ABSTRACT

Title of dissertation: IMAGE AND VIDEO ANALYTICS FOR DOCUMENT PROCESSING AND EVENT RECOGNITION

Sungmin Eum, Doctor of Philosophy, 2017

Dissertation directed by: Dr. David Doermann
University of Maryland Institute for Advanced Computer Studies
and
Professor Joseph F. JaJa
Department of Electrical and Computer Engineering

The proliferation of handheld devices with cameras is among many changes in the past several decades which affected the document image analysis community by providing a far less constrained document imaging experience compared to traditional non-portable flatbed scanners. Although these devices provide more flexibility in capturing, the users now have to consider numerous environmental challenges including 1) a limited field-of-view keeping users from acquiring a high-quality images of large sources in a single frame, 2) Light reflections on glossy surfaces that result in saturated regions, and 3) Crumpled or non-planar documents that cannot be captured effectively from a single pose.

Another change is the application of deep neural networks such as the deep convolutional neural networks (CNNs) for text analysis which is showing unprecedented performance over the classical approaches. Beginning with the success in character

recognition [1], CNNs have shown their strength in many tasks in document analysis as well as computer vision. Researchers have explored potential applicability of CNNs for tasks such as text detection and segmentation, and have been quite successful [2–7]. These networks, trained to perform single tasks, have recently evolved to handle multiple tasks. This introduces several important challenges including imposing multiple tasks on single architecture network and integrating multiple architectures with different tasks. In this dissertation, we make contributions in both of these areas.

First, we propose a novel Graphcut-based document image mosaicking method which seeks to overcome the known limitations of the previous approaches. Our method does not require any prior knowledge of the content of the document images, making it more widely applicable and robust. Information regarding the geometrical disposition between the overlapping images is exploited to minimize the errors at the boundary regions. We incorporate a sharpness measure which induces cut generation in a way that results in the mosaic including the sharpest pixels. Our method is shown to outperform previous methods, both quantitatively and qualitatively.

Second, we address the problem of removing highlight regions caused by the light sources reflecting off glossy surfaces in indoor environments. We devise an efficient method to detect and remove the highlights from the target scene by jointly estimating separate homographies for the target scene and the highlights. Our method is based on the observation that when given two images captured at different viewpoints, the displacement of the target scene is different from that of the highlight regions. We show the effectiveness of our method in removing the high-

light reflections by comparing it with the related state-of-the-art methods. Unlike the previous methods, our method has the ability to handle saturated and relatively large highlights which completely obscure the content underneath.

Third, we address the problem of selecting instances of a planar object in a video or set of images based on an evaluation of its "frontalness". We introduce the idea of "evaluating the frontalness" by computing how close the object's surface normal aligns with the optical axis of a camera. The unique and novel aspect of our method is that unlike previous planar object pose estimation methods, our method does not require a frontal reference image. The intuition is that a true frontal image can be used to reproduce other non-frontal images by perspective projection, while the non-frontal images have limited ability to do so. We show comparing 'frontal' and 'non-frontal' can be extended to compare 'more frontal' and 'less frontal' images. Based on this observation, our method estimates the relative frontalness of an image by exploiting the objective space error. We also propose the use of a K-invariant space to evaluate the frontalness even when the camera intrinsic parameters are unknown (e.g., images/videos from the web). Our method improves the accuracy over a baseline method.

Lastly, we address the problem of integrating multiple deep neural networks (specifically CNNs) with different architectures and different tasks into a unified framework. To demonstrate the end-to-end integration of networks with different tasks and different architecture, we select event recognition and object detection. One of the novel aspects of our approach is that this is the first attempt to exploit the power of deep convolutional neural networks to directly integrate relevant object

information into a unified network to improve event recognition performance. Our architecture allows the sharing of the convolutional layers and a fully connected layer which effectively integrates event recognition with the rigid and non-rigid object detection.

IMAGE AND VIDEO ANALYTICS FOR DOCUMENT PROCESSING AND EVENT RECOGNITION

by

Sungmin Eum

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Joseph F. JaJa, Chair/Advisor
Dr. David Doermann, Co-Advisor
Professor Rama Chellappa
Professor Larry S. Davis
Professor Ramani Duraiswami

Dedication

This dissertation is dedicated to my wife and my parents.

Acknowledgments

The past five years have been a period of intense learning and enlightenment, realizing how small and unwise I am.

First and foremost, I would like to express my sincere gratitude to my research advisor Dr. David Doermann for his unceasing support and guidance throughout the years of my Ph.D research. Apart from his immense knowledge and experience in technically guiding me throughout many research, his patience and kindness made him the best advisor and a mentor for my ph.D study.

Another big thanks to my academic advisor, Professor Joseph JaJa, who has been both passionate and understanding about my progress from the start of my Ph.D years.

I would like to thank my committee members: Professor Rama Chellappa, Professor Larry Davis, and Professor Ramani Duraiswami, for their wholehearted service and insightful comments to make this happen.

I would like to mention my colleagues in LAMP Lab and Computer Vision Lab who "suffered" together to survive in the challenging field of computer vision and machine learning: Hyungtae Lee, Jonghyun Choi, Hongyu Xu, Xianzhi Du, Rajiv Jain, Varun Manjunatha, Jayant Kumar, Kang Le, and Peng Ye.

I would like to send a huge "thank you" to my parents-in-law and my brother-in-law for their trust, support, and heart-warming kindness.

I owe my deepest thanks to my parents and my sister for their unconditional love, timely encouragements, and prayers along my life's journey.

A very special thanks goes out to my wife, Jihyun, who has always been a firm source of support. I cannot imagine how I could have completed the work without her love, prayers, and companionship.

Last but by no means least, I thank the almighty God for His grace and loving kindness.

Table of Contents

List of Figures

Chapter 1: Introduction

In the past several decades, there have been many changes that affected the document image analysis community and the problems it addresses. One is the proliferation of handheld devices including mobile phones with cameras which provide a far less constrained document imaging experience compared to traditional non-portable flatbed scanners. Table 1.1 identifies some of the challenges for cameras and shows a comparison between the scanners and the cameras in terms of various potential constraints. Although this provides more freedom and flexible experiences, the users now have to consider numerous environment settings such as lighting or camera motion, since these may degrade or limit the capture quality. Another change is the application of deep neural networks such as the deep convolutional neural networks (CNNs) for text analysis [1–6] which is not only introducing a new paradigm but also showing unprecedented performance. Since the initial success in character recognition [1], CNNs have been shown to perform well in many tasks in computer vision as well as document analysis. Researchers have explored potential applicability of CNNs for tasks such as text detection and segmentation, and have been quite successful. These networks, trained to perform single tasks, eventually evolved into networks which are targeted to handle multiple tasks. This

Table 1.1: Challenges for cameras and comparison with scanners

	Scanners	**Cameras**	**Challenges for cameras**
Source Size	Limited by device	Potentially unconstrained	Capturing large area with limited resolution
Lighting	Controlled	Uncontrolled	Removing lighting variation
Text Location	On documents	On documents or in scenes	Acquiring best pose
Surfaces	Planar	Planar or non-planar	Flattening or dewarping text

advancement introduces several important challenges such as, 'how to impose multiple tasks on single architecture network' or 'how to integrate different architectures with different tasks'.

In the first part of the dissertation (Chapters 2, 3, and 4), we will address several challenges that arise when capturing documents in unconstrained environments. These challenges include 1) a limited field-of-view keeping users from acquiring a high-quality images of large sources in a single frame, 2) reflections from bright lights on glossy surfaces that result in saturated regions, and 3) crumpled or non-planar documents that cannot be captured effectively from a single pose. In this dissertation, we will explore methods targeting each of these topics.

In our first topic, we address the problem of generating a high-quality document image by employing document image mosaicking. We introduce a novel Graphcut-based document image mosaicking method which seeks to lessen the known artifacts of the previous approaches such as ghosting effects or missing contents. Our method does not require any prior knowledge of the content of the given

document images, making it more widely applicable and robust. Information regarding the geometrical disposition between the overlapping images is exploited to minimize the errors at the boundary regions. Finally, we incorporate a sharpness measure which induces cut generation in a way that results in the mosaic including the sharpest pixels. Our method is shown to outperform previous methods quantitatively in terms of OCR accuracy, qualitatively based on visual appearance.

For our second topic, we address the problem of removing highlight regions caused by the light sources reflecting off glossy surfaces. We specifically target the cases where the highlights are saturated and the original contents are completely obscured. Our method is based on the observation that when two images are captured at different viewpoints, the displacement of the target content is different from that of the highlight regions (Motion parallax). Our method works with two images with slightly different viewpoints using a novel algorithm called, Joint Homography Estimation for Highlight Removal (JH2R) which performs a fast joint estimation of the two homographies, foreground and highlight. We show that our method provides a visually pleasing output with the highlights removed. We also show the effectiveness of our method in removing the highlight reflections by comparing it with the related state-of-the-art methods. Unlike the previous methods, our method has the ability to handle saturated and relatively large highlights which completely obscure the content underneath. Moreover, we stress that our method uses correspondence between the "highlight" regions for better localization of the highlights in multiple images.

For our third topic, we generalize the problem of "flattening" crumpled or

non-planar documents by assuming that each character or region-of-interest on a document is residing on a piecewise planar surface. We address the problem of selecting frames of a planar objects in a video or a set of images by analyzing their "frontalness". We exploit the idea of "evaluating the frontalness" by computing how close the surface normal of an object aligns with the optical axis of a camera. The unique and novel aspect of our method is that unlike previous planar object pose estimation methods, our method does not require a frontal reference image. Our approach was motivated by an observation that a true frontal image can be used to reproduce other non-frontal images by perspective projection, while the non-frontal images have limited ability to do so. We show comparing frontal and non-frontal can be extended to comparing 'more frontal' and 'less frontal' images. Our method estimates the relative frontalness of an image by exploiting the objective space error. We also propose the use of K-invariant space to evaluate the frontalness even when the camera intrinsic parameters are unknown (e.g., images/videos from the web). Our method outperforms a baseline method which uses the homography decomposition approach.

In the second part of this dissertation (Chapter 5), we will address a challenge one may face when trying to make use of multiple deep neural networks (specifically CNNs) with different architectures and different tasks within the same framework.

Here, we focus on the problem of "network integration" which is combining different networks (for different tasks) together in an end-to-end multi-task learning scheme. To demonstrate the integration of networks for "different" tasks, we select event recognition and object detection. Although many previous methods have

showed the importance of considering semantically relevant objects for performing event recognition, yet none of the methods have exploited the power of deep convolutional neural networks to directly integrate relevant object information into a unified network. We present a novel unified deep CNN architecture which integrates architecturally different, yet semantically-related object detection networks to enhance the performance of the event recognition task. Our architecture allows the sharing of the convolutional layers and a fully connected layer which effectively integrates event recognition, rigid object detection and non-rigid object detection.

This dissertation consists of the following chapters. Chapter 2 addresses the problem of mosaicking the document images using Graphcuts, considering sharpness and smooth transition between overlapped images. Chapter 3 describes the method to remove highlight regions on glossy surfaces caused by the light sources by jointly estimating separate homographies for the target scene and the highlights. Chapter 4 describes the method of selecting instances of planar objects in videos or sets of images by applying the concept of "frontalness" evaluation which uses object space error. Chapter 5 introduces a novel unified deep CNN architecture which integrates architecturally different, yet semantically-related networks for different secondary tasks (object detection) to enhance the performance of a primary task (event recognition). We conclude with future work and open questions, as well as a summary of theoretical contributions in Chapter 6.

Chapter 2: Sharpness-aware Document Image Mosaicking Using Graph-cuts

2.1 Introduction

In the field of document image analysis, document image mosaicing has received a great deal of attention as mobile devices with low cost built-in cameras are used to image printed materials. The idea of acquiring a single, high-quality, digital copy of a document from multiple overlapping shots has become very attractive, especially for documents which are difficult to scan or capture in a single pass. Some examples of such documents are shown in Figure 1 including long receipts, posters on display, and framed documents.

Numerous approaches were introduced which address the general issue of image mosaicking and many more now are even built into popular commercial software applications [12]. Although they seem to perform well on natural scene images, they typically show unsatisfactory results on document images. Unlike scene image mosaics where discontinuities are less noticeable, document images show noticeable errors because most of the content is small and very high contrast.

Figure 2 depicts examples of document image mosaics using two state-of-

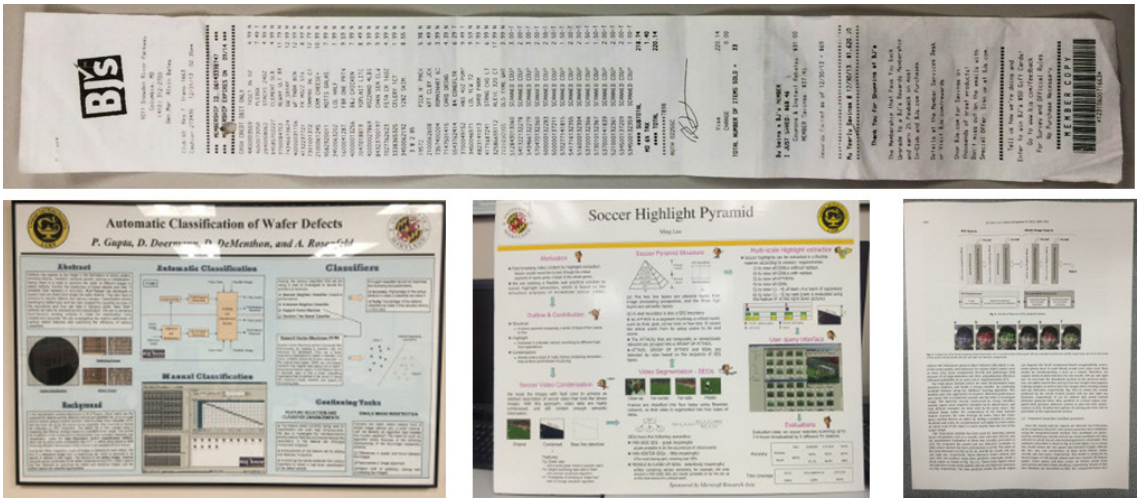
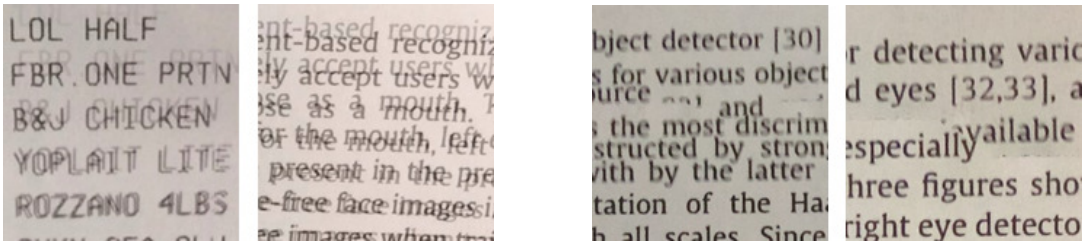


Figure 2.1: Documents which may require document image mosaicing

the-art scene image mosaicing approaches: AutoStitch [12] and iPhone5s built-in panorama. Figure 2(a) shows regions where the same texts appear twice with a slight offset. This is typically referred to as "ghosting". Figure 2(b) illustrates another erroneous result where contents are missing in the mosaic. Such artifacts are caused by two major components of a general image mosaicing process: image registration and image blending. The registration attempts to properly align the overlapping images, while image blending is responsible for compositing the images as naturally as possible.

