrişimli sinir ağı (CNN) yapısında olan Tiny YOLO v3 sistemi kullanılmıştır. İkinci bir magaza önü fotoğraf koleksiyonu toplanmış ve Tiny YOLO için eğitim veri seti oluşturmak amacıyla logo konumları elle işaretlenmiştir. Sistem bu veri seti ve uygun konfigürasyon ile eğitilmiştir. Sinir ağı, sol ve sağ görüntüleri girdi olarak alır ve logoları tanıyarak sınırlayıcı kutularını bulur. Bu sınırlayıcı kutularının orta noktaları koordinat sapma/fark değerlerinin hesaplaması için kullanılır, yani stereo eşleştirme sadece orta noktalarla yapılır. Daha sonra bu sapmadan mesafe kestirilir.

YOLO yöntemi, sınırlayıcı kutuların ve orta noktaların yeterince kesin olarak belirlenememesi ve dolayısıyla koordinat fark değerlerinin yüksek sapmalara sahip olması nedeniyle daha yüksek bir ortalama kestirim hatasıyla (%29) sonuçlanmıştır. Bununla birlikte, YOLO'nun logo tanıma performansı SURF'e yakındır. YOLO, farklı bakış açılarından ve şartlarda çekilen görüntüler ile eğitildiği için, eğitim veri seti yeterince çeşitli olduğu sürece kamera açısı ve diğer varyasyonlar tanıma konusunda problem oluşturmamaktadır.

Her ne kadar bir tripod üzerine monte edilmiş sabit dijital kameralarla geliştirilmiş ve test edilmiş olsa da, önerilen yöntemin koşulları sağlayan akıllı telefonlarda da uygulanması amaçlanmıştır. Akıllı telefonlar birden fazla kamerayla donatıldıkça, görme engellilere yardımcı olmak için önerilen yöntemi kullanabilir ve fazladan bir cihaza olan ihtiyacı ortadan kaldırabilir.

Elde edilen kestirim doğruluğu, sistemi geliştirmek ve başkaca navigasyon uygulamalarına entegre etmek için motive edici seviyededir.

1. INTRODUCTION

Vision is a key factor in perceiving and navigating through the environment, which is an everyday task for humans and a basic requirement in robotics and autonomous vehicles. Human visual impairment limits the information flow about the surroundings and restricts navigation activities, causing inconvenience and disadvantage. Computer vision provides the means to substitute eyesight, thus restoring the missing visual information. Stereo vision algorithms are ideally suited for inferring 3D spatial knowledge and compensate for the human binocular visual system, as well as being implemented in robotics and autonomous vehicles for the same purpose.

Featuring symbols and text, signboards are prominent elements of orientation and navigation, especially in unfamiliar environments. Hence, signboard recognition combined with 3D spatial information facilitates access to the destination. Shopping mall is an example of an indoor environment where we need directions. As we desire to access particular stores, we identify them on their logos to set our target and path. Visual impairment causes a major disadvantage in this respect, yet can be compensated by computer vision methods, and more specifically in our context by logo recognition.

In the past decade, computer vision algorithms and digital cameras have advanced substantially enabling convenient application in assistive technologies, which are now also available on smartphones. Recent research is focused on 3D vision for obstacle avoidance and detection/recognition algorithms for objects, landmarks, signs, and text. However, merging stereo vision with robust object recognition algorithms for blind navigation is at an early stage. Furthermore, store logo/signboard recognition and accessing to stores have not been introduced in any work related to facilitating assistance to the visually impaired.

This study is of a stereo vision based approach to estimate distances to stores in a shopping mall alongside logo recognition. The main objective is to detect and give directions to points of interest, which happen to be stores in this case. As the primary contribution, SURF [1] keypoint matching method is employed for both logo

recognition and stereo correspondence matching where distance estimations involve geometric calculations and refinements. Towards this end, a stereo dataset of storefront images with ground-truth distances is collected.

Although developed and tested with stationary digital cameras mounted on a tripod, the proposed approach is also intended to be implemented on qualified smartphones. As smartphones are becoming equipped with multiple cameras, they may run the proposed method to assist the visually impaired, eliminating the need for an extra device.

As the sensors and cues of the studied system are also to be encountered as on a mobile robot, the proposed method is further usable in robotic navigation applications.

2. RELATED WORK

2.1 Blind Navigation

Computer vision and image processing techniques aiming to provide visual data through auditory and haptic interfaces have been utilized in portable blind assistance systems since early 1990s.

The vOICe [2] is an image-to-sound conversion system introduced by Meijer, which scans images from left to right for sound generation. Frequency is a function of pixel row, where lower pixels have lower frequency while amplitude denotes pixel intensity. Processing all the columns results in a sequential superimposed audio signal.

Molton et al. [3] described a portable stereo vision-based obstacle detection system as a part of their Autonomous System for Mobility, Orientation, Navigation and Communication (ASMONC) project, where the user wears a backpack holding two cameras and other components. The system represents the ground plane with initial stereo disparity values and predicts the disparities for the following images as the user walks. Obstacles contradict the predicted disparities and are thereby detected. Kalman Filter is used when modeling the walking movement.

Zelek et al. [4] developed a prototype system featuring a dual camera setup and a tactile feedback glove. Images from the two cameras are used to calculate a depth map. The depth map is then divided into five vertical regions and each region is encoded as a vibration on the tactile unit, which happens to be a glove with pager motors attached to each finger.

SmartVision [5] navigation assistance system intends to incorporate a geographical information system with a stereo vision system. Its main objective is to help visually impaired people to navigate in unfamiliar environments as well as giving contextual information about obstacles and points-of-interest. The visual system uses Hough Transform to detect circular landmarks, which represent safe paths in the experiments. Trajectory angle is calculated based on the location of the detected circles. Providing

distance information from the stereo camera is stated as a future integration work. In the follow-up project Blavigator [6], range image segmentation is utilized to detect obstacles in 1m or 2m proximity.

Wang and Tian [7] proposed a method for detecting and recognizing restroom signs. Image regions containing signage are detected by comparing shape information of the connected components. The head part, modelled as a circle, has an approximate perimeter-to-area ratio. The body part is detected based on the relations to the head part. The detected signs are then recognized by SIFT [8] feature matching.

