2013

The joint effect of image blur and illumination distortions for Mobile Visual Search of print media

Yi Cao
University of Wollongong, yc833@uowmail.edu.au

Christian Ritz
University of Wollongong, critz@uow.edu.au

Raad Raad
University of Wollongong, raad@uow.edu.au

Publication Details

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
The joint effect of image blur and illumination distortions for Mobile Visual Search of print media

Abstract
Mobile Visual Search (MVS) is an emerging area of research given the explosion of smart and powerful mobile devices. Typically, the performance of MVS applications is influenced by the implemented image processing algorithms as well as various distortions that occur when capturing images by the mobile camera. This paper examines the joint effect of two common distortions, namely the illumination changes and image blurring on image matching accuracy for print media when using four state-of-the-art local feature algorithms. Results obtained for a database of real camera images captured by two different camera models show that the illumination changes have a more negative effect on matching accuracy compared to image blurring and this influence is camera-dependent.

Keywords
era2015, image, blur, illumination, distortions, mobile, joint, visual, effect, search, print, media

Disciplines
Engineering | Science and Technology Studies

Publication Details

This conference paper is available at Research Online: http://ro.uow.edu.au/eispapers/1948
The Joint Effect of Image Blur and Illumination Distortions for Mobile Visual Search of Print Media

Yi Cao, Christian Ritz, Raad Raad
ICT Research Institute/School of Electrical Computer and Telecommunications Engineering
University of Wollongong
Wollongong, NSW, Australia, 2522
ye833@uowmail.edu.au, critz@uow.edu.au, raad@uow.edu.au

Abstract—Mobile Visual Search (MVS) is an emerging area of research given the explosion of smart and powerful mobile devices. Typically, the performance of MVS applications is influenced by the implemented image processing algorithms as well as various distortions that occur when capturing images by the mobile camera. This paper examines the joint effect of two common distortions, namely the illumination changes and image blurring on image matching accuracy for print media when using four state-of-the art local feature algorithms. Results obtained for a database of real camera images captured by two different camera models show that the illumination changes have a more negative effect on matching accuracy compared to image blurring and this influence is camera-dependent.

Keywords—mobile visual search, image feature matching, joint distortion

I. INTRODUCTION

Mobile Visual Search (MVS) applications use image processing technologies to recognize captured images by smart phone cameras and then retrieve relevant multimedia contents from a remote server. With these emerging MVS applications, content-based image retrieval (CBIR) schemes are widely employed. There are several typical scenarios [1]. These include transmitting captured images to a remote server where the entire matching process is performed, extracting local features on the mobile device and then transmitting them to the server for matching or extracting features locally on the mobile device and matching to a local database of a given image dataset. Regardless of the scenario employed, feature extraction and matching to find the candidate visual content in a given dataset are the key tasks.

The discrimination of specific image features will be affected by distortions that occur when capturing and transmitting images as shown in Fig. 1. These distortions include the optical distortions such as lighting changes and blurring, geometric transformation such as rotation and zoom, and image compression. Therefore, the matching accuracy is influenced by the robustness of feature algorithms against these distortions. A lot of research have been conducted to develop robust local feature detectors and descriptors in computer vision like Scale-Invariant Feature Transform (SIFT) [2], Maximally Stable Extremal Region (MSER) [3], Oriented Fast and Oriented BRIEF (ORB) [4], Fast Retina Key-point (FREAK) [5], adaptive and generic accelerated segment test (AGAST) [6].

The SIFT feature is robust to scale and rotation changes while it requires high computation and memory resources due to the 128 dimension feature vectors. The ORB feature is based on the Features from Accelerated Segment Test (FAST) detector and has better performance than FAST by adding a fast and accurate orientation component to FAST. It achieves computational-efficiency, rotation invariance and robustness to image noise. The FREAK feature is the latest binary descriptor. It is fast to compute, more compact, and robust to scale rotation and noise. The MSER feature is robust to view changes in edge remarkable datasets such as buildings. The AGAST feature is also developed from the FAST detector and estimates corners in the image by finding an optimal tree [6].