Previous work can broadly be categorized into two groups based on which major component (registration or blending) they address. Most of the approaches [13–19] focus on enhancing the registration process. In early approaches [13,14], registration between overlapping images were estimated using methods such as image pyramid, image correlation or Least Median of Squares. These approaches target planar registration, typical of scanned documents. In [15], a sliding window reg-



(a) AutoStitch (b) iPhone5s

Figure 2.2: Document image mosaics with errors (a) AutoStitch showing ghosting artifacts (b) iPhone5s built-in panorama with missing contents

istration method was introduced, but is time consuming and only applicable for binary images. Kasar et. al. [16, 17] began using feature descriptor-based registration methods. In [16], the Harris corner detector and the discrete cosine transform feature descriptors were exploited, while [17] employed angular radial transform for the description of each connected component for registration. As mobile devices became more popular, a mobile-based, user-interactive mosaicing scheme [20] was introduced which incorporated SIFT features and RANSAC-based homography estimation. Most recently, two methods [18, 19] were proposed which focus on compensating for perspective distortion of the overlapping portions of documents.

We note that, most of the approaches addressing the registration problem [13, 15–19], adhere to using the conventional alpha-blending (weighted averaging). Although it is not explicitly stated in [13,15,18,19], we presume that they have used alpha-blending by carefully inspecting their experimental results.

Instead of focusing on the registration problem, Liang et al. [21, 22] addressed the blending problem by using "selective" image blending. The method was devel-

oped to handle text content, and thus performs binary morphology and word-level segmentation. It is likely to perform poorly when dealing with complex figures, tables or text with different sizes. Even if the given document image includes uniformly sized characters, words might appear jagged in the mosaicked image.

In this paper, we address the limitations introduced in the previous approaches by using a sharpness-aware document mosaicing based on Graphcuts performed at the pixel level. The contributions of our method are as follows. First, Graphcut-based blending method is a novel method which effectively stitches two overlapping images without requiring any prior knowledge of the document, thus being more robust and widely applicable. Second, boundary constraints are imposed which minimize discrepancy between overlapping and non-overlapping regions. Third, we incorporate a sharpness measure which promotes cuts which favor a mosaiced image with sharper pixels when blending the overlapping images.

2.2 Overall Document Mosaicing Approach

Although the novelty of our method is primarily in the image blending step, we briefly summarize the overall framework for completeness, and additional detail can be found in [20].

The mosaicing process begins with the capture of a portion of the document with a user interactive approach. Motion of the mobile device is estimated in real-time and the user is notified when to move and when to stop while scanning. The result is a series of images suitable for mosaicing.

Once the images are captured, scale and rotation invariant SIFT [23] features are extracted and matched. Matched features are then used to estimate the homography, or perspective projection, between pairs of overlapping images. Since there may exist outliers in the feature correspondences, we employ a robust homography estimation which efficiently eliminates the outliers through RANSAC [24] followed by a Levenberg-Marquardt refinement scheme.

Finally, we project the images onto a reference coordinate system, or a plane, followed by blending the images together where they overlap to generate the mosaiced result. A detailed description of the proposed Graphcut-based blending process is included in the next section.

2.3 Graphcut-based blending

As mentioned previously, the phenomenon of the same content appearing twice with a slight offset, referred to as the ghosting artifact, is caused by alpha-blending (weighted averaging) the two overlapping images when the homography estimation contains errors. It is very difficult to have zero error in the homography estimation throughout the overlapping region. Thus, the proposed method seeks to eliminate such ghosting artifacts by using a Graphcut-based blending scheme which performs well even when slight registration error exists. The proposed method also has advantages over the selective image blending [21, 22] in that it does not require any segmentation.

Our method is capable of acquiring a cut line where two overlapping images

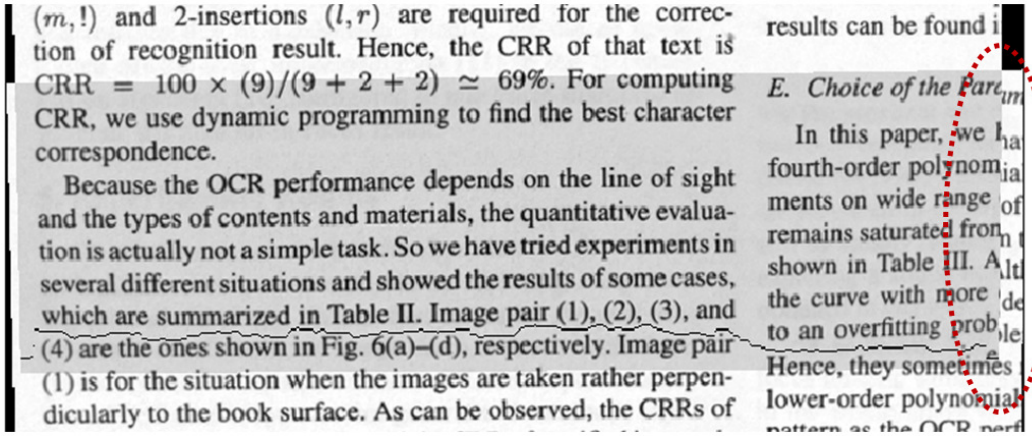


Figure 2.3: dynamic programming based horizontal cut blending

can be stitched together like two pieces of puzzle. Since documents tend to have empty space where text or other contents like figures or tables are not present, it is desirable for the cut to be generated in those empty regions as shown in Figure 3.

2.3.1 Boundary Constraints

This problem can be viewed as a 2-class labeling problem where each of the labels indicates which of the two images, pixels are being copied from. The energy function E which is being minimized in solving this labeling problem is represented by the sum of two terms: a smoothness term $\sum V_{p,q}$ and a data term $\sum D_p$, as shown in (1). The objective is to find a labeling f that labels each pixel $p \in P$ as f_p.

$$E(f) = \sum_{\{p,q\} \in N} V_{p,q}(f_p, f_q) + \sum_{p \in P} D_p(f_p) \tag{2.1}$$

The smoothness term is the sum of the penalty $V_{p,q}$ for all the pairs (p,q) included in N, where N, f_p, f_q, indicate the set of neighboring pairs of pixels, label

of pixel p, and label of pixel q, respectively. It can be described as the penalty imposed on the edge between pixel p and q, whenever a cut is being made. The data term is the sum of the penalty D_p for all the pixels in P. D_p measures the penalty imposed on pixel p when p is labeled as f_p. A detailed explanation of the energy function and the Graphcut algorithm can be found in [25–27].

In our method, we use the following equation [28] as the smoothness term, which is defined for edges between every pair of neighboring pixels in the overlapping region

$$V_{p,q}(p, q, A, B) = |A(p) - B(p)| + |A(q) - B(q)|, \tag{2.2}$$

where p and q are the neighboring pixel locations in images A and B.

For the data term in (1), we have incorporated two different terms, a boundary constraint and a sharpness measure, to guide the cut to minimize the discrepancy where the overlapped regions meet with the non-overlapping regions, and to favor the sharper image.

Our initial idea was to simply acquire a horizontal cut by using a method in [29]. This approach performs well in generating a seamless mosaic near the cut. However, a considerable number of discrepancies appear as shown in the dotted circular region of Figure 3.

In order to mosaic the two images with minimum discrepancies where the overlapped and non-overlapped regions meet, we have employed hard-constraints to constrain which image the boundary pixels are copied from. We have adaptively applied one of six different hard-constraints determined by the geometrical disposi-

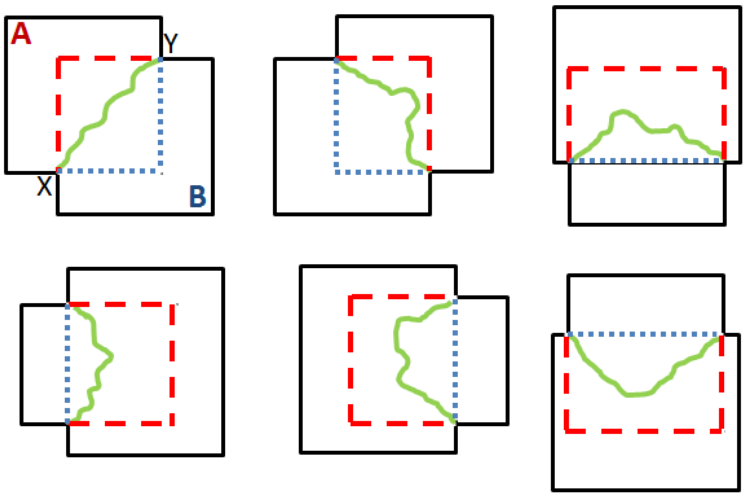


Figure 2.4: The six hard-constraints used for the data term commonly encountered for horizontal or vertical scanning

tion of the two overlapping images shown in Figure 4. This can also be viewed as designating the locations of the two end points, X and Y, of the cut being made within the overlapping region.

All the pixels located on the red dashed boundary line in Figure 4, are copied from image A, by setting $D_p(\mathrm{A}) = 0$, $D_p(\mathrm{B}) = \infty$. In the same way, pixels on the blue dotted boundary will be set with the data terms of $D_p(\mathrm{A}) = \infty$, $D_p(\mathrm{B}) = 0$.

2.3.2 Incorporating sharpness

The sharpness measure is also incorporated into the data term of the energy function. This, in turn, penalizes the blurred pixels with higher cost and the sharper pixels with lower cost when computing the energy function in (1).

The sharpness measure is computed for every pixel location within the overlapping region of the two images using a method introduced in [30] which is designed to

estimate sharpness for documents or scenes. For the proposed method, the penalty value, D_p(A) or D_p(B) for each of the pixels in the overlapping region is controlled by the difference of the sharpness of the two images as shown in (3). γ_{pA} and γ_{pB} are the sharpness value of image A and B, respectively, computed at the pixel location p.

$$\begin{aligned} &\text{Let,} \delta = |\gamma_{pA} - \gamma_{pB}| \\ &\text{If} \gamma_{pA} \geq \gamma_{pB}, \text{then } D_p(A) = -\delta \text{ and } D_p(B) = \delta \\ &\text{Else,} D_p(A) = \delta \text{ and } D_p(B) = -\delta \end{aligned} \tag{2.3}$$

Thus, the Graphcut favors the pixels with higher sharpness, which guides the cut so that regions with sharper pixels are included in the final mosaiced image.

2.4 Experimental Evaluation

2.4.1 Dataset

To the best of our knowledge, there are no publicly available datasets for document image mosaicing. Thus, we have constructed a dataset where each session is comprised of two partially overlapping shots of a document using the camera on the iPhone5s. The images were captured with the resolution of 3264(w) x 2448(h) in a reasonably lit, indoor environment.

Ten different documents were selected so that the method could be tested on not only the text lines but also on other types of frequently appearing contents such as equations, graphs, pictures, and tables. For each document, 6 sessions were captured, for a total of 60 sessions. The images in a session may have no blur or

Table 2.1: OCR Performance Comparison

	Alpha-blend	Selective blend	Proposed
character	72.31%	80.90%	**83.70%**
word	62.25 %	71.98 %	**77.25%**

blur in one of the two images. Note that the blur is added to the dataset for the purpose of verifying the performance of sharpness-aware approach.

2.4.2 Performance Comparison

Our experiments compare alpha-blending and selective blending to our method using our dataset. As the target objects for the mosaicing are documents which typically include text contents, OCR performance was used as a measure for quantitative performance comparison. Character and word level OCR accuracy were obtained using the OCR Frontiers Toolkit [31]. Table 1 shows that our method significantly outperforms the previous methods in both character and word-level OCR accuracy.

Figure 5 shows the resulting mosaics of two documents generated by three different blending approaches.The gray regions indicate the overlaps between pairs of images. Observe that the ghosting artifacts clearly occur when using the alpha-blending as depicted in Figure 5(a) and (d). Meanwhile, the selective blending approach generates several different types of artifacts due to its binary morphology based procedures which incorporate dilation, thresholding and connected component labeling. In result, the mosaic shows unwanted fragments of contents as seen in

Figure 5(b).

Moreover, neither of the previous approaches demonstrate the smooth transition between the overlapping and the non-overlapping regions, thus generating text or figures on the boundary with improper alignment. Such phenomenon can also be seen in Figure 5(b). The selective blending may even lose some of the contents which reside on the boundary. Notice that almost an entire text line is missing in Figure 5(e), while the same portion is properly recovered in Figure 5(f).

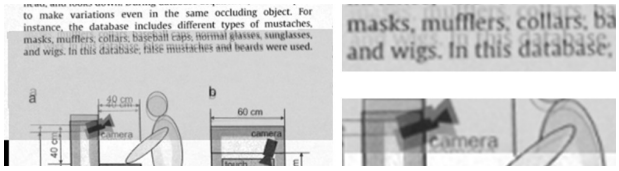
2.4.3 Limitations

Although our method outperforms the previous approaches, Graphcut-based blending does not address registration errors. In other words, if the registration error is considerably large, the resulting mosaicked image may contain duplicate contents as shown in Figure 6. The red crosses and blue dots indicate the corresponding feature points.

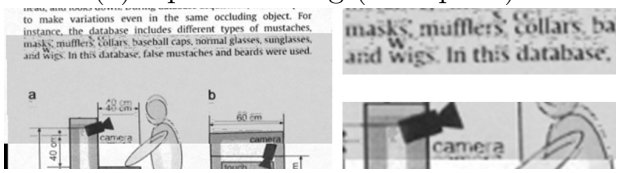
Note that in an ideal case, only one of the two matching features should appear. However, such problems arise when the cut runs between the corresponding feature points. The relative positions of the cut and the matching features could be used do an automatic check on the mosaicing quality.