Schauerte et al. [9] presented an assistive vision system to find displaced objects. Specific textured objects are detected using SIFT feature matching, while a color-based approach is used to extend the detection capability towards a broader range. The objective is to guide the user's hand to the object locations through acoustic feedback. A small hand-held camera, which can also be attached to the wrist, facilitates hands-free movement. The object location is encoded as sound, of varying pan and pitch.

The Sound of Vision system [10] introduces a sensory substitution device employing data fusion. The system comprises of a stereo camera, an IR-based depth camera and an inertial measurement unit (IMU) to provide depth information in any environment at any illumination condition. Stereo stream along with IMU sensor enables 3D environment representation. Ground plane detection and object segmentation are also performed. The system further identifies doors, stairs, signs, and text through edge/line detection, support vector machine (SVM) classification, and optical character recognition (OCR).

2.2 Visual Odometry

Distance measurement is an essential problem of computer vision, featuring single or multiple camera algorithms in the field.

Mrovlje and Vrancic [11] implemented an algorithm to calculate the object distance based on stereoscopic images and basic trigonometric calculations. Images of the target object is taken with two identical and horizontally aligned cameras. The target object is selected manually on the left image and searched on the right image by matrix

subtraction, namely calculating pixel value differences. The subtraction occurs in a restricted search area, a horizontally expanded window of the selected object, since the proper horizontal alignment of the cameras. After the target object is located on the right image, its distance is calculated from its displacement on the images.

Neto et al. [12] proposed a distance measurement system to use in mobile robots using stereo vision and SIFT. The system features two identical cameras with a baseline (distance between cameras) of 120 mm. Two images are captured from the respective cameras. SIFT algorithm is applied to the images in order to detect and match keypoints. The scene is divided into 9 sub-regions, of which 5 of them (middle, top, bottom, left, right) are used for distance calculation. Distance of each keypoint is calculated using camera parameters and triangulation. In each sub-region the average distance of keypoints is used as the resulting distance.

Xiong et al. [13] presented a depth estimation method based on stereo vision and color segmentation with region matching in CIE Lab color space. Left and right images are first undergo a color segmentation to distinguish objects from the background. An automatic seeded region growing approach is used for color segmentation, where the seed pixel is selected according to its color difference to its neighbours and the region grows again according to the similar metric. For each segmented region, geometrical properties including area, perimeter and centroid are also calculated. The segments on the left image are then matched to the segments of the right image based on color and these geometrical properties. Finally, the object distance is estimated thru 3D reprojection using camera parameters, pixel coordinates and disparity which is obtained after matching.

Nunez et al. [14] described a system to estimate robot motion from temporal stereo images. The system aims to calculate the displacement and estimate the 6DOF pose of a moving robot, thus solve the localization problem. The proposed approach consists of two matching steps. Left and right images coming from the stereo camera of the mobile robot are first matched to find keypoints and calculate their respective 3D locations. This matching and calculation is applied to each stereo pair, as continuous data come from the camera. Then, between each consecutive frames, the calculated keypoints are matched to estimate robot displacement. Both matching steps employ SIFT keypoints

and descriptors for feature matching, so the output of the first matching step is the input to the second matching step.

Kamencay et al. [15] proposed a hyrbrid approach to estimate depth map from stereo image pairs. A color based hybrid segmentation method using K-Means Clustering and Belief Propagation algorithms is implemented to split the image into regions. The resulting segments are the basis for the stereo matching stage. A combination of SIFT and Sum of Absolute Differences (SAD) algorithms called SIFT-SAD is applied to these segments for stereo matching. The results of the matching are then used for depth map estimation thru 3D reconstruction.

Kim et al. [16] suggested a concept for estimating distance between vehicles with a single camera. The vehicle in the back is equipped with a camera facing forward, thus getting rear-lamp images of the front vehicle. Based on the perspective that the positions of the left and right rear lamps change with respect to the distance as well as the apex angle of the triangle formed by left lamp, camera, right lamp; the distance is estimated from this camera viewpoint angle. The rear lamps of the front vehicle is detected by red color detection. Then the angle is calculated from the aforementioned triangle. Relying on the assumption that the distance and the angle vary between a certain range, a look-up table of angles and distances are made. The distance is then estimated using the calculated angle and the look-up table.

Güzel and Nattharith [17] introduced a method to estimate distance using a single camera and a neural network. Image coming from the camera is first matched against a dataset using SIFT algorithm to detect the target object. Using the scale parameter of the SIFT descriptor and the camera zoom factor, the distance is then estimated by a neural network. The target objects for the training phase are images of book covers. Images of these books are taken from various distances with particular zoom factors. SIFT algorithm is applied on these images to obtain the scale parameter. A neural network is trained with these scale paramaters and zoom factors, which is used for distance estimation.

Kordelas et al. [18] presented a dense disparity estimation approach from stereo images using a combined matching metric consisting of RGB color information, Census transform and SIFT. Rectified stereo pair images are first segmented using an external

mean-shift segmentation tool. The resulting segmentation maps define the application region of the matching algorithms. Absolute differences of each metric; RGB values, Census coefficients and SIFT features constitute the initial step of pixel matching. The matching cost (absolute difference) for each metric is aggregated over the neighbouring pixels using adaptive weights for each neighbour. These costs are then combined into one metric. The calculated disparities are further refined by outlier detection, disparity histogram analysis and plane fitting to estimate the final disparity image.

Pramote and Piamsa-nga [19] proposed a method to compute disparity map from stereo images using SIFT and YCbCr color model. SIFT keypoints and descriptors are calculated on the stereo image pair and matched between the left and right image. These matches are used to restrict the search region in the next steps. SIFT serves to speed up the matching process and the YCbCr color model serves to improve accuracy. Each pixel in the search region of the left image is matched against the right image using SAD algorithm, where the difference metric is a weighted function of YCbCr color model components. Thus, a patch window centered at that pixel is color compared to its counterpart in the other image of the stereo pair to obtain the disparity value.

3. METHOD

3.1 Stereopsis

The perception of depth and thereby 3D structure thru visual information coming from two eyes is denoted as stereopsis. As the two eyes are located at separate points, each of them has its own viewpoint and there is a slight difference between their perception of the same scene, nevertheless the two views overlap significantly. This phenomenon is imitated in computer vision by taking images with two cameras from different viewpoints, and called stereo vision. Figure 3.1 shows the concept of stereo vision.