However, the existing research has only focused on the repeatability, distinctiveness, and efficiency of feature algorithms used in visual search under a specific single distortion. In practical mobile applications, joint distortions will occur when capturing an image: 1) illumination changes due to the ambient lighting condition; 2) image blurring due to object motion or camera motion and out-of-focus; 3) rotation and scale; 4) partial occlusion. These distortions combine together to affect the matching accuracy in visual search. The geometric distortions such as rotation and scale, and partial occlusion are easier to overcome if efficient instructions are provided to users when they are capturing images. However, the illumination changes and blurring are harder to avoid when capturing images from mobile phone cameras because of the uncontrolled lighting environment, poor shooting quality, and limited inbuilt image enhancement functionality and quality of the mobile phone camera.
In this paper, the effect of joint illumination changes and blurring are studied with several state-of-art image feature algorithms. Section 2 provides a problem description of joint distortion. Section 3 describes the image matching method and evaluation criterion. Section 4 presents the simulation results and analysis. Section 5 concludes the paper and provides suggestions for future work.

II. PROBLEM FORMULATION

The target application in this paper is to use a mobile phone camera to capture print media and then find the matching candidate visual content in a remote server. The joint effect of illumination change and blurring applied to print media is analyzed.

A. Effect of illumination changes

In the target application with a stable light source, in the applications, the illumination will not dramatically change across a specific capture scene, no matter what kind of capture modes are used (i.e. still image mode or video mode). Thus, according to the study in [7], it can be assumed that the trichromatic (RGB) color \( I(x, \lambda) \) of pixel \( x \) of a certain image or frame under a certain wavelength \( \lambda \) is computed as:

\[
I(x, \lambda) = W_d(x)S_d(x, \lambda)E(\lambda) + W_s(x)E(\lambda) \quad (1)
\]

where \( W_d(x) \) and \( W_s(x) \) are geometric parameters of the diffuse and the specular reflections resulting from the 3D surface variations, \( S_d(x, \lambda) \) is the diffuse reflectance function and \( E(\lambda) \) is the spectral energy distribution function of the illumination. In the target application, assuming a single illuminant, \( \lambda \) can be approximately treated as invariant across the surface of the print media. In addition, the common print media examined in this work has flat surfaces, such as book covers. Therefore, \( W_d(x) \), \( W_s(x) \) and \( S_d(x, \lambda) \) can be treated as constant for the print media investigated in this work. For the image sensors used in a mobile phone camera, the output resulting from \( I(x, \lambda) \) at each sensor (for each pixel) can be modeled as follows [8]:

\[
Q_{ij}(x, t) = \int_0^t (I_{ph}(t) + i_{dc})dt + U(t) + V(t) + C \quad (2)
\]

where \( Q_{ij}(x, t) \) is the output of pixel at the end of exposure time \( t \), \( I_{ph}(t) \) is the response of \( I(x, \lambda) \), \( i_{dc} \) is the reference signal, \( U(t) \) is shot noise, \( V(t) \) is the readout circuit noise and \( C \) is reset noise. Because of the high quality CMOS image sensors are typically used in current mobile phone cameras and this paper, these last three types of noise can be assumed minimal and insignificant. Therefore, according to equation (1) and (2), when the ambient lighting condition is changing, \( I(x, \lambda) \) varies as a function of \( E(\lambda) \), this in turn induces a shift of the histogram of image intensities (i.e. increase/decrease the amplitude of all pixels by a certain magnitude) unless \( Q_{ij}(x, t) \) reaches \( Q_{sat} \). We consider such influence as the global illumination changes applied to print media. Because of camera saturation, this affects the surfaces to certain extent [2]. After the pixel level analog-to-digital converter (ADC), the output of the CMOS image sensor \( Q_{ij}(x, t) \) is digitized to \( F \) which represents the digital image output. Theoretically, there is quantization loss in the process of ADC. However, such quantization loss becomes a minor influence because of the high precision ADC (more than 8 bit ADC [9], [10], [11]) used in contemporary digital cameras. Therefore, this influence is not considered in this paper.

The global illumination changes normally affect the performance of feature detectors and feature descriptors. The global illumination variation will influence the detected keypoints in terms of their position and number. It will also decrease the matching accuracy of feature descriptors (i.e. the correct feature matching pairs decrease with increasing distortion), although several local descriptors employ local illumination normalization to achieve robustness against illumination changes [12][13].