2.5 Summary

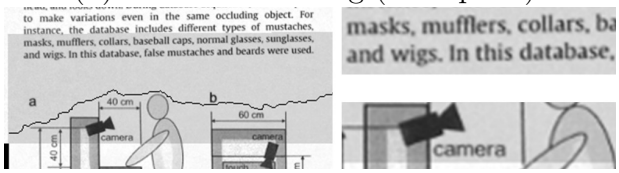
In this work, we have proposed a novel method for document image mosaicing based on Graphcuts. We have focused on comparing the proposed blending approach



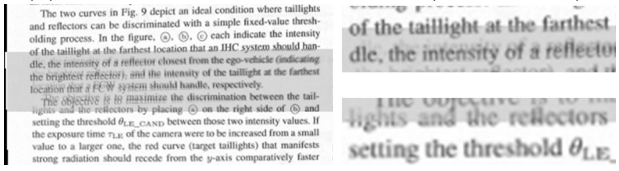
(a) alpha-blending (example 1)



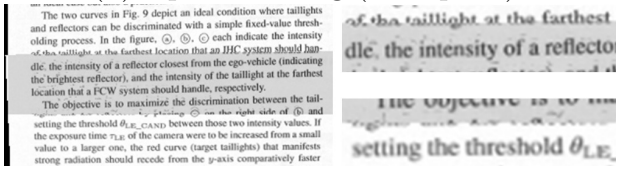
(b) selective blending (example 1)



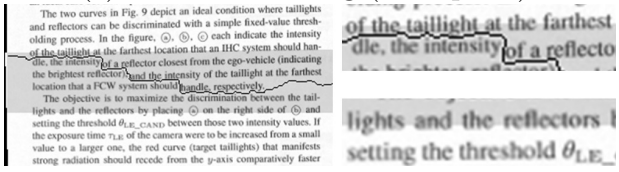
(c) Graphcut-based blending (example 1)



(d) alpha-blending (example 2)



(e)selective blending (example 2)



(f) Graphcut-based blending (example 2)

Figure 2.5: Resulting mosaic documents using (a),(d) alpha blending, (b),(e) selective blending, (c),(f) proposed

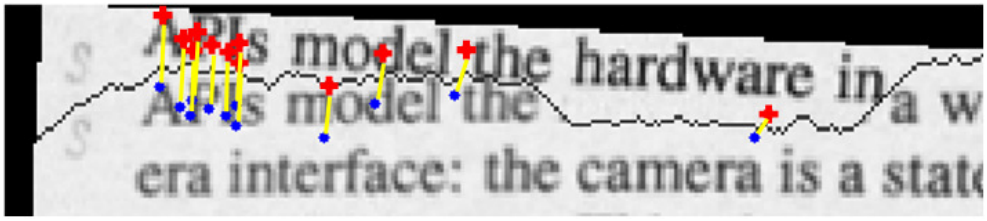


Figure 2.6: A failure case with duplicate content

with the two blending approaches used in previous literature. We have verified that our approach outperforms other approaches both qualitatively and quantitatively, showing its advantage in eliminating the ghosting effects and being capable of handling various types contents other than text. In the future, it will be worthwhile to devise a method which incorporates the registration error along with the matching correspondence information when running the Graphcut so that the result could effectively avoid having duplicate contents in possible erroneous cases.

Chapter 3: Joint Homography Estimation for Highlight Removal

3.1 Introduction

Imagine being in an art museum or any other indoor environment where there are numerous paintings, pictures, documents or posters held inside glass-frames for protection. There are pieces which you wish to capture using a camera, but you experience difficulty avoiding highlights which are generated by bright indoor lighting reflected off the glossy surfaces. Similar problems occur when trying to capture contents off of whiteboards, documents printed on glossy surfaces or objects such as books or CDs with plastic covers. Figure 3.1a illustrates typical examples.

In this work, we address the problem of removing unwanted highlight regions in images generated by reflections of light sources on glossy surfaces. Although there have been efforts made to synthetically fill in the missing regions using the neighboring patterns by applying methods like inpainting [32,33], it is impossible to recover the actual missing information in completely saturated regions. Therefore, it is prudent to consider using multiple images where corresponding regions are not covered by the saturated highlights.

We make the following observations in devising our approach:

(a) (b)

Figure 3.1: (a) Examples of highlights shown on the glossy surfaces obscuring the desired content and degrading visual quality (b) Result (right) obtained using our algorithm to remove the highlights using two images (left and middle) captured at different viewpoints

- The distance between the camera and the virtual location of the light source is typically larger than the distance between the camera and the target content. (Figure 3.2). Thus, it is reasonable to use two separate homographies in distinguishing the objects at different distances. [34]

- When two images are captured with a change of view point, the displacement of the desired content is different from the displacement of the highlight regions. This is referred to as 'motion parallax'.

Our method works with two images with slightly different viewpoints and applies a novel algorithm called, Joint Homography Estimation for Highlight Removal (JH2R) which performs a fast joint estimation of the two homographies, foreground and highlight, and provides a visually pleasing output with the highlights removed. (Figure 3.1b)

To the best of our knowledge, no previous work has addressed an approach which can successfully handle relatively large and saturated highlight regions ob-

scuring the content underneath. We show the effectiveness of our approach by comparing it with closely related state-of-the-art methods.

3.2 Related Work

Several methods have been suggested to explicitly address highlight issues based on the dichromatic reflection model [35]. Tan et al. [36] uses a user-assisted inpainting and show that highlight pixels contain useful information for highlight removal. Similarly, [37] asserts that the color texture data lying outside the highlights can assist in filling in the missing diffuse surface colors inside the highlights. Yang et al. [9, 38] introduced a method which propagates the diffuse color information into the highlight regions using an iterative bilateral filter. Tan et al. [39] proposed a local operation based method which does not require explicit color segmentation. They strongly assume that surface color is chromatic and ignores cases with saturated regions.

Solutions based on reflection removal or layer separation can also be taken into consideration. Some suggest that it is possible to solve this ill-posed problem using a single image supported by additional priors. Levin et al. [40] showed that layer decomposition can be performed by minimizing the total number of edges and corners. In [41], the prior information for layer separation is strengthened by bringing the user into the loop for manual gradient labeling. Li and Brown [8] recently suggested an approach which assumes that one layer is smoother than the other. Since all of these methods use only one image as input, it is virtually

impossible to recover the content obscured by the highly-saturated or large highlights unless the region is homogeneous and smooth.

Numerous approaches to exploit multiple images have also been explored. Some approaches have used the polarizing effect on specularities [42–45] while others have used focus [46] or flash [47] as priors. However, using polarizers, different focuses or flash may require use of additional hardware which is not always feasible or convenient for typical users.

Techniques using multiple images with different viewpoints have also proven effective. Szeliski et al. [48] showed that relative motion between the layers in multiple images can be used effectively. In [10], gradients across the aligned image set are used to distinguish pixels in different layers. Lin et al. [49, 50] integrated color analysis and multi-baseline stereo. This, however, requires large set (>50) of images captured by moving the camera along a linear path with constant velocity. The approach also suffers when images contain color saturations. Recently, Guo et al. [11] showed that by harnessing correlation, sparsity, and the independence prior, reflection separation can be performed.

These methods share a similar perspective with our approach in that they use multiple viewpoints and incorporate the relative motion difference in different layers. However, our method does not employ any sophisticated optimization which usually requires significant processing time [10,11], nor does it require any user intervention [11]. Most importantly, unlike others, our method uses the relationship between the highlight regions resulting in more robust removal of saturated highlights. A detailed comparison is presented in the experiments section where our method is

shown to outperform the representative state-of-the-art.

3.3 Our Method

3.3.1 Overview

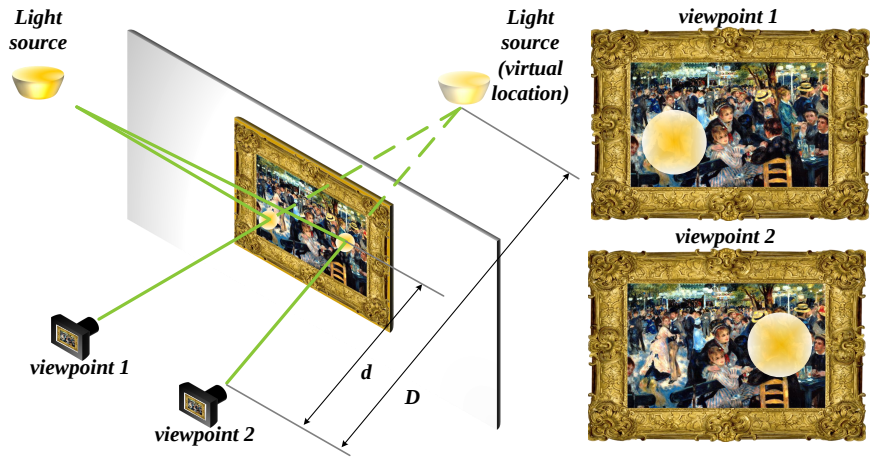


Figure 3.2: The illustration depicts the overhead view of the camera, the desired content, and the light source.

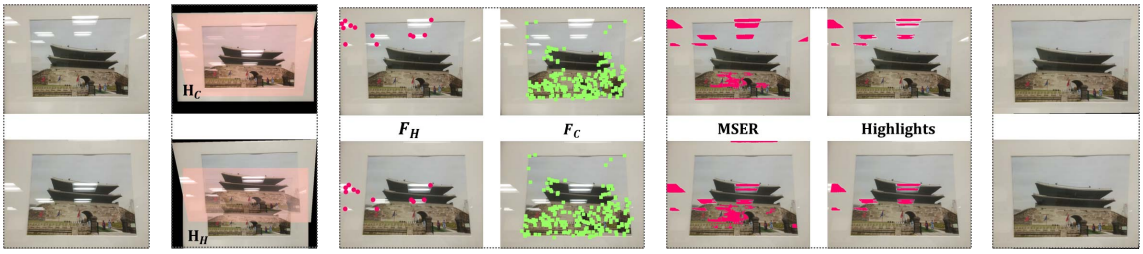
Our method was motivated by a widely acknowledged physical phenomenon known as 'motion parallax'. Motion parallax states that as the viewer moves, the movement of the objects in the vicinity is greater across the field of view than those in the distance. A driver can easily observe that the objects close to the window (e.g., roadside traffic signs) pass by quickly while those in the distance (e.g., clouds) remain in one's field of view longer.

Without loss of generality, we can view the relationship between the desired content (e.g., a painting) and the highlights as shown in Figure 3.2. Since the

highlights caused by the light source are the result of the reflection on the glossy surface before they reach the camera, the light source can be modeled to virtually exist on the other side of the content. Note that the distances of the two sources (target content and light) from the camera are different. Unless the light source is attached on the same wall as the painting, in which case no reflection would exist, the distances can never be the same. In fact, the distance from the light source is always larger than the distance from the content ($D > d$, in Figure 3.2).

In order to distinguish the movements of the highlights, we need at least two images captured in different views. We detect where the highlights are by searching for the two separate homography matrices: one for the content ($\mathbf{H}_C$) and the other for the highlights ($\mathbf{H}_H$). Applying two different homographies for scenes at different distances proved to be effective by Gao et al. in [34]. We exploit the fact that the homography ($\mathbf{H}_C$) which can properly overlay the desired contents in the two images will generate an erroneous overlap between the corresponding highlight regions. Similarly, the desired contents will display incorrect overlap when $\mathbf{H}_H$ is employed. This is shown in the second step of Figure 3.3b.

Unlike the intrinsic layer separation problem, removing the saturated highlights from images requires another image which can provide the corresponding non-highlight pixels. To perform such "pixel-transfer", it is necessary to have the pixel-level detection results of the highlights. In our approach, we first detect the highlight regions at the feature level by jointly estimating the two homographies using the proposed JH2R algorithm. Then $\mathbf{H}_H$ is used to estimate the highlight regions at the pixel-level. Finally, we remove the highlights in both of the images by



(a) (b) (c) (d) (e)

Figure 3.3: Schematic overview of our method (a) Input images (b) Joint homography estimation (c) Feature-level labeling (d) Pixel-level labeling (e) Final results

transferring the corresponding pixels from the complementary image using Poisson blending [51]. Figure 3.3 shows the schematic overview of our method. Details on each steps of the algorithm are explained in the following subsections.

3.3.2 Joint homography estimation and highlight feature labeling

In our approach, we attempt to estimate the two different homographies. We devise a novel, yet efficient algorithm which only requires feature correspondences between the two images along with Maximally Stable Extremal Region (MSER) [52] features for those images as input. Although we have utilized the SIFT [23] features in our implementation, any type of feature extractor and descriptor can be used as long as the features can be stably matched throughout the image including the highlight regions. Before triggering our algorithm, a set of all the feature correspondences (F) is acquired by thresholding the Euclidean distances between tentative feature pairs as described in [23]. Our algorithm is shown in Algorithm 1. Note that

F and M represent the set of all feature correspondences and the set of all MSER features, respectively. The framework of our algorithm was inspired by the Random Sample Consensus algorithm [24].

Algorithm 1: Joint homography estimation for highlight removal (JH2R)

Input : F, M

Output: $\mathbf{H}_C, \mathbf{H}_H, F_C, F_H$

k $\leftarrow$ 1 `/* iteration index */`

repeat

  Randomly select 4 correspondences $\in F$, Compute $\mathbf{H}_C$

  for $\forall F_i \in F$ **do**

    if $e(F_i, \mathbf{H}_C) > T$ **then**

      $F_O \leftarrow F_O \cup F_i$

    end

  end

  Randomly select 4 correspondences $\in F_O$, Compute $\mathbf{H}_H$

  for $\forall F_i \in F$ **do**

    if $e(F_i, \mathbf{H}_C) \leq T$ & $e(F_i, \mathbf{H}_C) \leq e(F_i, \mathbf{H}_H)$ **then**

      $F_C \leftarrow F_C \cup F_i$

    end

    if $e(F_i, \mathbf{H}_H) \leq T$& $e(F_i, \mathbf{H}_C) > e(F_i, \mathbf{H}_H)$ & $F_i \in M$ **then**

      $F_H \leftarrow F_H \cup F_i$

    end

  end

  Compute J_{curr} (Eqn. 3.2)

  if J_{curr} ¡ J **then**

    $J \leftarrow J_{curr}$ and update $\mathbf{H}_C, \mathbf{H}_H, F_C, F_H$

  end

  k $\leftarrow$ k + 1

  compute and update N (Eqn. 3.5)

until $k < N$

Our algorithm begins by estimating the homography for the content ($\mathbf{H}_C$)

using four randomly selected feature correspondences from F. Using $\mathbf{H}_C$, we temporarily label all the feature correspondences in F as either the content feature F_C or the outlier feature F_O by thresholding (T) their symmetric transfer errors [53]. The threshold T is empirically acquired. For estimating the symmetric transfer error of a feature correspondence F_i, we consider both the forward ($\mathbf{H}_C$) and backward ($\mathbf{H}_C^{-1}$) transformations and use them to compute the sum of geometric errors as follows:

$$e(F_i, \mathbf{H}_C) = d(x_i, \mathbf{H}^{-1}x_i')^2 + d(x_i', \mathbf{H}x_i)^2, \tag{3.1}$$

where x_i and x_i' are the corresponding feature points in F_i, while $d(p, q)$ represents the Euclidean distance between the inhomogeneous points p and q.

At this point, we assume that the set of outlier correspondences, F_O, should include the highlight feature correspondences since they do not follow the homography for the desired content ($\mathbf{H}_C$). Based on that, a second random sampling from set F_O is carried out to compute the homography for the highlights ($\mathbf{H}_H$). The results for the joint homography estimation is depicted in Figure 3.3b.

Once both of $\mathbf{H}_C$ and $\mathbf{H}_H$ are estimated, all the feature correspondences are relabeled into three different mutually exclusive sets: F_C, F_H and F_O. Figure 3.3c depicts a sample result of the feature-level labeling step. If a feature correspondence F_i is not labeled as either desired content or highlight, it is labeled as an outlier. In order for a correspondence F_i to be categorized into the desired content correspondence set (F_C), the symmetric transfer error using $\mathbf{H}_C$ (i.e., $e(F_i, \mathbf{H}_C)$) should be smaller than the threshold T. At the same time, the error using $\mathbf{H}_C$ has to be smaller than the error using $\mathbf{H}_H$, which indicates that F_i favors $\mathbf{H}_C$ over $\mathbf{H}_H$. If F_i does not get categorized into F_C, the algorithm checks if it can be categorized as one of the highlights by evaluating the symmetric transfer error using the highlight homography ($\mathbf{H}_H$) in a similar manner.