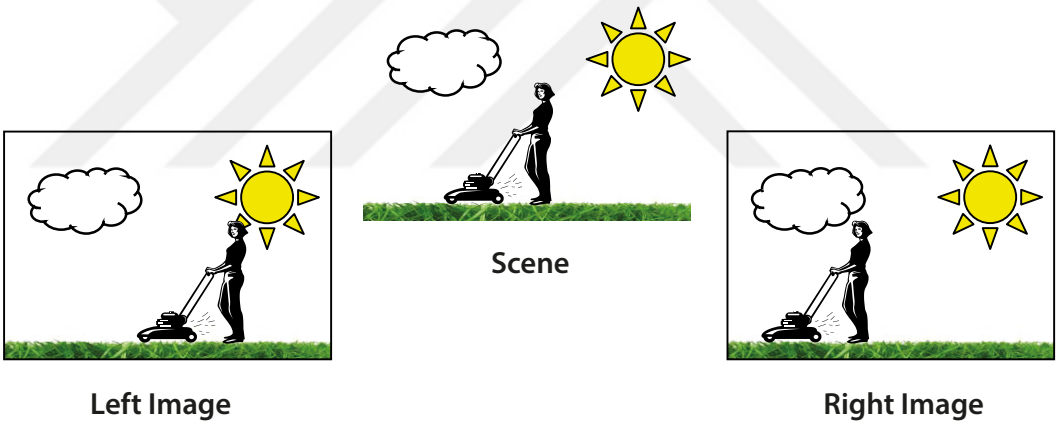


Figure 3.1 : Left and Right Images of a Scene [20].

Closer objects have a larger displacement between the left and right image, while distant objects have a very slight displacement and are approximately at the same position. This displacement is called disparity and constitutes the basis of depth estimation in stereo vision algorithms, since it is proportional to the distance of the target object. Disparity together with camera parameters leads to distance calculation as explained in detail in Section 3.4.2.

3.2 Image Rectification

Cues of the 3D world are not kept true when imaging with cameras. Camera images have distortions due to the spherical shape of the camera lens, called radial distortions. Straight lines are curved when passing thru the lens and appear bent on the image. This *barrel distortion* effect is nonexistent at the center of the image but increases noticeably when moving towards the edges. Figure 3.2 demonstrates the formation of barrel distortion.

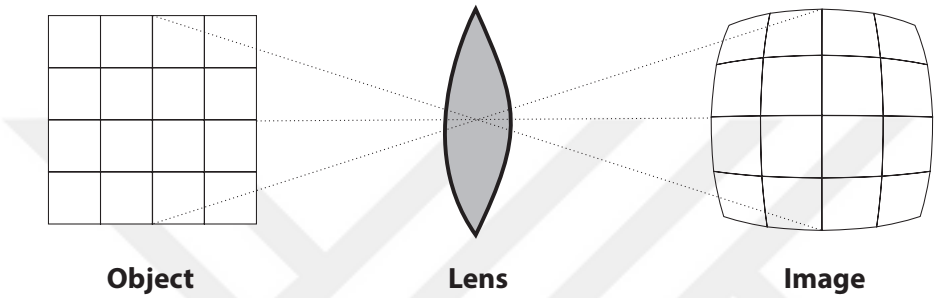


Figure 3.2 : Barrel Distortion.

Stereo correspondence matching is subject to epipolar constraint, i.e. a pixel's match in the other image can only lie on the respective epipolar line, a specific line defined for each pixel. Figure 3.3 shows the epipolar geometry and epipolar line (red) for a point X. If the left and right cameras of the stereo system have the same 6DOF pose except a horizontal offset, then the epipolar line of a pixel is horizontal and lies on same image row as that pixel. Thus, stereo matching is simplified by only searching on a specific image row.

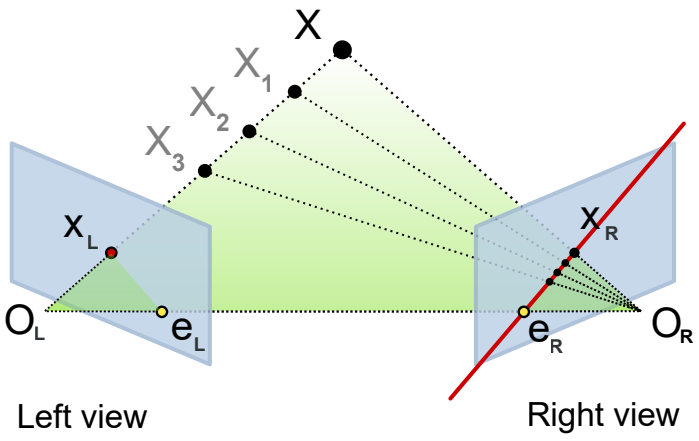


Figure 3.3 : Epipolar Geometry [21].

Image rectification is the step where radial distortions are corrected and epipolar lines become horizontal before moving on with the stereo matching, ensuring image coplanarity. This procedure calculates distortion coefficients, intrinsic parameters (focal lenghts and optical centers) and extrinsic parameters (rotation and translation between the two cameras). The calculation is based on a set of images of a pattern like checkerboard, where the relative positions of specific points (e.g. corners) and real-world sizes are known beforehand. All these calculated parameters are then used for rectification.

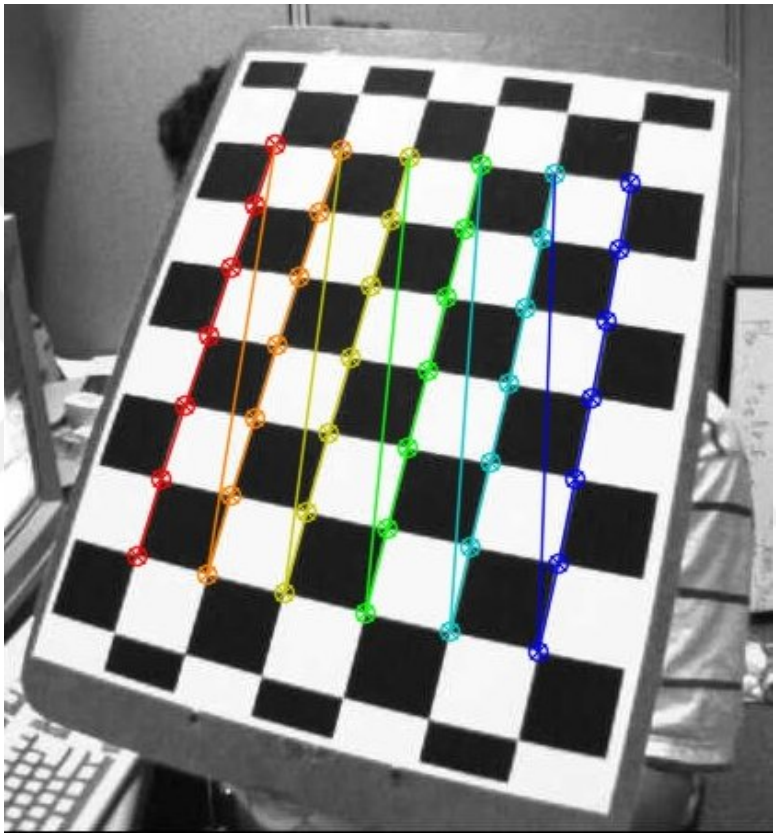


Figure 3.4 : Checkerboard Calibration [22].

3.3 SURF

SURF is a widespread approach in object recognition and a faster alternative to another widespread local feature detector, SIFT. It is robust to common image transformations including rotation, scaling, illumination and slight viewpoint change. Its first step is the keypoint detection, where the image is searched at different scales. Blobs and corners are appropriate candidates in this sense, since they are likely to be found in different images of a particular object. SURF uses a blob detector based on Hessian matrix

and box filters to speed up keypoint detection. Second step is the calculation of the descriptors. The SURF descriptor of each keypoint is generated using horizontal and vertical Haar Wavelet responses of the region around the keypoint. These wavelet responses indicate significant changes in horizontal and vertical directions. The resulting descriptor is a 128-dimensional vector. Figure 3.5 shows detected keypoints on a sample image with SURF.

Figure 3.5 : SURF Keypoints [23].

The idea behind scale invariant features and thereby SURF is that they are local maxima of the image at the respective scale σ_1, so that the same feature is also to be found as a local maxima on a different size of the image at a different scale σ_2, where the ratio of the two σs corresponds to ratio of the image sizes.

In order to detect these maxima and consequently keypoints, SURF uses the Hessian matrix. The Hessian matrix defines the local curvature of a function and its determinant quantifies the intensity of this curvature. The Hessian matrix is calculated at each pixel and defined as:

$$H(x,y,\sigma) = \begin{bmatrix} L_{xx}(x,y,\sigma) & L_{xy}(x,y,\sigma) \\ L_{xy}(x,y,\sigma) & L_{yy}(x,y,\sigma) \end{bmatrix} \tag{3.1}$$

Each element of the matrix is the convolution of the respective Gaussian second order derivative with the image at point (x,y), called Laplacian of Gaussian (LoG). SURF

approximates these second order derivatives with box filters as shown in Figure 3.6. These box filters allow the use of integral images and speed up the computation dramatically. Further acceleration is achieved by scaling only the box filters and keeping the image constant for scale space search, eliminating the requirement for up- and down-sampling.

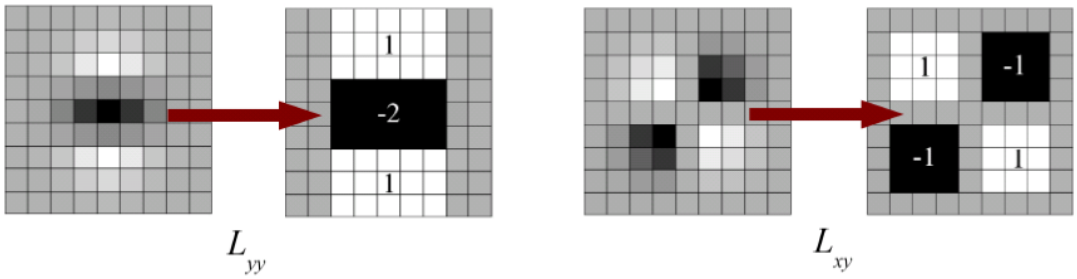


Figure 3.6 : Box Filter Approximations of LoG [23].

After keypoint detection, SURF descriptor is calculated using simple Haar Wavelets, measuring vertical and horizontal local intensity differences in an interest region of size 20σx20σ around the keypoint. This square region is divided into 16 square subregions. For each subregion, wavelets are computed and summed forming the vector $\left[\sum dx \quad \sum dy \quad \sum |dx| \quad \sum |dy|\right]$. Computing the sums dx and $|dx|$ for $dy < 0$ and $dy \geq 0$ individually; and the sums dy and $|dy|$ for $dx < 0$ and $dx \geq 0$ likewise yields a 8-dimensional vector. Since there are 16 subregions, the resulting descriptor is a 128-dimensional vector.

3.4 Implementation

The method makes use of calibrated camera parameters and trained logo images. Stereo images are first converted to grayscale and rectified using the camera parameters. The left image is queried against the trained logos using SURF keypoint descriptor matching. Through image homography, the bounding box of the logo is calculated, which defines the search region in the right image for stereo matching, where again SURF keypoint descriptor matching is employed. Disparity values of the keypoints along with camera parameters enable calculation of the 3D coordinates, hence the distances. The mean and standard deviation of the distances are utilized to filter possible outliers, allowing a more accurate distance estimation. Figure 3.7 shows an overview of the proposed method.