One additional criterion is employed for F_i to be categorized into F_H. It constrains the features in F_i to be present on the "bright-on-dark" MSERs [52]. The "bright-on-dark" MSER regions indicate the MSER regions which are brighter than the vicinity. As the intensity values in highlight regions tend to be stable and lighter than the neighboring regions, MSER is a reasonable choice for obtaining potential highlight regions. Yet, MSER also detects some other non-highlight regions as shown in Figure 3.3d which will be eliminated by the pixel-level labeling and the

blending scheme in Section 3.3.3.

Having obtained the labeling for all the feature correspondences along with the two homographies, the cost J for the current iteration is computed as

$$J \ = E(F_C, \mathbf{H}_C) + E(F_H, \mathbf{H}_H) \ - \gamma\left(\frac{n(F_C) + n(F_H)}{n(F)}\right). \tag{3.2}$$

The first and the second term incorporate the symmetric transfer error while the third incorporates the number of inlier (desired content and highlights) feature correspondences. γ is a parameter which balances the three terms. $n(F_C)$ and $n(F_H)$ indicates the number of feature correspondences in each of the sets, respectively, while n_{tot} represents the total including the outliers. The first term which measures the average symmetric transfer error for the set (F_C) is computed using Equation 3.3. The second term is computed in the same manner.

$$E(F_C, \mathbf{H}_C) = \sum_{F_i \in F_C} \frac{e(F_i, \mathbf{H}_C)}{n(F_C)} \tag{3.3}$$

If the cost for the current iteration is smaller than the best previous case, the two homographies along with the two feature correspondence sets are updated. This process is repeated until the termination criteria are met.

Termination criteria We determine a maximum iteration number N adaptively after every iteration. We define w_C as the probability that any correspondence randomly selected from F is included in F_C. We assume that w_H is the probability that any correspondence randomly selected from $F - F_C$ is included in F_H. These probabilities can be iteratively updated at the end of each iteration as $w_C = n(F_C)/n(F)$ and $w_H = n(F_H)/(n(F) - n(F_C))$. In Equation 3.4, p is defined as the probability that 4 randomly selected samples are from F_C in the first selection and 4 randomly selected correspondences are from F_H in the second selection within N interations, at least once.

$$p = \ 1 - ((1 - w_C^4) + w_C^4(1 - w_H^4))^N = \ 1 - (1 - w_C^4 w_H^4)^N \tag{3.4}$$

Here, $\left(1 - w_C^4\right)$ is the probability that all 4 correspondences in the first selection are not from F_C. $w_C^4\left(1 - w_H^4\right)$ indicates the probability that the 4 correspondences from the first selection are from F_C but at least one sample from the second selection

are from the outlier set. Therefore, the adaptive maximum iteration number N can be derived from equation 3.4 as

$$N = \frac{log(1-p)}{log(1-w_C^4 w_H^4)}. \tag{3.5}$$

3.3.3 Pixel-level highlight detection and blending

Using the JH2R algorithm, the two homographies (Figure 3.3b) along with the two feature correspondence sets (Figure 3.3c) for the desired content and the highlight regions can be acquired. However, the feature-level detection of the highlight regions is insufficient to properly eliminate the highlights. Instead, it needs to be extended up to the pixel-level so that the non-highlight pixels can be transferred complimentarily to recover the obscured contents.

We make use of two previously acquired results which make this step computationally efficient: the estimated homographies ($\mathbf{H}_C$, $\mathbf{H}_H$) and the MSER detection. Pink regions in the left column of Figure 3.3d depict the MSER detection result. Then the homography $\mathbf{H}_H$ is used to warp the two MSER images onto a common plane. This overlays the highlight regions from one image onto the corresponding highlight regions on the other. Thus, the intersection between the two MSER images, when projected onto the same plane using $\mathbf{H}_H$, should be the estimated region for the highlights in pixel sense. The right column of Figure 3.3d shows the final highlight detection result. Note that we are assuming that the two images both contain the highlights which we wish to eliminate.

Given the pixel-level highlight regions in both of the images, $\mathbf{H}_C$ is used to project the two images onto a common plane so that the desired contents are overlaid properly while the highlight regions do not overlap. In other words, highlight regions in one image are layed over the non-highlight regions in the other image. This enables us to easily recover missing information for all the highlight regions in both of the images. Lastly, Poisson blending [51] is applied to assist the pixel transfers at the highlight regions with smooth boundaries. Figure 3.3e shows the sample result with all the highlights eliminated with visually pleasing quality.

3.4 Experimental Evaluation

(a) Input (b) Ours (c) Li et al. (d) Yang et al. (e) Li et al. (f) Guo et al.

Figure 3.4: Five examples of highlight removal results using (b) our method compared with those produced by (c) Li et al. [8], (d) Yang et al. [9], (e) Li et al. [10], (f) Guo et al. [11]

Our method is implemented in Matlab and run on Intel Core i5 PC (2.6GHz CPU, 4GB RAM). All the data used in the experiments are captured in real world scenes under different indoor lighting conditions. Each input image set contains two images with two different viewpoints.

Comparison with state-of-the-art We have compared our method with four state-of-the-art algorithms [8–11]. They were chosen to represent three different approaches to solving the given problem : 1) highlight removal [9], 2) single image-based reflection removal [8], and 3) multiple image-based reflection removal [10, 11]. We have used the implementations provided by the authors using author-recommended parameters. Since [8] and [9] only use a single image, we have used only one of the two images per set as input.

Figure 3.4 shows five sample results of real world images. As can be observed in Figure 3.4c and 3.4d, both [8] and [9] are incapable of removing the highlights due to the lack of information within the saturated regions. Li et al. [8] fails to

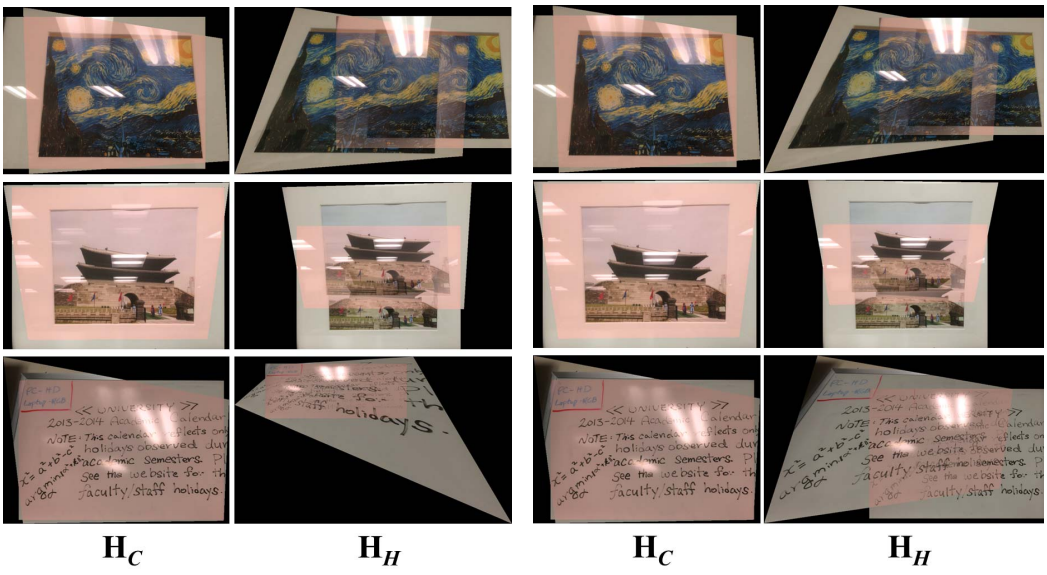
obtain a sufficient amount of gradient information which they use to separate the reflection layer. Yang et al. [9] also suffers since the saturated highlights are void of diffuse color information which is supposed to change smoothly from outside the highlights to the inside.

Multiple-image based approaches by [10, 11] produce results where the highlights are only partially removed. In [10], gradients with variation across the aligned images are assumed to belong to the reflected scenes while constant gradients are assumed to belong to the desired scene. Thus, when the gradients on the highlights are too weak to be distinguished from the underlying smooth texture, this approach may suffer as shown in Figure 3.4e. While [11] uses several priors including the independence between the desired content and the reflection to separate the two layers, none of the priors explain the inherent characteristics of highlights. Thus, in most cases (Figure 3.4f), color components were falsely categorized into the reflection layers, generating unnaturally colored results.

Our method, unlike others, specifically uses the relationship between the highlight regions resulting in more precise detection and removal. One may observe from Figure 3.4 that our method can also handle dim highlights as there still exist geometrical distinction between desired contents and dim highlights in terms of homography. In overall, our method produces the most visually pleasing results.

Homography estimation evaluation In Figure 3.5, we show the efficacy of JH2R by comparing the warped images using the estimated homographies with those using the groundtruth. The estimated $\mathbf{H}_C$ for the desired content are very accurate. Although the estimated $\mathbf{H}_H$ may not be equivalent to the groundtruth as illustrated in the third example, notice that the highlight regions are still well aligned. As long as the highlights overlap properly, pixel-level labeling can be performed. The groundtruth homographies are computed using manually labeled correspondences for content and highlights, separately.

Processing time Our method spends 25.3 seconds on average which is much faster than Li et al. [10] and Guo et al. [11] by almost the order of magnitude as shown in Table 3.1. Although Li et al. [8] and Yang et al. [9] both spend less processing time compared to ours, their performance in removing the highlights are unsatisfactory. We have used a single image ([8, 9]) or a pair of images (ours, [10, 11]) according to each methodology. The size of the images used in the experiments is 640 x 480.



(a) (b)

Figure 3.5: (a) Estimated homographies compared with the (b) groundtruth. These estimated homographies are used to generate the results in the top three rows of Figure 3.4b. Overlapped regions between the pairs are shaded in red.

In Figure 3.6, we show more results produced by our method including a failure case. The red arrow indicates the region which is obscured by the highlights in both of the input images which leaves no information to recover from. This violates our assumption that the highlights in the input images should not cover the same content. However, this assumption is known to be reasonable when targeting saturated regions as stated in [49, 50], and such cases can easily be avoided with user cooperation.

3.5 Summary

In this work, we have devised an efficient method for removing highlights reflected off glossy surfaces of the target scene generated by bright sources. Our algorithm jointly estimates the two representative homographies for the target scene and the highlights to effectively detect and remove the highlights. Unlike some of

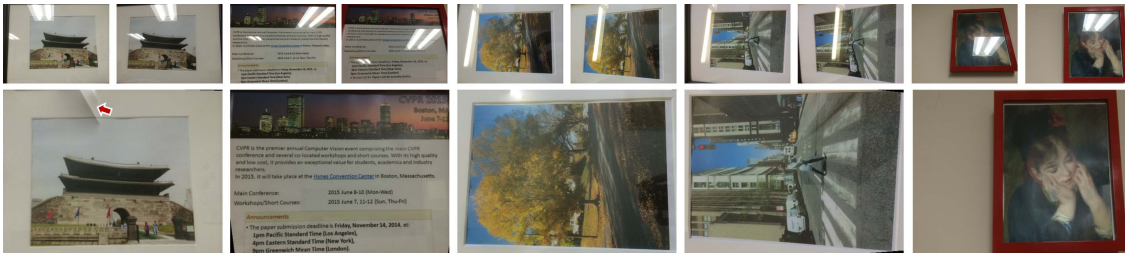


Figure 3.6: More highlight removal results produced by our method. Red arrow indicates a failure case.

Method	**Num of Imgs**	**Processing Time**
Ours	2	25.3 s
Li et al. [8]	1	24.5 s
Yang et al. [9]	1	$< 1s$
Li et al. [10]	2	$221.7s$
Guo et al. [11]	2	$260.2s$

Table 3.1: Quantitative processing time comparison with previous methods

the previous approaches that use homography between non-highlight regions, we newly use correspondences between "highlight" regions for better localization.

We have verified that our approach outperforms closely-related approaches, showing its state-of-the-art quality in handling highly saturated highlights which obscure the underlying content. It requires fewer constraints in image acquisition and is faster than any other multi-view methods [54]. It will be worthwhile to further investigate an automatic capture scheme which can smartly overcome the challenging scenarios.

Chapter 4: Content Selection Using Frontalness Evaluation of Multiple Frames

4.1 Introduction

Consider a crumpled receipt or a folded document which one would like to capture and save using a mobile device. It is often impossible to find the precise location and pose to capture the entire source with perfect quality. This is because some portions of the documents would not be directly facing the image plane while other portions may be out of focus, or experience inconsistent lighting. (Figure 4.1)

One possible solution is to capture and model the 3-D structure to "flatten" the document using dewarping algorithms to reconstruct the original planar surface. However, these methods either require external sensors such as structured light [55,56] or light grid projectors [57] which makes them inconvenient or even impossible for typical users or cannot handle complex distortion. It also may not be desirable in outdoor environments. Instead of seeking to recover the whole document at once, an alternative approach may be to attempt to recover locally "optimal" portions of the image, from a collection of possible poses.

In another task, consider having an interest in a planar object, such as a book cover or business logo, in a movie or a long video. If one wants to find a frame which best depicts that object with respect to its pose, one may have to manually browse through the entire video. An example set of frames for such a case is shown in Figure 4.2.

As suggested in the case of crumpled documents, one may assert that this can be handled by applying a pose estimation solution for planar objects which seeks to estimate the relative pose of an object with respect to a reference (frontal) image. This has been addressed in a number of article including [58–60] which were shown to have reliable and stable performance. However, these methods all share the same

Figure 4.1: Set of frames showing a folded document in different poses representing the case of crumpled document.

Figure 4.2: Set of frames extracted from a video which shows different poses of an object of interest.

limitation in that they assume the reference model (frontal image) is provided a priori. This makes them unsuitable for handling this case because the assumption of having an ideal frontal image beforehand directly conflicts with the very purpose of our goal. Homography decomposition [61–63], on the other hand, does not require this assumption and can estimate the surface normal of a planar surface with respect to the optical axis of a camera when given a pair of images. However, it suffers from highly unstable performance and also provides results which are ambiguous.

We claim that these problems can be handled in a common framework which relies on analyzing the poses of the local planar targets and selecting the best one when given images or a video which span different viewpoints. Without loss of generality, the best shot of a planar target can be considered as the one capturing the pose closest to the frontal pose of the target.

In this paper we develop the concept of *evaluating the frontalness* of the image of a planar source by measuring how well the surface normal of a planar object aligns with the optical axis of a camera. We show that measuring the relative frontalness can be analyzed by noting that if an image is assumed to be a true frontal image (as a reference), but is not, it shows limited ability to represent other non-frontal images. In other words, a less frontal image has less representability for different poses of an object than a more frontal image. Based on this observation, we estimate

the relative frontalness by comparing the objective space errors for a given image pair, first setting one of the two images as the true frontal image (reference image), then setting the other. Objective space error values are acquired by applying a state-of-the-art pose estimation algorithm for planar objects [58].

4.2 Our Method

4.2.1 Overview

We assume that we have a short video or camera burst of a planer source, captured from different orientations, sufficient to adequately capture at least one instance that would be considered acceptably "frontal". Given a pair of candidates, our goal is to evaluate the relative frontalness of the images and select the one which is more frontal. Through multiple pairwise comparisons, we can ultimately find the best or most frontal candidate. Since our method does not use any temporal information, it can be applied to any unordered set of images in an equivalent manner.