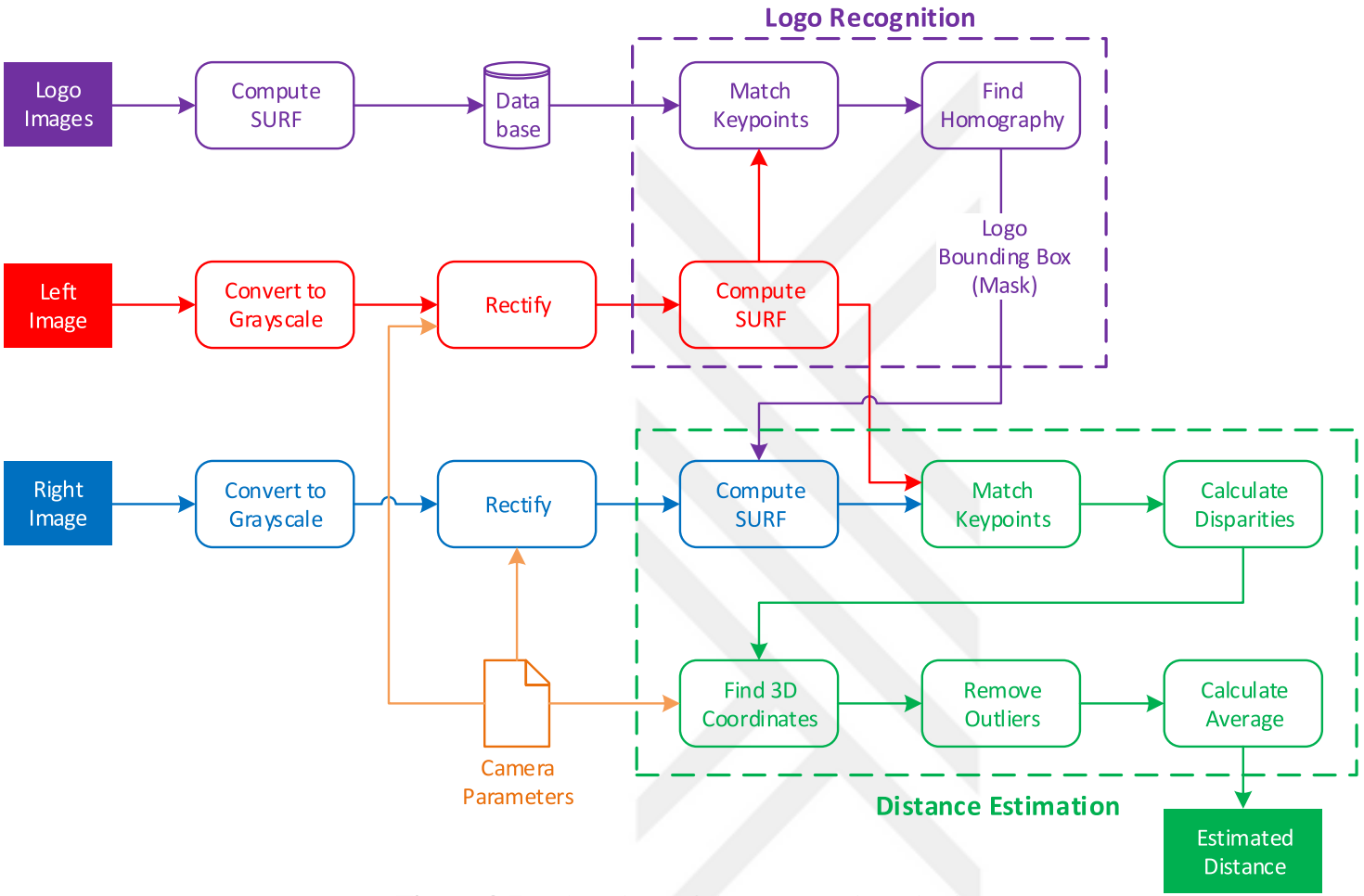
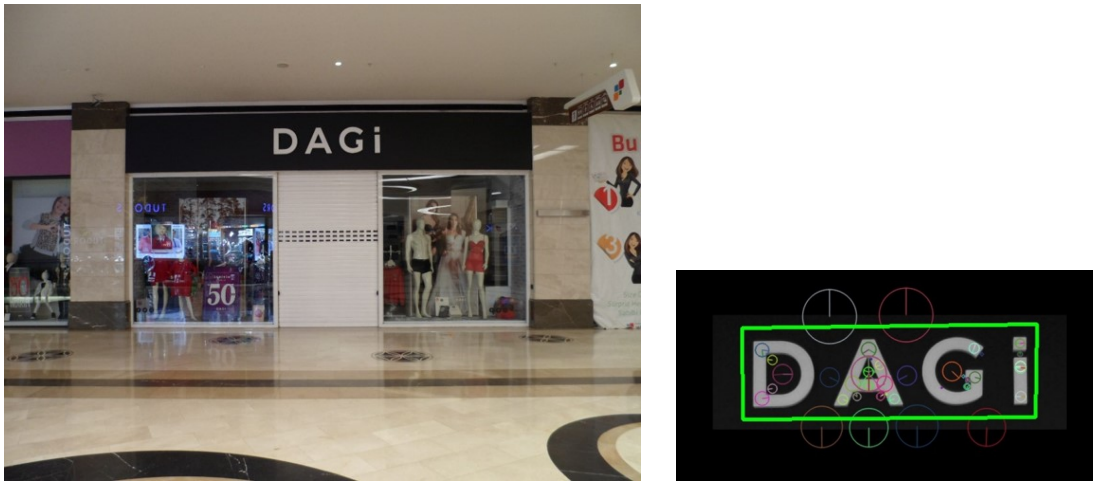


Figure 3.7 : Flowchart of the Proposed Method.

The following assumptions are made for an efficient implementation of the method. Firstly, camera is not tilted and store signboards are situated overhead. Thus only the upper half of the images are searched for logos. Secondly, signboards have a rectangular/planar shape and upright orientation, so there is no need to consider rotation during recognition. Since signboards are usually located above the entrances of the stores, they are used as targets when calculating distance and directions.

3.4.1 Logo recognition

Logo recognition employs SURF, which is a widespread approach in object recognition due its robustness. SURF descriptors of the logo images are calculated beforehand and stored in the logo database. When a query image is searched for logos, the same descriptor calculation is performed. Keypoints of the query image are matched with logo keypoints based on the Euclidean distance of the descriptors. In order to avoid false matches, a ratio test is applied as suggested in [8]. The matching process returns two nearest matches for each keypoint. If the ratio of nearest-distance to the second-nearest-distance is above a threshold, the match is rejected. After this filtering, remaining point correspondences are used to estimate the image homography and find the bounding box of the logo. Figure 3.8 shows a sample run of logo recognition where the query image is searched and the detected logo is indicated.



(a) Query Image (b) Recognized Logo

Figure 3.8 : Logo Recognition.

3.4.2 Distance estimation

Distance estimation succeeds stereo correspondence matching and 3D reprojection via triangulation. When the coordinate difference of the same point on the left and right images -*disparity*- and the camera parameters are known, then the distance of the point can be calculated using Equation 3.2 in an *ideal* stereo system.

$$Z = \frac{f.T}{d} \quad \text{where } d \text{ is disparity} \tag{3.2}$$

Figure 3.9 shows the geometry of a perfectly aligned and rectified stereo system, where T is the horizontal distance between the cameras, f is the focal length, and disparity is defined as the x-coordinate difference of p_l and p_r. The left camera center O_l is the origin of the 3D world coordinate system.