In order to evaluate the relative frontalness of a target, we use a pose estimation error-based method. Typically, pose estimation is used to estimate the pose of an object with respect to a set of model points which are assumed to be known beforehand. However, in our case, the pose estimation algorithm is employed to measure the pose estimation error, or objective space error for an image with respect to another image. Thus, to compare the pose estimation errors for each image in a pair, the error is computed twice, once with the first image as the reference model and the second time with the other image as the reference model.

The intuition behind this process is that, when the true (or more) frontal image is used as a reference image, the pose estimation error is smaller than the case where non (or less) frontal image is used as the reference. This occurs because a true-frontal image can be used to reproduce non-frontal images by perspective projection, whereas the non-frontal image has a limited ability to reproduce other non-frontal images. Detailed explanation on our method is explained in the following subsection.

4.2.2 Frontalness evaluation with known intrinsic camera parameters (K)

Let us first summarize the typical approach for a pose estimation procedure. Consider n coplanar model points $\mathbf{p}_i = \begin{bmatrix} p_{ix} & p_{iy} & 0 \end{bmatrix}^T$ in reference coordinate system. These points can be transformed into the camera coordinates $\mathbf{v}_i$ by:

$$\mathbf{v}_i \propto R\mathbf{p}_i + \mathbf{t}, \tag{4.1}$$

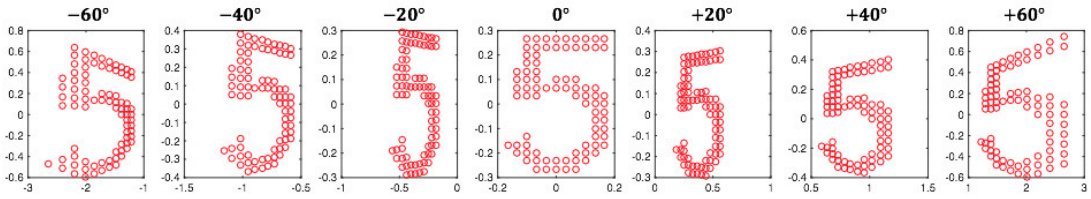
where $\propto$ indicates that the left hand side is directly proportional to the right hand side, due to the fact that $\mathbf{v}_i$ can only be computed up to a scale. Note that R and $\mathbf{t}$ indicate the 3 dimensional rotation and translation vectors, respectively, which are also known together as extrinsic camera parameters. Under the assumption that the image coordinate system aligns with the reference coordinate system, the task of estimating the pose of a camera with respect to the reference coordinate system, is to estimate R and $\mathbf{t}$. So in principle, a pose estimation algorithm seeks to find the values for R and $\mathbf{t}$ that minimizes an error function. We use the object-space error, as used by [58, 60, 64], which can be written as:

$$E_{os}(\hat{R}, \hat{\mathbf{t}}) = \sum_{i=1}^{n} \parallel (I - \hat{V}_i)(\hat{R}\mathbf{p}_i + \hat{\mathbf{t}}) \parallel^2 \text{ with } \hat{V}_i = \frac{\hat{v}_i \hat{v}_i{}^t}{\hat{v}_i{}^t \hat{v}_i}. \tag{4.2}$$

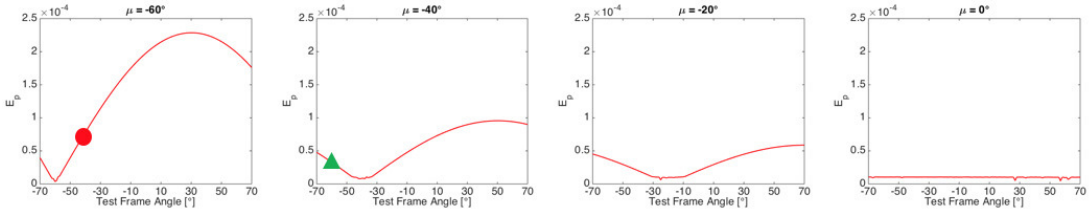
For evaluating frontalness, we exploit the objective error itself which is being minimized in the pose estimation process instead of utilizing $\hat{R}$ or $\hat{\mathbf{t}}$. When given a pair of images, we first acquire a set of corresponding features from both images (in our case, SIFT [23] and RANSAC [24]). These feature coordinates are then normalized (i.e., transformed to camera coordinates) using the camera intrinsic parameters (represented by the matrix K) which are assumed to be known.

Using the transformed feature coordinates, we perform the pose estimation (Eq. 4.2) twice. In each case, one of the two images is chosen as the reference image. Lastly, we compare the two error values to decide which one better fits as the reference image or "which one is more frontal". Note that the smaller error value indicates that the reference image has been chosen well and this image serves better as a representative for the other image.

The overall process of frontalness evaluation given a set of corresponding feature coordinates extracted from a pair of images (image i and image j) is computed



(a)



(b)

Figure 4.3: (a) Synthetic images of number "5" with various rotations captured by perspective camera model. (b) Objective space error plot for different reference images. X-axis: Test image angle ($-70°$ to $+70°$), Y-axis: E_p.

as shown below:

$$f^* = \begin{cases} i, & E_p(j|\mu = i)/E_p(i|\mu = j) \leq 1 \\ j, & \text{otherwise} \end{cases} \tag{4.3}$$

where f^* and μ indicate *image to be chosen (more frontal of the two)* and the *model frame*, respectively. Also note that $E_p(j|\mu = i)$ indicates the pose estimation error of image j when image i is set as the reference image.

To verify that our method of comparing E_p is a reasonable approach for frontalness evaluation, we have run a simulation using a synthetic dataset generated by a perspective camera model with known K. The images of a number "5" in various poses were captured by rotating the camera between $-70°$ to $70°$ with respect to the y axis (Figure 4.3a). Each graph in Figure 4.3b is acquired by plotting the objective space error (E_p) for all the images in the dataset with respect to a reference model (μ). Observe that the E_p values generate a smoothly changing plot which is minimum when the reference model (μ) is used as as the test model.

Now consider one example of comparing the E_p values which correspond to the two locations with the circle and the triangle marks in Figure 4.3b. It clearly shows that $E_p(-60°|\mu = -40°)$ is smaller than $E_p(-40°|\mu = -60°)$, and this verifies

that the image with $-40°$ angle is indeed "closer to the true frontal" than the image with $-60°$. By comparing any two E_p values in two different plots, one can verify that the method can be applied in general.

4.2.3 K-Invariant projective space

In applying the method described in the previous subsection, we assume that the camera intrinsic parameters (K) are known. This means it remains a challenge for uncalibrated cameras where K is inconsistent or unknown [65]. There may be a case where K is constantly changing due to zooming even if a same camera is used. When the goal is to evaluate a set of randomly collected images from a web search, K is also unknown and most likely different for each image. In such cases, we need to transform the points from two images onto a space to make them invariant to the camera intrinsic parameters. This can be done by using a projective transformation as used in [66, 67].

Consider three non-collinear points in one image $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3)$ and their corresponding points in a second image $(\mathbf{p}_1^*, \mathbf{p}_2^*, \mathbf{p}_3^*)$, both in image coordinates. The image coordinates of these points are acquired by equations:

$$\mathbf{P} = \mathbf{KV} \quad \text{and} \tag{4.4}$$

$$\mathbf{P}^* = \mathbf{K}^*\mathbf{V}^*. \tag{4.5}$$

where $\mathbf{P} = [\mathbf{p_1}\ \mathbf{p_2}\ \mathbf{p_3}]$, $\mathbf{V} = [\mathbf{v_1}\ \mathbf{v_2}\ \mathbf{v_3}]$, $\mathbf{P}^* = [\mathbf{p_1^*}\ \mathbf{p_2^*}\ \mathbf{p_3^*}]$, and $\mathbf{V}^* = [\mathbf{v_1^*}\ \mathbf{v_2^*}\ \mathbf{v_3^*}]$. Here, v_i is a point represented in camera coordinates as in Eq. 4.1. Since we assumed that these three points are not collinear, matrices $\mathbf{P}$ and $\mathbf{P}^*$ are non-singular which can define two different projective spaces, for example, γ and γ^*. Thus, we can transfer the points in the images on to those spaces as $\mathbf{w} = \mathbf{P^{-1}p}$ and $\mathbf{w}^* = \mathbf{P^{*-1}p}^*$, respectively. Thus, if we consider these equations with Eq. 4.6 and Eq. 4.7, we can observe that $\mathbf{w}$ and $\mathbf{w}^*$ are you below:

$$\mathbf{w} = \mathbf{P^{-1}p} = \mathbf{V^{-1}K^{-1}Kv} = \mathbf{V^{-1}v} \qquad \text{and} \tag{4.6}$$

$$\mathbf{w}^* = \mathbf{P^{*-1}p}^* = \mathbf{V^{*-1}K^{*-1}K^*v}^* = \mathbf{V^{*-1}v}^*. \tag{4.7}$$

For generating the matrices $\mathbf{P}$ and $\mathbf{P}^*$, three non-collinear points from each images need to be chosen. These points are automatically chosen so as to maximize

Figure 4.4: Sample images from the dataset including scanned frontal images (top row) and corresponding non-frontal images (bottom two rows).

the spacing as recommended in [68]. We will show in the following section, that this approach indeed increases the accuracy of frontalness evaluation on images with unknown K.

4.3 Experimental Evaluation

The experimental evaluation was carried out by targeting two real data scenarios based on the availability of camera intrinsic parameters (K). First, we evaluated our method assuming that the camera is calibrated (i.e., K is known). The camera intrinsic parameters were obtained beforehand using the the calibration method introduced in [69]. We compare the performance of our method with the homography decomposition-based method [62].

Second, we performed an evaluation on images under the assumption that K is unknown. In this scenario, we compare the performance of two different methods: 1) our method with a known or fixed K and 2) our method which uses a K-invariant space.

For both experiments, frontalness evaluation was performed on each possible

pair of images deciding which of the two images is more frontal. The overall accuracy is computed as the percentage of the correct pairwise decisions over all possible pairs in the given dataset.

Lastly, we include two samples results which qualitatively verify that our method performs well in selecting the most frontal image from a set of images.

4.3.1 Experiment 1: Calibrated Camera, Known K

We have constructed a new dataset as there are no public dataset available targeting the evaluation of frontalness. The dataset consists of 1200 images which were captured using the camera on iPhone5s with the resolution of 3264 x 2448 (w x h). This includes 30 different planar objects (books, documents, boxes), with each object being captured in 40 different camera angles and distances. The images were captured so that the angle between the optical axis of the camera and the surface normal of the plane ranges between 0° to 50°, approximately, distributed in various random directions.

To evaluate the performance of each decision, the angle between a test image and the optical axis of the camera should be provided as groundtruth. Since it is difficult to directly measure and work with the optical axis of a camera, we computed the angle between each image in the dataset (non-frontal) with respect to its corresponding true frontal shot. The pose estimation method in [58], which is known to be one of the state-of-the-art in robustness and accuracy, was used to compute the angles and be saved as the groundtruth. The true frontal image of each planar object was acquired by scanning the frontal surface of the object using a flatbed scanner. Figure 4.4 shows some of the selected images of frontal (scanned) and non-frontal shots from the dataset.

Each decision is made in a pairwise manner. Thus, testing was performed on every possible image pair in the dataset, which sums up to 23.4k pairs. The frontalness evaluation accuracy of our method and the homography decomposition-based method (baseline) for the overall dataset is shown in Table 1. Our method clearly outperforms the baseline method.

To better analyze the capability of our method with different difficulty levels, we have defined the measure of difficulty ν which can be computed for each image

Table 4.1: Frontalness Evaluation Accuracy

Homography-decomp	**68.35%**
Ours	**86.04%**

pair. We use the cosine similarity as the measure which is shown below:

$$\nu = \frac{\mathbf{A} \cdot \mathbf{B}}{\| \mathbf{A} \| \| \mathbf{B} \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{4.8}$$

where **A** and **B** are the two surface normal vectors of the two given images which are provided by the groundtruth.

The plot in Figure 4.5 shows the performance of our method with respect to the 7 different difficulty levels along with two sample pairs with minimum and maximum difficulty. The accuracy goes up to 97% for the easiest pairs while it performs 71% for the most difficult ones. Note that, however, as the difficulty level goes up, the appearance of the image pairs begin to resemble with each other, thus having low risk even if the decision is incorrect.

The frontalness evaluation of each pair of images requires less than a second (0.54 seconds in average for the given dataset) with MATLAB implementation on Intel Core i5 PC (2.6GHz CPU, 4GB RAM) excluding the feature extraction time.

4.3.2 Experiment 2: Randomly Collected Images, Unknown K

Our method explained in Section 4.3.1 which assumes that K is given, is not suitable for handling images captured with cameras with unknown intrinsic parameters. To validate the effectiveness of using K-invariant space with a pose estimation-based method, we have collected images of 3 different planar objects (a FedEx logo, a UPS logo, and a Wall Street sign), each at various rotations. For each planar object, 20 non-planar images are included along with the one true frontal image for each object. Note that there are 190 possible pairs for each object for evaluation. The groundtruth for each pair was generated in an equivalent manner as described in Experiment 1. The images were downloaded from the internet and sample images are shown in Figure 4.6.

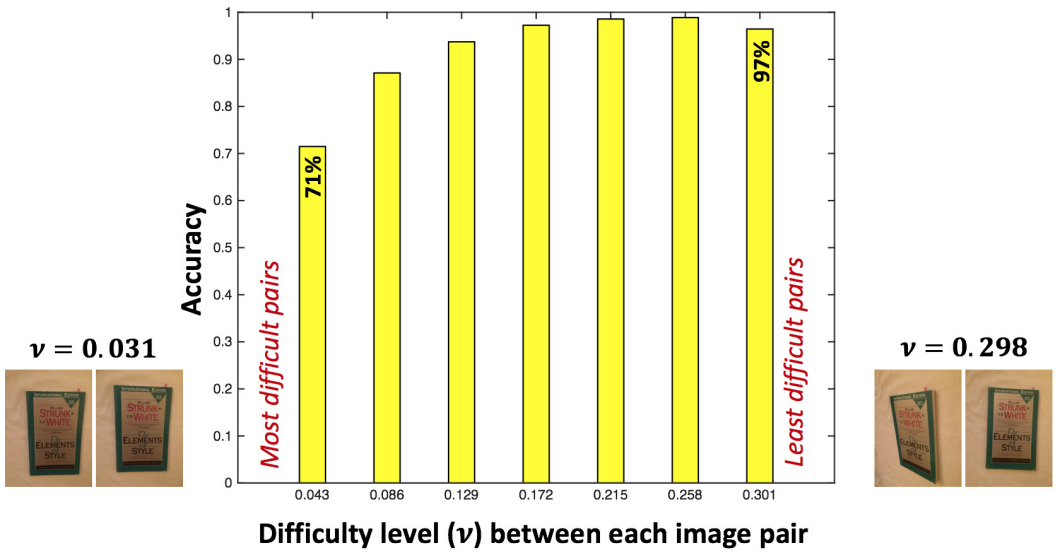


Figure 4.5: Frontalness evaluation accuracy with respect to difficulty levels. Testing dataset size = 23.4k pairs.

We compare the performance of two different methods: our method which assumes known/fixed K, our method which uses K-invariant space (KIS). When applying the method which assumes known/fixed K, we have used the K of our pre-calibrated camera to transform the points to camera coordinates in order to make a fair comparison. The performance comparison is shown in Figure 4.7 and it depicts the effectiveness of applying the K-invariant space. However, the overall performance does not quite reach the accuracy shown in the known/fixed K cases.

4.3.3 Qualitative Results

In addition, we show that our method can be used in selecting the best characters from a set of 40 images with various viewpoints. The sample images are shown in Figure 4.8a. Each character in different images are assumed to be residing on piecewise planar surfaces. Bounding boxes for the characters were manually assigned so that the evaluations are carried out within the same set of characters. Compare the best set of characters with the worst set of characters in Figure 4.8b and Figure 4.8c, respectively.

In addition, our method of performing the pairwise comparison of the E_p values can easily be used on a set of images to order them in terms of their frontalness. We

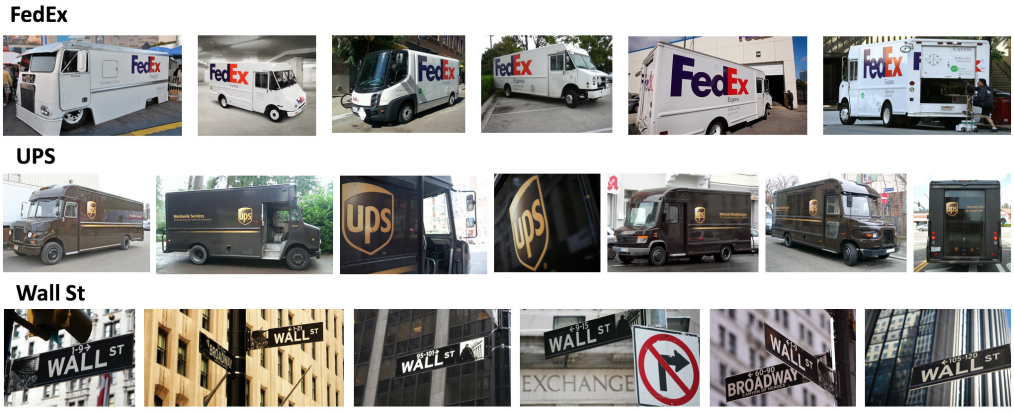


Figure 4.6: Sample images from the dataset for cases with unknown K.

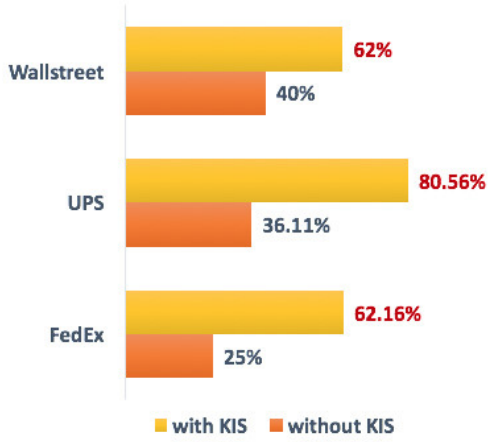
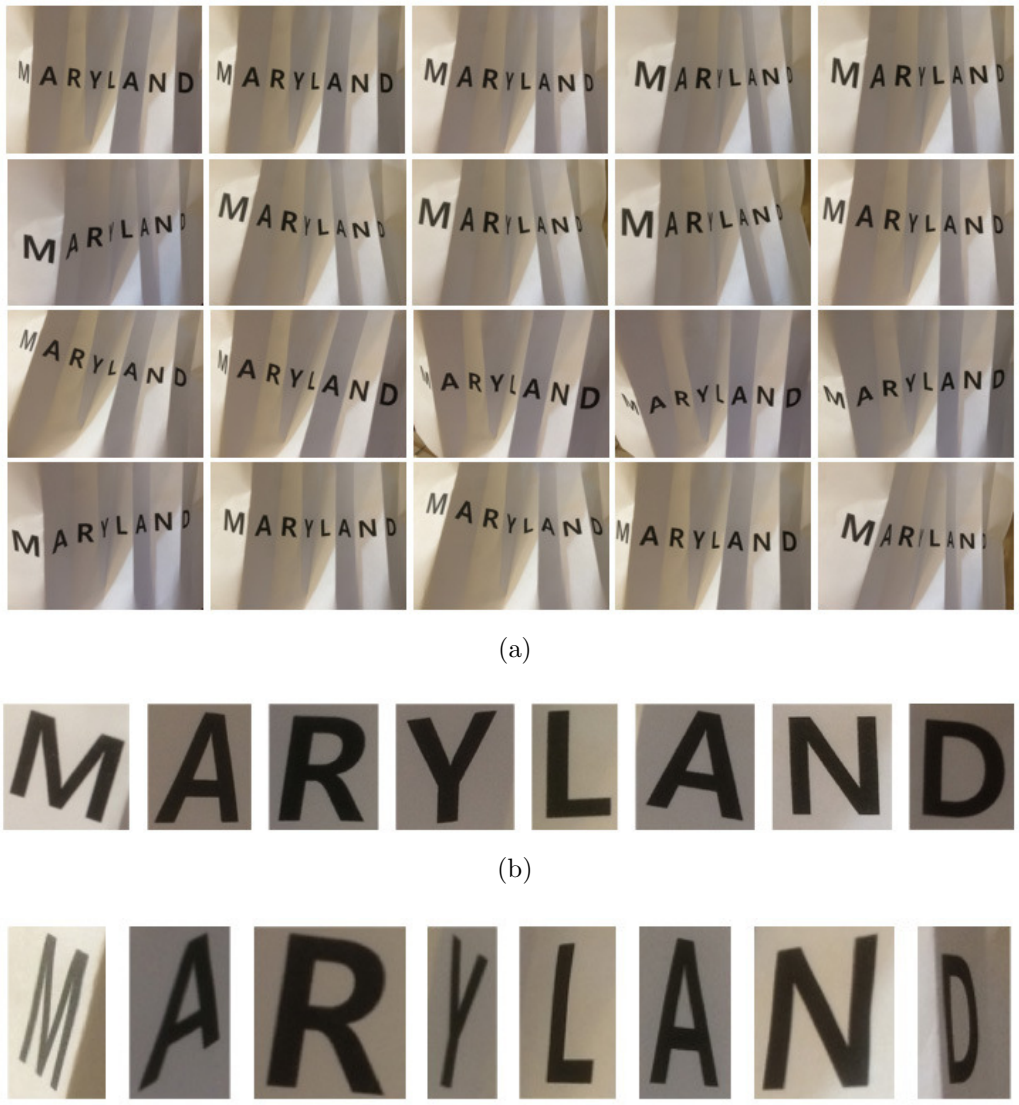


Figure 4.7: Frontalness evaluation accuracy on dataset with unknown K. Using K-invariant space (KIS) shows its effectiveness.

have selected one of the objects from the dataset introduced in 4.3.1 and applied our method. The resulting ordered images are shown in Figure 4.9.

4.4 Summary

In this paper, we have devised a novel method for evaluating the frontalness of planar objects. Our method takes a pair of images at a time to measure the relative frontalness between the two by exploiting the objective space error. Each run only requires a fraction of a second which makes it possible to be applied in real applications. Unlike the previous pose estimation methods that strictly require a true frontal image of the target object as a reference model, our method does not



(a)

(b)

(c)

Figure 4.8: (a) Sample Images of a folded document captured in different viewpoints. (b) Characters with highest frontalness. (c) Characters with lowest frontalness.

Figure 4.9: Ordered images with respect to their frontalness, from high to low.

require any reference model. Moreover, by introducing K-invariant space, we show that the proposed method can be applied even when the camera intrinsic parameters are unknown. The approach can be applied to optimizing the reconstruction of severely crumpled documents from a short video scan, especially the cases where a character or any continuous content reside on two or more piecewise planar surfaces. In addition, bringing more efficiency in terms of computation time would trigger real time applications or auto capturing of planar objects using mobile devices.

Chapter 5: IOD-CNN: Integrating Object Detection Networks for Event Recognition

5.1 Introduction

To better perform event or action recognition, recently introduced approaches have exploited the importance of considering semantically relevant and distinctive objects. For example, Althoff et al. [70] showed that the statistics derived from object detection results can better represent events. Joel et al. [71] claimed that event recognition performance can be enhanced by incorporating semantically related keywords which represent the salient objects. Jain et al. [72] showed that objects do matter for actions by encoding object categories that benefit action recognition as well as object localization.

Recently, Wang et al. [73] presented an approach which uses two separate deep convolutional neural networks (CNNs): an object CNN and a scene CNN. They used a simple late fusion to combine the fully connected (FC) layer outputs from the networks and applied a support vector machine (SVM) for classification. Similarly, an enhanced network was introduced in [74] by incorporating the local features (TDD: Transformed Deep-convolutional Descriptor), because the features from the FC layers were found to be weak in capturing the local information in the images. Both approaches use separate networks which are integrated with a late fusion.

In our approach, we exploit the power of deep convolutional neural networks (CNNs) in combining different networks (for different tasks) together in an end-to-end multi-task learning scheme. Learning a unified network allows better harvesting of the semantically relevant object information to boost event recognition. We incorporate event recognition as a primary task and relevant object detections as secondary tasks. This approach is motivated by previously methods [75–79] which have

demonstrated that a task can be better learned assisted by appropriate secondary tasks.

There are several technical challenges in constructing a unified deep network which integrates image classification (event recognition in our case) and object detection which are architecturally different in nature. First, the image classification system must pass an input image through the sequential layers of a network and generate class probability scores as an output [80–83]. Second, object detection must generate local candidate object region of interests (RoIs) which are evaluated to compute their scores. We inherit a widely used object detection approach called the Fast R-CNN [84] for this. This object detection approach uses RoI generation and RoI pooling steps which are the two primary differences when compared to the aforementioned image classification.

To integrate these architectures, we devised a unified CNN framework which enables the sharing of the convolutional layers, one FC layer and one RoI pooling layer between image classification and object detection. As the CNN is integrated by object detection modules, we refer to it as an Integrated Object Detection (IOD)-CNN. The fact that the image classification also uses the RoI pooling layer (which is different from typical image classification) not only makes the network differ in appearance, but also adds beneficial functionality. With the help of the shared RoI pooling layer, it is no longer necessary to resize the input images to a fixed size. This allows the use of high-resolution images as input, providing room for classification performance enhancement.

For image classification, the input to the RoI pooling (i.e., RoI), is the entire region of the input image. For object detection, object proposals generated by the selective search (rigid objects) or by the multi-scale sliding window search (non-rigid objects) are used as inputs to the RoI pooling.

Our contributions can be summarized as:

1. The introduction of a novel unified deep CNN architecture which integrates architecturally different, yet semantically-related networks for different secondary tasks to enhance the performance of a primary task

2. A demonstration of the effectiveness of the novel approach by showing that the performance of event recognition (primary task) can be boosted by incorporating rigid and non-rigid object detection.

3. The fact that our architecture can be further enhanced by appending a late fusion, indicating that early-sharing of the layers is complementary to the late fusion.

5.2 Our Approach

5.2.1 IOD-CNN: Integrated Object Detection CNN

In this section, we elaborate on three tasks of event recognition, rigid object detection and non-rigid object detection followed by the modifications we made to architectures which implement them. We then explain how these different architectures are integrated into a unified network.

Event Recognition. We use a common classification architecture, known as ConvNet [80], for event recognition. As shown in Fig. 5.1a, the network typically consists of a number of convolutional layers followed by several FC layers. The input is an image with predefined fixed width and height for both training and testing, while the output is the softmax probability estimates over all of the classes.

Rigid Object Detection. As shown in Fig. 5.1b, the Fast R-CNN (FRCN) [84] was chosen to perform the rigid object detection. Unlike the deep ConvNet which requires resized images as input, the original FRCN architecture takes in a full image as input and passes it through a series of convolutional layers to generate a feature map. This map along with approximately 2000 object proposals generated by selective search are then fed into a Region of Interest (RoI) pooling layer. The output from the RoI pooling is fed into the FC layers which are followed by two output layers: one for the softmax class-wise probability estimation and the other for the bounding box regression.

The bounding box regression is removed from our architecture (dotted box in Fig. 5.1b) because the primary task of event recognition does not benefit from it. This is due to the fact that the power of bounding box regression in the original FRCN is exhibited in the post-processing which is separate from the learning process. We have experimentally observed that when object detection is learned along with the bounding box regression in a multi-task scheme, the performance degrades

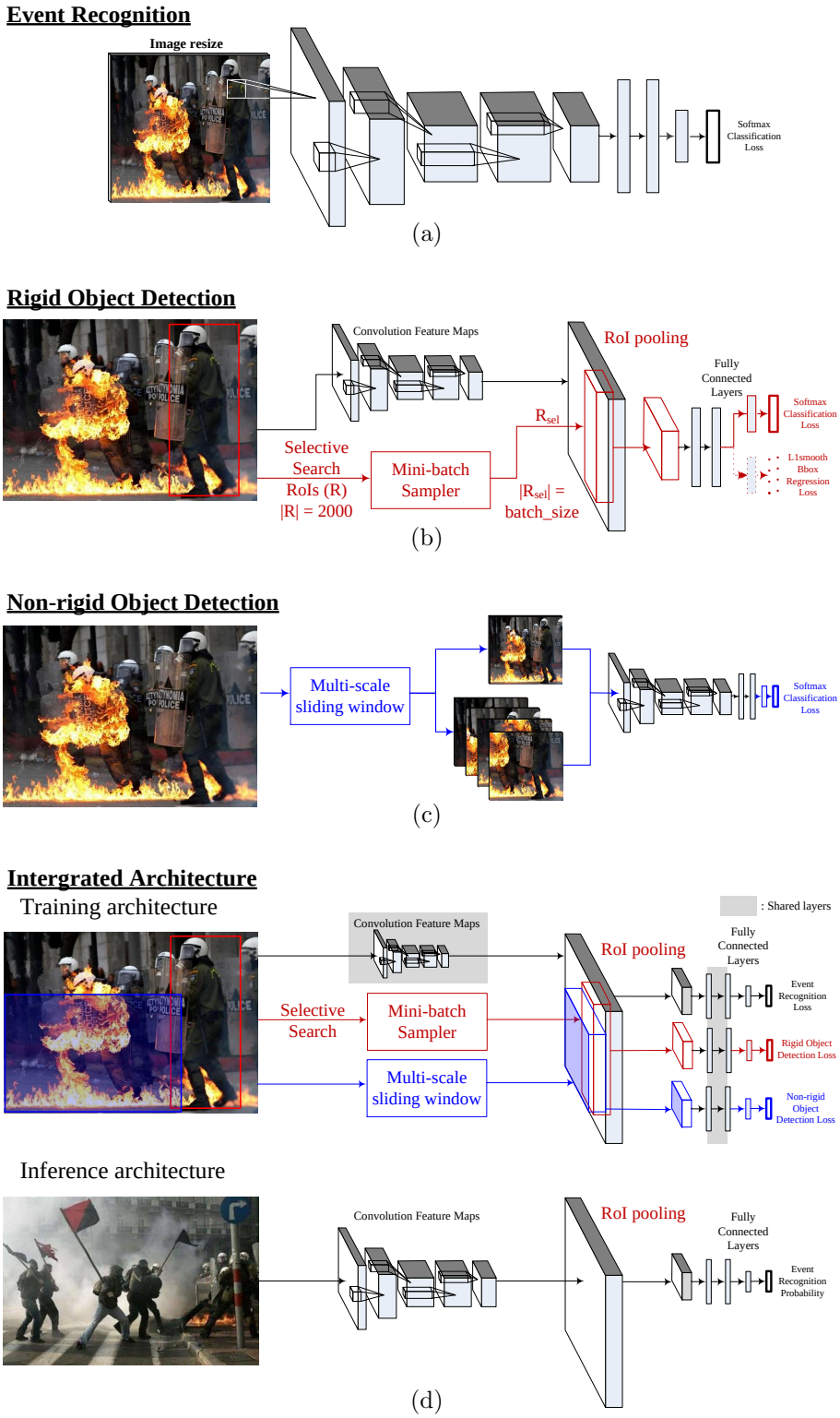


Figure 5.1: IOD-CNN architecture. (a, b, c) Architectures for three separate tasks before the integration (d) A novel architecture which integrates event classification with rigid and non-rigid object detection.