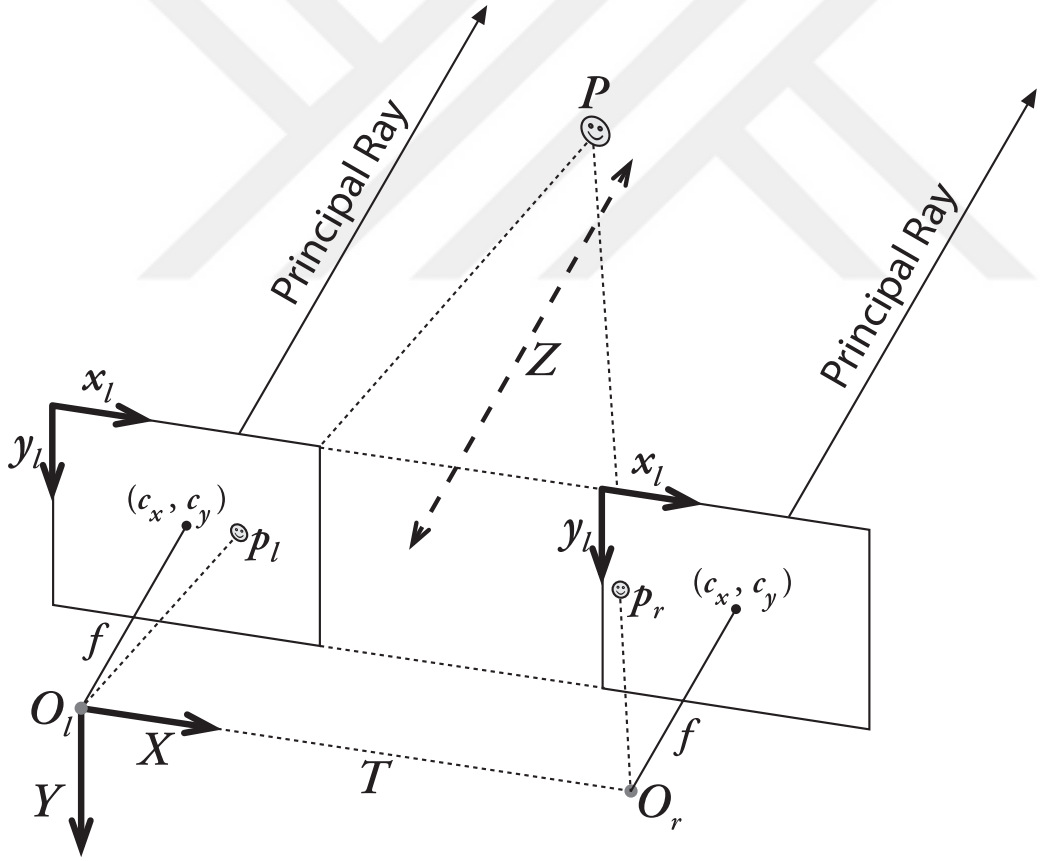


Figure 3.9 : Stereo Geometry [24].

Both left and right images are rectified and aligned to eliminate lens distortions to realize an ideal stereo system before proceeding with SURF. Keypoints and SURF descriptors of the right image are calculated on a restricted area defined by the logo bounding box found in the recognition step. Keypoint matching between the left and

right image is accomplished using the same methodology above. The resulting point correspondences enable calculation of disparity values. Using camera parameters and disparities, 3D coordinates of the keypoints are calculated. Since there may be mismatches and miscalculations, the points not satisfying the following condition (outliers) are filtered out:

$$\mu - \sigma \leq P_Z \leq \mu + \sigma$$

where μ is the mean and σ is the standard deviation of the distances P_Z. Averaging over the remaining points yields the estimated distance. The direction angle can be estimated geometrically by dividing the average of P_X values (X coordinates) by the estimated distance, as the 3D world coordinates are relative to the left camera center.

3.5 You Only Look Once (YOLO)

YOLO [25] is an object detection system based on Darknet [26] neural network framework and has a Convolutional Neural Network (CNN) architecture. It handles the object recognition task differently and processes the whole image on a single network at one pass, where prior systems first detect the possible regions containing objects and then perform classification. Hence, it outperforms other methods in terms of speed.

YOLO divides the image into a grid of cells and each cell predicts if it contains any object and bounding box of the objects alongside with class probabilities if any found. The network consists of convolutional and max pooling layers. As of version 3, YOLO has 75 convolutional layers and others include route, upsample, shortcut layers. YOLO also features a small network called Tiny YOLO, which is fit for limited conditions, runs faster and has only 13 convolutional layers.

For comparison purposes, a stereo distance estimation system implemented using Tiny YOLO v3. First another set of storefront images are collected and manually annoted for logo locations to generate a training set for Tiny YOLO v3. A network is then trained with this dataset and appropriate configuration to detect logos in the stereo dataset. The neural network takes the left and right images as inputs and recognizes logos and their bounding boxes. The middle points of these bounding boxes are used for disparity calculation, i.e. stereo matching is done only on the middle points. The distance is then estimated from this disparity.

4. EXPERIMENTS

In order to evaluate the performance of the proposed method, a stereo dataset of frontal store images is generated. Two digital compact cameras are mounted on a bracket side by side and attached to a tripod. Making sure that the stereo rig is parallel to the ground, frontal images of 11 stores are captured from various distances and viewpoints, resulting in 28 stereo pairs. For each image pair, ground truth distance is measured with a laser rangefinder while registering the stereo rig location. Figure 4.1 shows the experimental setup.

Figure 4.1 : Stereo Rig and the Laser Rangefinder.