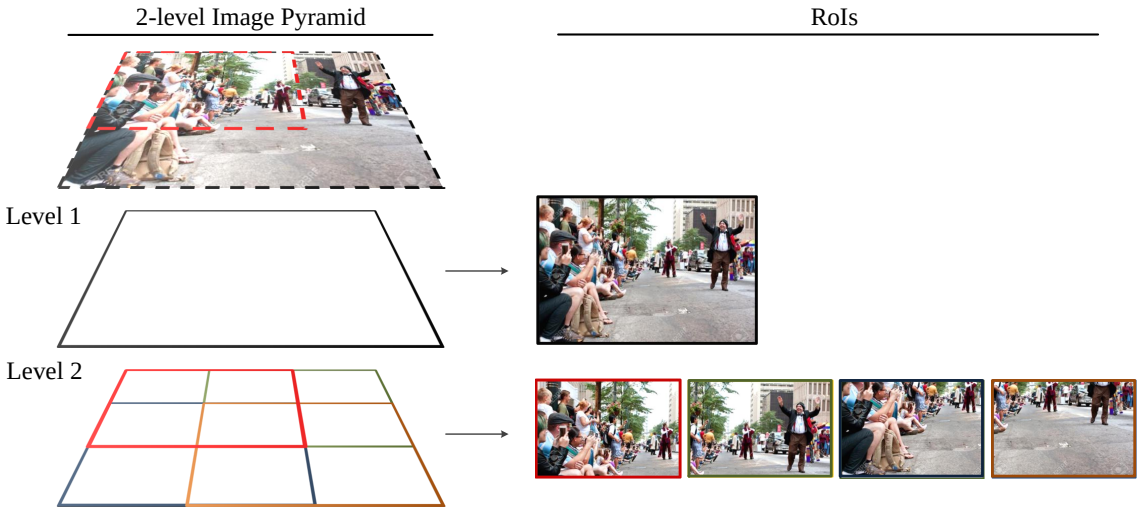


Figure 5.2: Multi-scale sliding window for non-rigid object detection

unless the bounding box regression post-processing is existent. In short, incorporating the bounding box regression into our architecture will have a negative effect for the primary task, as there is no chance to perform the post-processing to make up for the loss.

Non-rigid Object Detection. Modeling the "objectness" for objects with non-standard or non-rigid shape, such as smoke or fire, is not only difficult but also computationally expensive. Thus, instead of using a fine scanning method such as selective search which is used for rigid object detection, we use a multi-scale sliding window strategy as shown in Fig. 5.2. For one input image, five RoIs are generated: one covering the whole image region and the others covering the four overlapping regions with 2/3 height and 2/3 width of the whole region. These five RoIs are fed into the network shown in Fig. 5.1c.

Integrating the Different Architectures. The unified network for training and inferencing are shown in Fig. 5.1d. The training architecture consists of a series of convolutional layers, a RoI pooling layer, and three separate modules responsible for event recognition, rigid-object detection and non-rigid object detection, respectively. Each module consist of one shared and two non-shared FC layers. For testing, only the components responsible for the event recognition (primary task) are included in

the architecture.

The training network takes an input image and passes it through a series of convolutional layers until it reaches the RoI pooling layer. At the same time, the input image goes through two different sample generators: a selective search and a multi-scale sliding window search, generating samples for rigid and non-rigid object detection, respectively. The output of the convolutional layers along with the outputs of the two sample generators are fed into the shared RoI pooling layer. The three task-specific streams go through the FC layers. Each stream is connected to an appropriate loss function at the end.

The effective integration of these architectures was made possible by sharing the convolutional layers and the first FC layer (known as *fc6*) which are learned to serve all three tasks. Note that, the other two "task-specific" FC layers (*fc7* and *fc8*) are learned separately for different tasks. By sharing these layers, we provide each task a means to associate the information from the other tasks. In the experiments (Section 5.3), we show that the performance of our primary task is indeed boosted by this integration. In addition, although the RoI pooling layer is not a layer to be learned, it serves a crucial role in allowing full-size input images to be fed into the convolutional layers without resizing.

It is noted in [80] that the first convolutional layer (*conv1*) is more generic and task independent than other convolutional layers. In our case, we share a similar philosophy, but we also show that the network can be better learned when the overall set of convolutional layers is shared and learned together between the semantically-related tasks.

5.2.2 Learning the Unified Network

We have found the network introduced by Krizhevsky et al. [80] suitable for the single-task event recognition architecture. In our experiments, we used the Malicious Crowd Dataset [71], which is described more fully in Section 5.3.1. To label the RoI for training in the rigid and non-rigid object detection, we have used 0.5 and 0.2 as the thresholds for the intersection over union (IoU) metric. While the *fc6* and *fc7* are fine-tuned, the weights for *fc8* are initialized by samples from a Gaussian distribution with zero mean and 0.1 standard deviation.

For every iteration, a batch of two images is used. We made sure that each

batch is comprised of one sample with a benign label (a normal scene) and one with a malicious one (which would draw attention of law enforcement). For training the rigid object detection, the network takes 64 RoIs from each image which is the selected subset of the initial RoI set provided by the selective search. For event recognition and non-rigid object detection, 1 and 5 RoIs, respectively, are generated per image, thus 2 and 10 RoIs are used as one batch.

Cascaded Optimization. One technical challenge in learning the IOD-CNN is selecting the appropriate learning parameters. Naively using the parameters optimized for one of the three modules may not be suitable for acquiring the best performance out of the unified network. For the event recognition and non-rigid object detection, all the RoIs acquired from one image are used for one batch. However, for the rigid object detection, approximately 2000 RoIs are generated per image and only the subset of those RoIs (i.e., 64 for malicious and 64 for benign) are used per batch. To allow more training iterations for the rigid object detection module, we have employed a three step cascaded optimization strategy. The initial CNN network is first trained on the Places Dataset [85]. Then only the rigid object detection module is learned/fine-tuned on the target Malicious Crowd Dataset using the learning rate of 0.01, 30k iterations, and the step size of 20k. Lastly, the unified network (i.e., IOD-CNN with all the modules) is trained with the learning rate of 0.0001, 12k iterations, and the step size of 8k.

5.3 Experimental Evaluation

5.3.1 Dataset

To demonstrate the effectiveness of our architecture, we use the Malicious Crowd Dataset introduced in [71]. This dataset was chosen as it contains not only the crowd event images but also the ground truth labels for relevant objects which are suitable for testing our architecture which requires both image classification and object detection. The dataset contains 1133 crowd images, equally split into *malicious* and *benign* classes (Sample images are shown in Fig. 5.3. The *malicious* label is said to have been assigned to an image when the scene would be alarming to a passerby or a law enforcement personnel. For both classes, the images contain two

(a) benign examples

(b) malicious examples

Figure 5.3: Sample images from the Malicious Crowd Dataset with two classes: (a) benign and (b) malicious events

different types of objects: rigid (e.g., cars) and non-rigid (e.g., smoke). The dataset also provides the bounding boxes of the frequently appearing "*malicious*-related" objects which are police, helmet, car, fire, and smoke. The bounding boxes are used to train and evaluate the rigid and non-rigid object detection. Details on how the objects are selected is given in [71].

5.3.2 Performance Evaluation

We have carried out a set of experiments to demonstrate how our architecture integration approach can boost event recognition performance to a new state-of-the-art. For all the experiments described in this subsection, we have used the Malicious Crowd Dataset briefly described in the previous subsection.

The first six rows of Table 5.1 show that IOD-CNN without any fusion processing outperforms all the baseline single CNNs. The results indicate that integrating rigid (R), non-rigid (N), or both (R,N) object detections into the network all show superior performance, and integrating both works the best. Moreover, we verify that incorporating the RoI pooling layer which allows the input images of arbitrary size, increases the performance.

Table 5.1: **Event recognition** average precision (AP). All methods use [80] as the baseline architecture. Task: **E**: Event Recognition, **R**: Rigid Object Detection, **N**: Non-rigid Object Detection. [71]* reproduces the result of [71] with our network learning strategy.

Method	Tasks	AP
Single CNN [71]	-	72.2
Single CNN [71]*	-	82.5
Single CNN+RoI pooling	-	90.2
IOD-CNN	E, R	91.8
IOD-CNN	E, N	91.9
IOD-CNN	E, R, N	**93.6**
2 CNNs&DPM+Score Fusion [71]	-	77.1
OS-CNN+fc7&TDD Fusion [74]	-	92.9
3 Separate CNNs+Score Fusion	-	92.9
IOD-CNN+Score Fusion	E, R, N	93.9
IOD-CNN+fc7&TDD Fusion	E, R, N	**94.2**

Table 5.2: **Single task versus multitask performance.** **C**: Classification, **D**: Detection, **R** and **N** used mean average precision (mAP) as the evaluation metric.

Method	C/D	Single-task (AP/mAP)	Multi-task (AP/mAP)
E	C	90.2	**93.6**
R	D	**11.8**	11.0
N	C	27.7	**82.1**

In the last five rows of Table 5.1, we have also compared IOD-CNN with two baselines [71, 74] which use multiple CNNs and exploit fusion strategies. To make a fair comparison with the baselines, we use the same fusion techniques, i.e., score fusion [86] and fc7&TDD fusion. To generate a two stream network, we prepared two networks pretrained on the ImageNet [87] and the Places [85] Datasets, as in [74]. By applying the same score fusion or fc7&TDD fusion used in [71] and [74], the performance of *pre-fusion* IOD-CNN is improved by 0.3 and 0.6 AP, respectively. This indicates that the early-sharing of the network layers (convolutional and one FC) is complementary to the late fusion in terms of the performance. The IOD-CNN with either of the fusion strategies outperforms all the baselines and the case where 3 separate CNNs (E,R,N) are score-fused.

We have also carried out an experiment to analyze how the performance of each task changes when all the tasks are learned together using the IOD-CNN. Table 5.2 shows that the event recognition and the non-rigid object detection performance is boosted when learned together. Notably, the non-rigid object detection performance improved drastically by almost three fold.

5.4 Summary

We presented a novel unified deep CNN architecture which integrates architecturally different, yet semantically-related networks for different tasks to enhance the performance of event recognition. The experimental results show that each of the newly incorporated architecture components are crucial in boosting the performance. The architecture which integrates the two object detections with the

event recognition outperforms the previous object-aware event recognition CNNs. As one unified network is learned in an end-to-end fashion, the training can also be performed more efficiently. Moreover, the performance of our architecture can be further improved by appending a late fusion approach. This indicates that the within-network sharing of the layers is complementary to the late fusion.

Chapter 6: Summary of Thesis Contributions and Open Problems

In the first part of this dissertation, we have addressed challenges in capturing documents in unconstrained environments including 1) a limited field-of-view keeping users from acquiring a high-quality images of large sources in a single frame, 2) light reflections on glossy surfaces that result in saturated regions, and 3) crumpled or non-planar documents that cannot be captured effectively from a single pose. In the second part, we have addressed the challenge of effectively integrating multiple CNNs with different architectures, yet with relevant tasks.

The following subsections summarize our approach to each topic, our contributions and open problems.

6.1 Sharpness-aware Document Image Mosaicking Using Graphcuts

6.1.1 Overview of Approach

To address the unique problems associated with document image mosaicking, we used a novel Graphcut-based document image mosaicking method which focuses on lessening the known artifacts such as ghosting effects and missing contents. The major contribution is that we have incorporated a sharpness measure into the Graphcut formula which induces the cut generation in a way that results in selecting the sharpest pixels from the source images. We also incorporated geometrical disposition between the overlapped images to minimize the errors at the boundary regions. Proposed method not only generates visually pleasing mosaicked results but also outperforms previous methods quantitatively, in terms of OCR accuracy.

6.1.2 Summary of Contributions

- A novel document image mosaicking method is presented which allows the acquisition of a single, high-quality, digital copy of a document from multiple overlapping shots.
- Graphcut-based blending is introduced which effectively stitches two overlapping images without requiring any prior knowledge of the document, thus being more robust and widely applicable.
- Boundary constraints are imposed which minimize discrepancy between overlapping and non-overlapping regions.
- A sharpness measure is incorporated which promotes cuts which favor mosaicked images with sharper pixels when blending the overlapping images.
- As there are no publicly available datasets for document image mosaicking, we have newly constructed a dataset where each test case is comprised of two partially overlapping shots of different types of documents (including equations, graphs, pictures, and tables) using a mobile device.

6.1.3 Open Problems

One of the remaining challenges is how to effectively deal with duplicate contents caused by mis-registration. The idea of incorporating the locations of the corresponding features into the Graphcut formula so as to effectively avoid generating a cut that runs between the regions with duplicate contents can be explored. We also need to consider ways of realigning regions that are detected as having missing content with a second pass.

6.2 Joint Homography Estimation for Highlight Removal

6.2.1 Overview of Approach

To address the problem of having saturated highlights induced by light reflections on glossy surfaces, we proposed a method which exploits the fact that the reflections and the target contents reside on two separate virtual planes. We have

devised a novel algorithm which jointly estimates the two representative homographies for the target scene and the highlights. Unlike previous methods that only use the relationship between the target scenes in multiple images, we considered the relationship between the corresponding highlights as well the target scenes. The proposed method was shown to outperform previous approaches, especially showing its state-of-the-art quality in recovering the underlying contents in saturated highlight regions.

6.2.2 Summary of Contributions

- Our method is the first to successfully handle relatively large and saturated highlight regions obscuring the content underneath.
- Unlike some of the previous approaches that use homography between non-highlight regions, we newly use correspondences between "highlight" regions for better localization.
- We have exploited the observation that the distance between the camera and the virtual location of the light source is typically larger than the distance between the camera and the target content. Thus, it is reasonable to use two separate homographies in distinguishing the objects at different distances.
- We have shown that when two images are captured with a change of view point, the displacement of the desired content is different from the displacement of the highlight regions ('Motion parallax').
- We have verified that our approach outperforms closely-related approaches, showing its state-of-the-art quality in handling highly saturated highlights which obscure the underlying content. It requires fewer constraints in image acquisition and is faster than any other multi-view methods.

6.2.3 Open Problems

To lessen the potential artifacts with our current reflection removal method, we should explore the viability of using a video as input instead of using a pair of images. A method to automatically select a set of images which could reconstruct the optimal "reflection-removed" result can also be explored.

6.3 Content Selection Using Frontalness Evaluation of Multiple Frames

6.3.1 Overview of Approach

To address the problem of selecting best instances in a set of images in terms of their "frontalness", we proposed a novel method to evaluate the relative frontalness of an object by computing and comparing the objective space error of a pair of images. The novelty of our method is based on the observation that a true frontal image can be used to reproduce other non-frontal images by perspective projection, while the non-frontal images have limited ability to do so. To handle the cases where the intrinsic camera parameters (K) are unknown, we additionally propose the use of K-invariant space.

6.3.2 Summary of Contributions

- We have devised a novel method for evaluating the "frontalness" of planar objects by computing the objective space error for a pair of images.
- The novelty of our method is based on the observation that a true frontal image can be used to reproduce other non-frontal images by perspective projection, while the non-frontal images have limited ability to do so.
- Unlike the previous pose estimation methods that strictly require a true frontal image of the target object as a reference model, our method does not require any reference model.
- By incorporating K-invariant space, we show that the proposed method can be applied even when the camera intrinsic parameters are unknown.
- Each run only requires a fraction of a second which makes it possible to be applied in real applications.

6.3.3 Open Problems

Addressing the problem of "flattening" the crumpled document by employing a deep neural network which includes our local frontalness evaluation will be a good topic to follow. Incorporating unpooling layers and deconvolutional layers

could be effective for the purpose of learning the reconstruction process of degraded characters.

6.4 IOD-CNN: Integrating Object Detection Networks for Event Recognition

6.4.1 Overview of Approach

To address the problem of exploiting the power of deep convolutional neural networks (CNNs) in combining different networks (image classification and object detection in our case), we proposed a unified CNN framework which enables the sharing of the convolutional layers, one FC layer and one region of interest (RoI) pooling layer between image classification and object detection. As the CNN is integrated by object detection modules, we call it the Integrated Object Detection (IOD)-CNN. The major contribution is that we have introduced a novel unified CNN architecture which integrates architecturally different, yet semantically-related networks for different secondary tasks (rigid and non-rigid object detection) to enhance the performance of a primary task (event recognition). The IOD-CNN can be further enhanced by appending a late fusion strategy, indicating that early-sharing of the layers is complementary to the late fusion.

6.4.2 Summary of Contributions

- We introduced a novel unified deep CNN architecture which integrates architecturally different, yet semantically-related networks for different secondary tasks to enhance the performance of a primary task.
- Unlike the previous deep CNN-based event recognition approaches, our approach uses an architecture which shares the early portion of deep CNN layers.
- We demonstrated the effectiveness of the novel approach by showing that the performance of event recognition (primary task) can be boosted by incorporating rigid and non-rigid object detection.
- The fact that our architecture can be further enhanced by appending a late fusion, indicating that early-sharing of the layers is complementary to the late

fusion.

- Our network which integrates multiple tasks outperforms all the baselines and the case where 3 separate CNNs (Event recognition ,rigid object detection, non-rigid object detection) are score-fused.

6.4.3 Open Problems

Based on the successful integration of single image classification and object detection (which are semantically related) to boost the object detection performance, it would be meaningful to devise a unified network which is targeted to better recognize human actions or activities within a video by the help of semantically relevant tasks such as object detection, scene recognition or human pose estimation. Similarly, we might also construct a network which contains multiple experts, where each expert is specialized to better detect/recognize objects with different scales or different appearances.

Bibliography

[1] Y. Lecun, L. D. Jackel, H. A. Eduard, N. Bottou, C. Cartes, J. S. Denker, H. Drucker, E. Sackinger, P. Simard, and V. Vapnik, "Learning algorithms for classification: A comparison on handwritten digit recognition," in *Neural Networks: The Statistical Mechanics Perspective*. World Scientific, 1995, pp. 261–276.

[2] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, ser. ICDAR '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 958–963.

[3] M. Jaderberg, A. Vedaldi, and A. Zisserman, *Deep Features for Text Spotting*. Springer International Publishing, 2014, pp. 512–528.

[4] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[5] S. Sudholt and G. A. Fink, "Phocnet: A deep convolutional neural network for word spotting in handwritten documents," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Oct 2016, pp. 277–282.

[6] F. Wang, Z. Li, and Q. Liu, "Coarse-to-fine human parsing with fast r-cnn and over-segment retrieval," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 1938–1942.

[7] Y. Tang and X. Wu, "Scene text detection and segmentation based on cascaded convolution neural networks," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1509–1520, March 2017.

[8] Y. Li and M. Brown, "Single image separation using relative smoothness," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, 2014, pp. 2752–2759.

[9] Q. Yang, J. Tang, and N. Ahuja, "Efficient and robust specular highlight removal," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 6, pp. 1304–1311, 2015.

[10] Y. Li and M. Brown, "Exploiting reflection change for automatic reflection removal," in *Computer Vision (ICCV), IEEE International Conference on*, 2013, pp. 2432–2439.

[11] X. Guo, X. Cao, and Y. Ma, "Robust separation of reflection from multiple images," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, 2014, pp. 2195–2202.

[12] M. Brown and D. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007.

[13] A. Whichello and H. Yan, "Document image mosaicing," in *Proc. Int. Conf. on Pattern Recognition*, 1998, pp. 585–595.

[14] M. Pilu and F. Isgr, "A fast and reliable planar registration method with applications to document stitching," in *British Machine Vision Conference*, 2002, pp. 688–697.

[15] P. Shivakumara, G. Kumar, D. Guru, and P. Nagabhushan, "Sliding window based approach for document image mosaicing," *Image and Vision Computing*, vol. 24, no. 1, pp. 94–100, Jan 2006.

[16] T. Kasar and A. Ramakrishnan, "Block-based feature detection and matching for mosaicing of camera-captured document images," in *TENCON 2007 - 2007 IEEE Region 10 Conference*, Oct 2007, pp. 1–4.

[17] ——, "Ccd: Connected component descriptor for robust mosaicing of camera-captured document images," in *Document Analysis Systems, 2008. DAS '08. The Eighth IAPR International Workshop on*, Sept 2008, pp. 480–486.

[18] M. Ligang and Y. Yongjuan, "Automatic document image mosaicing algorithm with hand-held camera," in *Intelligent Control and Information Processing (ICICIP), 2011 2nd International Conference on*, July 2011, pp. 1094–1097.

[19] T. Lijing, Z. Yan, and Z. Huiqun, "A warped document image mosaicing method based on registration and trs transform," in *Computer and Information Science (ICIS), 2011 IEEE/ACIS 10th International Conference on*, May 2011, pp. 179– 183.

[20] J. Hannuksela, P. Sangi, J. Heikkila, X. Liu, and D. Doermann, "Document image mosaicing with mobile phones," in *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*, Sept 2007, pp. 575–582.

[21] J. Liang, D. DeMenthon, and D. Doermann, "Camera-based document image mosaicing," in *Proc. Int. Conf. on Pattern Recognition*, 2006, pp. 476–479.

[22] ——, "Mosaicing of camera-captured document images," *Computer Vision and Image Understanding*, vol. 113, no. 4, pp. 572–579, Apr 2009.

[23] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Computer Vision and Image Understanding*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[24] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun 1981.

[25] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, Nov 2001.

[26] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 2, pp. 147–159, Feb 2004.

[27] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1124–1137, Sept 2004.

[28] V. Kwatra, A. Schodl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: Image and video synthesis using graph cuts," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 277–286, Jul 2003.

[29] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, Jul 2007.

[30] J. Kumar, F. Chen, and D. Doermann, "Sharpness estimation for document and scene images," in *Pattern Recognition (ICPR), 21st International Conference on*, Nov 2012, pp. 3292–3295.

[31] A. Bagdanov, S. Rice, and T. Nartker, "OCR Frontiers Toolkit," http://code.google.com/p/isri-ocr-evaluation-tools/, 1999.

[32] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *ACM Transactions on Graphics*, 2000, pp. 417–424.

[33] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *Image Processing, IEEE Transactions on*, vol. 13, no. 9, pp. 1200–1212, Sep 2004.

[34] J. Gao, S. Kim, and M. Brown, "Constructing image panoramas using dual-homography warping," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, 2011, pp. 49–56.

[35] S. Shafer, "Using color to separate reflection components," *Color Research and Application*, vol. 10, no. 4, pp. 210–218, 1985.

[36] P. Tan, S. Lin, L. Quan, and H.-Y. Shum, "Highlight removal by illumination-constrained inpainting," in *Computer Vision (ICCV), IEEE International Conference on*, 2003, pp. 164–169.

[37] P. Tan, S. Lin, and L. Quan, "Separation of highlight reflections on textured surfaces," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, 2006, pp. 1855–1860.

[38] Q. Yang, S. Wang, and N. Ahuja, "Real-time specular highlight removal using bilateral filtering," in *Computer Vision (ICCV), IEEE International Conference on*, 2010, pp. 87–100.

[39] R. T. Tan and K. Ikeuchi, "Separating reflection components of textured surfaces using a single image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 2, pp. 178–193, Feb 2005.

[40] A. Levin, A. Zommet, and Y. Weiss, "Separating reflections from a single image using local features," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, 2004, pp. 306–313.

[41] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 9, pp. 1647–1654, Sept 2007.

[42] N. Kong, Y.-W. Tai, and J. Shin, "A physically-based approach to reflection separation: From physical modeling to constrained optimization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 2, pp. 209–221, Feb 2014.

[43] B. Sarel and M. Irani, "Separating transparent layers through layer information exchange," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2004, pp. 328–341.

[44] H. Farid and E. Adelson, "Separating reflections and lighting using independent components analysis," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, 1999, p. 267.

[45] Y. Schechner, J. Shamir, and N. Kiryati, "Polarization-based decorrelation of transparent layers: The inclination angle of an invisible surface," in *Computer Vision (ICCV), IEEE International Conference on*, 1999, pp. 814–819.

[46] Y. Schechner, N. Kiryati, and R. Basri, "Separation of transparent layers using focus," *International Journal of Computer Vision*, vol. 39, no. 1, pp. 25–39, Aug 2000.

[47] A. Agrawal, R. Raskar, S. Nayar, and Y. Li, "Removing photography artifacts using gradient projection and flash-exposure sampling," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 828–835, Jul 2005.

[48] R. Szeliski, S. Avidan, and P. Anandan, "Layer separation from multiple images containing reflections and transparency," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, 2000, pp. 246–253.

[49] S. Lin and H. Shum, "Separation of diffuse and specular reflection in color images," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, 2001, pp. 341–346.

[50] S. Lin, Y. Li, S. Kang, X. Tong, and H. Shum, "Diffuse-specular separation and depth recovery from image sequences," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2002, pp. 210–224.

[51] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 313–318, Jul 2003.

[52] J. Matas, O. Chum, M. Urban, and T. Padjla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2002, pp. 36.1–36.10.

[53] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.

[54] A. Artusi, F. Banterle, and D. Chetverikov, "A survey of specularity removal methods," *Computer Graphics Forum*, vol. 30, no. 8, pp. 2208–2230, 2011.

[55] M. S. Brown and W. B. Seales, "Document restoration using 3d shape: a general deskewing algorithm for arbitrarily warped documents," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, 2001, pp. 367–374.

[56] M. Pilu, "Undoing page curl distortion using applicable surfaces," in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 1, 2001, pp. 237–240.

[57] A. Doncescu, A. Bouju, and V. Quillet, "Former books digital processing: image warping," in *Document Image Analysis, 1997. (DIA '97) Proceedings., Workshop on*, Jun 1997, pp. 5–9.

[58] G. Schweighofer and A. Pinz, "Robust pose estimation from a planar target," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2024–2030, Dec 2006.

[59] Z. Jia, A. Gallagher, and T. Chen, "Cameras and gravity: Estimating planar object orientation," in *2013 IEEE International Conference on Image Processing*, Sept 2013, pp. 3642–3646.

[60] C. P. Lu, G. D. Hager, and E. Mjolsness, "Fast and globally convergent pose estimation from video images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 610–622, Jun 2000.

[61] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[62] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. SpringerVerlag, 2003.

[63] O. Faugeras and F. Lustman, "Motion and Structure From Motion in a Piecewise Planar Environment," *Intern. J. of Pattern Recogn. and Artific. Intelige.*, no. 3, pp. 485–508, 1988.

[64] P. Wunsch and G. Hirzinger, "Registration of cad-models to images by iterative inverse perspective matching," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, vol. 1, Aug 1996, pp. 78–83 vol.1.

[65] B. P. Wrobel, *Calibration and Orientation of Cameras in Computer Vision*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, ch. Minimum Solutions for Orientation, pp. 7–62.

[66] E. Malis, "Visual servoing invariant to changes in camera intrinsic parameters," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1, 2001, pp. 704–709 vol.1.

[67] E.Malis, "Visual servoing invariant to changes in camera-intrinsic parameters," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 1, pp. 72–81, Feb 2004.

[68] E. Malis and F. Chaumette, "2 1/2 D visual servoing with respect to unknown objects through a new estimation scheme of camera displacement," *International Journal of Computer Vision*, vol. 37, no. 1, pp. 79–97, 2000.

[69] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov 2000.

[70] T. Althoff, H. O. Song, and T. Darrell, "Detection bank: An object detection based video representation for multimedia event recognition," in *ACM Multimedia (ACMMM)*, 2012, pp. 1065–1068.

[71] J. Levis, H. Lee, H. Kwon, J. Michaelis, M. Kolodny, and S. Eum, "Joint deep exploitation of semantic keywords and visual features for malicious crowd image classification," *arXiv preprint arXiv:1610.06903*, 2016.

[72] M. Jain, J. C. Gemert, and C. G. M. Snoek, "What do 15,000 object categories tell us about classifying and localizing actions?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 46–55.

[73] L. Wang, Z. Wang, W. Du, and Y. Qiao, "Object-scene convolutional neural networks for event recognition in images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 30–35.

[74] L. Wang, Z. Wang, S. Guo, and Y. Qiao, "Better exploiting os-cnns for better event recognition in images," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2015, pp. 45–52.

[75] S. T. Kim, D. H. Kim, and Y. M. Ro, "Spatio-temporal representation for face authentication by using multi-task learning with human attributes," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2996–3000.

[76] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *13th European Conference on Computer Vision (ECCV)*, 2014, pp. 94–108.

[77] S. Li, Z.-Q. Liu, and A. B. Chan, "Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network," *International Journal of Computer Vision (IJCV)*, vol. 113, no. 1, pp. 19–36, 2015.

[78] Z. Kang, K. Grauman, and F. Sha, "Learning with whom to share in multi-task feature learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011, pp. 521–528.

[79] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[80] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *26th Annual Conference on Neural Information Processing Systems (NIPS) 2012.*, 2012, pp. 1106–1114.

[81] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[82] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[84] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[85] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2014, pp. 487–495.

[86] H. Lee, H. Kwon, R. M. Robinson, W. D. Nothwang, and A. M. Marathe, "Dynamic belief fusion for object detection," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.

[87] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.