In the resulting dataset, store distances range from 5.48 to 14.63 meters whereas the camera yaw angle varies from -45 to 45 degrees. The logo recognition performance and distance estimation accuracy of the proposed method are evaluated. The method can recognize logos *en face* and handle camera yaw variations up to 30 degrees. The recognition fails if the logo has an apparent 3D structure and the camera yaw is above 30 degrees, which is the case with one particular store (see Figure 4.4). The average error in estimated distances is 12%, corresponding to 1.2 meters as the average ground truth distance is 10 meters. Mrovjle and Vrancic [11] had 8.3%, 5.3%, 0.1% and 3.4% estimation errors in their four experiments, selecting the target object manually and using SAD matching for a distance of 10 meters. Detailed estimation results can be seen in Table 4.1. Sample stereo matching after a successful logo recognition is shown in Figure 4.2, where the query logo is Figure 4.3

Table 4.1 : Distance Estimation Results

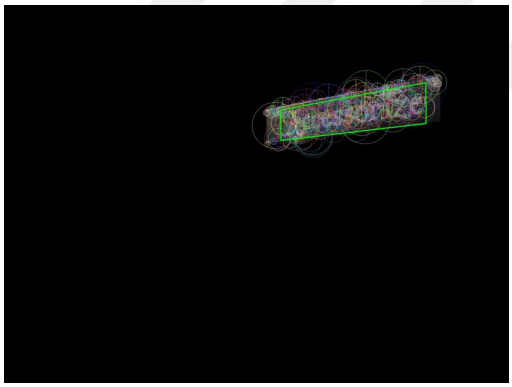
Logo	Camera Yaw[°]	Actual Distance[m]	Estimation Error (SURF)	Estimation Error (YOLO)
Lacoste	0	12.86	15%	Fail
Starbucks	0	9.95	8%	40%
Starbucks	30	14.03	1%	891%
W	0	5.48	12%	4%
W	45	7.61	12%	59%
Altınyıldız	0	11.20	6%	51%
Altınyıldız	30	11.67	25%	25%
English Home	30	9.10	4%	22%
English Home	45	9.10	21%	91%
Derimod	0	8.28	3%	29%
Derimod	-15	8.70	9%	39%
Inci	0	7.91	8%	24%
Inci	-30	10.29	2%	13%
Inci	30	9.06	32%	23%
Roman	0	8.32	12%	Fail
Roman	-45	9.79	Fail	Fail
Roman	-30	9.79	Fail	Fail
Accessorize	0	8.85	25%	2%
Accessorize	-45	10.16	1%	26%
Accessorize	-15	10.16	1%	Fail
Dagi	0	9.02	19%	23%
Dagi	-30	10.04	20%	14%
Dagi	-15	10.04	25%	46%
Boyner	0	13.16	16%	8%
Boyner	-30	14.63	6%	5%
Boyner	-15	14.63	4%	171%

(a) Left Image

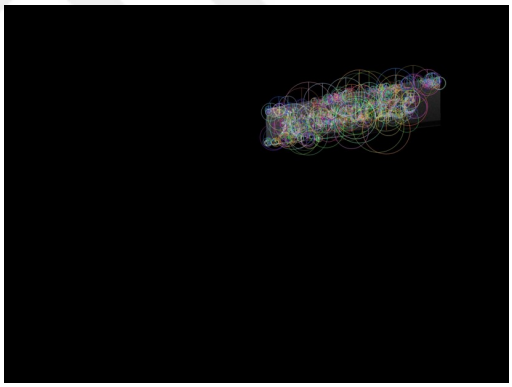
(b) Right Image

(c) Rectified Left Image

(d) Rectified Right Image



(e) SURF on Left Image



(f) SURF on Right Image

(g) Stereo Matching

Figure 4.2 : Logo Recognition and Stereo Matching.

Figure 4.3 : Query Logo for the Sample Run in Figure 4.2.

Estimation errors in the experiments are likely caused by the focus incoordination of the cameras. The proposed method can be benchmarked to aforementioned results of [11] only in terms of stereo distance estimation (SURF vs. SAD), having a similar baseline and a ground truth distance but lacking object detection, hence the manual object selection. As comparison, the distances are also estimated using YOLO, which resulted in a higher average estimation error (29%) because the logo bounding boxes and middle points are not determined precise enough and therefore the disparity values have high deviations. However the logo recognition performance of YOLO is similar to SURF in terms of identifying the logo. Since YOLO is trained with images from various viewpoints and of different conditions, camera yaw and other variations are not a concern regarding recognition, as long as the training data is diverse enough. The astronomical errors 891% and 171% in Table 4.1 are due to very small disparity values resulting from the coarse bounding boxes in stereo images. A sample run of the YOLO method is shown in Figure 4.5.

(a) Upright View

(b) Side View

Figure 4.4 : Failed Logo Recognition with Apparent 3D Structure.

(a) Left Image

(b) Right Image

(c) Rectified Left Image

(d) Rectified Right Image

(e) YOLO on Left Image

(f) YOLO on Right Image

Figure 4.5 : Logo Recognition with YOLO

Along with the storefront dataset, a calibration set consisting of stereo pair images of a checkerboard pattern is generated to calculate the camera parameters and use them in the experiments. Figure 4.6 shows one pair of the calibration set.

(a) Left Image (b) Right Image

Figure 4.6 : A Calibration Image Pair.

5. CONCLUSION

Logo recognition with stereo vision can assist the visually impaired to navigate more easily in an unfamiliar environment. Robust algorithms can handle variations and recognize logos in most instances. As the experimental results demonstrate, the method achieves a reasonable distance estimation accuracy. A finer calibration and alignment e.g. factory calibrated stereo rig and coordinated focusing of the cameras will improve the accuracy. Further integration with a text-to-speech engine can create a convenient navigation aid.

As the proposed method employs the SURF algorithm, which can detect any kind of textured object, further visual navigation applications for humans as well as autonomous robots/vehicles are likely to be developed.

Regarding the recognition failures arising out of wider camera yaw angles or apparent 3D structures of logos, the trained logo set can be expanded to include different viewpoints and angles, so that the system proceeds with the expanded set in case of such failure.

As YOLO achieves a similar accuracy for logo recognition, it can be further trained to detect the bounding boxes more precisely, provided that a comprehensive dataset is available. A neural network can also be trained to estimate the distances directly, again a comprehensive and distance-annoted dataset is necessary.

The two methods can also be used complementarily to increase accuracy and overcome cases of failure.

REFERENCES

[1] **Bay, H.**, **Tuytelaars, T. and Van Gool, L.** (2006). Surf: Speeded up robust features, *Computer vision–ECCV 2006*, 404–417.

[2] **Meijer, P.B.** (1992). An experimental system for auditory image representations, *IEEE transactions on biomedical engineering*, *39*(2), 112–121.

[3] **Molton, N.**, **Se, S.**, **Brady, J.**, **Lee, D. and Probert, P.** (1998). A stereo vision-based aid for the visually impaired, *Image and vision computing*, *16*(4), 251–263.

[4] **Zelek, J.S.**, **Bromley, S.**, **Asmar, D. and Thompson, D.** (2003). A haptic glove as a tactile-vision sensory substitution for wayfinding, *Journal of Visual Impairment and Blindness*, *97*(10), 621–632.

[5] **Fernandes, H.**, **Costa, P.**, **Filipe, V.**, **Hadjileontiadis, L. and Barroso, J.** (2010). Stereo vision in blind navigation assistance, *World Automation Congress (WAC), 2010*, IEEE, pp.1–6.

[6] **Costa, P.**, **Fernandes, H.**, **Martins, P.**, **Barroso, J. and Hadjileontiadis, L.J.** (2012). Obstacle detection using stereo imaging to assist the navigation of visually impaired people, *Procedia Computer Science*, *14*, 83–93.

[7] **Wang, S. and Tian, Y.** (2012). Camera-based signage detection and recognition for blind persons, *Computers helping people with special needs*, 17–24.

[8] **Lowe, D.G.** (2004). Distinctive image features from scale-invariant keypoints, *International journal of computer vision*, *60*(2), 91–110.

[9] **Schauerte, B.**, **Martinez, M.**, **Constantinescu, A. and Stiefelhagen, R.** (2012). An assistive vision system for the blind that helps find lost things, *Computers Helping People with Special Needs*, 566–572.

[10] **Caraiman, S.**, **Morar, A.**, **Owczarek, M.**, **Burlacu, A.**, **Rzeszotarski, D.**, **Botezatu, N.**, **Herghelegiu, P.**, **Moldoveanu, F.**, **Strumillo, P. and Moldoveanu, A.** (2017). Computer Vision for the Visually Impaired: The Sound of Vision System, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1480–1489.

[11] **Mrovlje, J. and Vrancic, D.** (2008). Distance measuring based on stereoscopic pictures, *9th International PhD workshop on systems and control: young Generation Viewpoint*, volume 2, pp.1–6.

[12] **Neto, E.C.**, **Maya, H.C.**, **Varela, A.T.**, **Carvalho, M.M. and de ALEXANDRIA, A.R.**, (2010). Depth Calculation using Computer Vision and Sift,

Technological Developments in Networking, Education and Automation, Springer, pp.253–256.

[13] **Xiong, G.**, **Li, X.**, **Xi, J.**, **Fowers, S.G. and Chen, H.** (2010). Object distance estimation based on stereo vision and color segmentation with region matching, *International Symposium on Visual Computing*, Springer, pp.368–376.

[14] **Núñez, P.**, **Vázquez-Martín, R. and Bandera, A.** (2011). Visual odometry based on structural matching of local invariant features using stereo camera sensor, *Sensors*, *11*(7), 7262–7284.

[15] **Kamencay, P.**, **Zachariasova, M.**, **Breznan, M.**, **Jarina, R.**, **Hudec, R.**, **Benco, M. and Matuska, S.** (2012). A new approach for disparity map estimation from stereo image sequences using hybrid segmentation algorithm, *International Journal of Modern Engineering Research (IJMER)*, *2*(5), 3201–3206.

[16] **Kim, H.C.**, **Song, D.W.**, **Lee, J.H.**, **Yoo, J.H. and Jung, S.Y.** (2014). Inter-vehicular distance estimation scheme based on image sensor, *2014 International Conference on Electronics, Information and Communications (ICEIC)*, IEEE, pp.1–2.

[17] **Guzel, M.S. and Nattharith, P.** (2014). New technique for distance estimation using sift for mobile robots, *2014 International Electrical Engineering Congress (iEECON)*, IEEE, pp.1–4.

[18] **Kordelas, G.A.**, **Alexiadis, D.S.**, **Daras, P. and Izquierdo, E.** (2015). Enhanced disparity estimation in stereo images, *Image and Vision Computing*, *35*, 31–49.

[19] **Park, K.Y. and Hwang, S.Y.** (2014). Robust range estimation with a monocular camera for vision-based forward collision warning system, *The Scientific World Journal*, *2014*.

[20] **Olofsson, A.**, (2010), Modern stereo correspondence algorithms: investigation and evaluation.

[21] **Nordmann, A.**, (2007), Epipolar geometry, `https://commons.wikimedia.org/wiki/File:Epipolar_geometry.svg`, [Online; accessed April 10, 2019].

[22] **OpenCV Tutorials**, Camera Calibration, `https://docs.opencv.org/3.4.3/dc/dbb/tutorial_py_calibration.html`, [Online; accessed April 10, 2019].

[23] **OpenCV Tutorials**, Introduction to SURF, `https://docs.opencv.org/3.0-beta/doc/py_tutorials/py_feature2d/py_surf_intro/py_surf_intro.html`, [Online; accessed April 10, 2019].

[24] **Bradski, G. and Kaehler, A.** (2008). *Learning OpenCV: Computer vision with the OpenCV library*, " O'Reilly Media, Inc.".

[25] **Redmon, J. and Farhadi, A.** (2018). YOLOv3: An Incremental Improvement, *arXiv*.

[26] **Redmon, J.**, (2013–2016), Darknet: Open Source Neural Networks in C, `http://pjreddie.com/darknet/`.

CURRICULUM VITAE

Name Surname: Mehmet Biberci

Place and Date of Birth: İstanbul, 1989

E-Mail: biberci@itu.edu.tr

EDUCATION:

- **B.Sc.:** 2012, Istanbul Technical University, Department of Computer Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2012-2014 Research Fellow, Artificial Intelligence and Robotics Laboratory, Istanbul Technical University

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Biberci, M.** and Bayazit, U. (2018). Stereo Vision Based Distance Estimation and Logo Recognition for the Visually Impaired. *Proceedings of the 16th international conference on computers helping people with special needs (ICCHP 2018)*.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- Ersen, M., Ozturk, M. D., **Biberci, M.**, Sariel, S., and Yalcin, H. (2014). Scene interpretation for lifelong robot learning. *Proceedings of the 9th international workshop on cognitive robotics (CogRob 2014) held in conjunction with ECAI-2014*.

- Ozturk, M. D., Ersen, M., **Biberci, M.**, Sariel, S., and Yalcin, H. (2014, April). World modeling for cognitive robots. *Signal Processing and Communications Applications Conference (SIU), 2014 22nd* (pp. 1670-1673). IEEE.