B. Effect of image blurring

Image blurring can generally be categorized as: 1) out-of-focus blurring; and 2) motion blurring. The former blurring can be caused by user’s amateur operation or autofocus of the camera while the latter blurring can be caused by camera and object motion. In the target applications, the object and camera are relatively still or have minor motion and hence blurring caused by out-of-focus is the main consideration in this paper. The image blurring can be modeled as [14], [15]:

\[
G = H \ast F + N \quad (3)
\]

where \( H \) is the blur function representing the above mentioned types of blur and where matrix \( G, F, N \) and * represent the blurred image, original image, noise and the convolution operation, respectively. For out-of-focus blurring, \( H \) can be approximated using a Gaussian kernel [15], [16]. The additive noise, \( N \) is typically considered to have a zero mean and white distribution and orthogonal to the original image [16] and can hence be neglected. The image blurring has a great effect on image edges sharpness. In some cases, for a given set of images, the edge information will disappear and the sharpness will be mitigated. As a consequence, since the most common content-based feature keypoint detectors used in image recognition and explored in this paper rely on image edge information, the number and position of the keypoints detected by these detectors will also be influenced and the distance between feature descriptors will be affected as well.

C. Joint lighting variation and blurring model

Based on the effects of lighting changes and image blurring as discussed in Section 2.1, 2.2, when the global illumination changes \( \Delta I(x, \lambda) \) during a certain exposure time, the response \( i_{ph}(t) \) has corresponding changing \( \Delta i_{ph}(t) \) . Thus, the illumination changes can be described as:

\[
Q_{ij}(x, t) = \int_0^t (I_{ph}(t) + i_{dc})dt + U(t) + V(t) + C \leq Q_{sat} \quad (4)
\]
Where \( i_{ph}(t) = i_{ph}(t) + \Delta i_{ph}(t) \), \( \Delta U(t) = \bar{U}(t) - U(t) \), \( \Delta V(t) = \bar{V}(t) - V(t) \).

Thus, the digital output of \( Q_{ni}(x, t) \) is defined as:

\[
\hat{F} = F + \delta
\]

(5)

Where the \( \delta \) is the digital approximation of the summation of \( \int_0^t \Delta i_{ph}(t) dt, \Delta U(t) \) and \( \Delta V(t) \).
Table 1. Summary of local feature algorithms

<table>
<thead>
<tr>
<th>Keypoint detection</th>
<th>Descriptor extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGAST</td>
<td>FREAK</td>
</tr>
<tr>
<td>MSER</td>
<td>SIFT</td>
</tr>
<tr>
<td>ORB</td>
<td>ORB</td>
</tr>
<tr>
<td>SIFT</td>
<td>SIFT</td>
</tr>
</tbody>
</table>

When the out-of-focus blur occurs simultaneously, the joint effect of illumination changes and blurring can be modeled as [15], [16], [17]:

$$\mathbf{g} = \mathbf{h} * \mathbf{x} = \mathbf{h} * (\mathbf{f} + \delta)$$

(6)

We use this equation to generate the distorted images in our experiments.

III. MATCHING METHOD

This section describes the feature matching method based on clustering to find the closest matching query and candidate image pairs in a given dataset. The whole matching method can be divided into two steps. Firstly, a cluster is found in a query image and then matched to a cluster derived from a given candidate dataset to find a corresponding matching pair. Secondly, an image pair-wise matching and Geometric Verification (GV) are performed to validate the retrieved image.

A. Keypoint clustering in the query image

Keypoint clustering is motivated by the fact that the keypoints of an object in an image have the tendency to cluster [18], [19], [20]. Such clustered keypoints are typically more useful for matching than isolated keypoints. In addition, isolated keypoints have a great possibility to be outlier keypoints because the feature descriptors generated from these isolated keypoints often produce incorrect feature matching pairs [18]. Thus, using keypoint clustering on the query image is helpful for removing these isolated keypoints resulting in improved matching accuracy as well as faster performance. The Mean Shift algorithm is used for clustering in this paper. For a given query image, the n detected keypoints are $X = \{x_1, x_2, \ldots, x_n\}, x_i \in \mathbb{R}^2$, where $x_i$ is the coordinate of the $i$-th keypoint. Then the Mean Shift algorithm which can be modeled as a nonparametric kernel density estimator [21] is performed on $X$ to find the clusters:

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K \left( \frac{x-x_i}{h} \right)$$

(7)

where $h$ is the bandwidth (i.e. kernel size), $d$ is the dimension of variable (i.e. coordinate of keypoint), $K$ is the Gaussian kernel function. In our experiment, the kernel size, $h$, is set to 20 pixels: 1) It is $1/20$ of the resolution of the processed image; 2) the histogram of detected region size is mainly distributed around 20 [13]. This process results in $k$ clusters, and every $x_i$ is indexed with corresponding closest cluster center $\bar{x} = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k\}, \bar{x}_k \in \mathbb{R}^2$ as judged by the mean squared error.

The keypoints around the cluster center have a large probability of belonging to the same object [1][17]. Therefore, a 2-D location histogram of $X$ is used to study the neighboring relationship around each cluster center [1]. The image is divided into spatial bins and the number of keypoints is counted within each spatial bin. The bin containing the cluster center is also located. In the experiment, to provide sufficient keypoints for GV, the size of the bin is set to 4 pixels and the bin with less than 3 keypoints is filtered because 3 keypoints are the minimum requirement for GV [2]. Then, the adjacent bins around the bin containing the cluster center are reserved. The keypoints within these bins are merged to form the final keypoint set $\bar{X}$, which is then used to generate local feature descriptors for feature matching. The size of $\bar{X}$ varies for different images.

B. Feature training in candidate image datasets and K-Nearest Neighbor (KNN) search

Comparing the query image against each image in a large image dataset using exhaustive pair-wise feature matching is time-consuming and unsuitable for mobile visual search applications. To speed up the feature matching, the Fast Library for Approximate Nearest Neighbors (FLANN) is employed to train the extracted descriptors from candidate image datasets and then get the FLANN index using a KD-tree [22]. The KNN search is performed by using this index to find the nearest neighbor for each query descriptor. Here, $K$ is set to 1. After examining the matching pairs, the candidate images are ordered by the number of matching pairs. The first 5 images are chosen for further pair-wise matching and GV.

C. Pair-wise matching and GV

A cross-check method [22] is employed to perform pair-wise image matching after KNN. This method only returns matching pairs $(i,j)$ where the $i$-th query descriptor is nearest for $j$-th descriptor in the matcher’s collection and vice versa. This method does not require a threshold of Euclidean distance or hamming distance and can also improve the matching accuracy. The equation is shown below, $F_{A\rightarrow B}$ is the matcher set from image A to B, $F_{B\rightarrow A}$ is the matcher set from image B to A. $M_{AB}$ is the intersection of $F_{A\rightarrow B}$ and $F_{B\rightarrow A}$:

$$M_{AB} = F_{A\rightarrow B} \cap F_{B\rightarrow A}$$

(8)
After the cross check method, the RANSAC algorithm is used for GV [23]. The equation of RANSAC is shown below:

\[ ||\text{dstpt}_i - (M \cdot \text{srcpt}_i)|| > T_d \]  \hspace{1cm} (9)

\( \text{dstpt}_i \) is the coordinates of the points in the query image, \( \text{srcpt}_i \) is the coordinates of the points in the candidate image, \( M \) is the matrix which represents perspective transformation such as rotation [2], \( T_d \) is the maximum allowed re-projection error to treat a point pair as a true positive. \( T_d \) is set as 3 in the experiment which distinguishes the major inliers from outliers across the given image dataset. The evaluation criterion for matching result is Mean Average Precision (MAP) score [24]:

\[ \text{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} p(q) \]  \hspace{1cm} (10)

\( p(q) = 1 \), if the matched image is correct (as judged by ground truth image labels), otherwise 0. \( Q \) is the number of query images.

IV. SIMULATION RESULT AND ANALYSIS

A. Image dataset description

A subset of the Stanford Mobile Visual Search (MVS) dataset [25] was created containing 291 clean versions of candidate images to be matched, which consist of 100 book cover images, 100 DVD cover images and 91 museum painting images. The MVS dataset also contains corresponding query images captured by Motorola Droid mobile phone camera (5 MP) under varying indoor lighting conditions. However, the lighting condition will have more variation when outdoors. Therefore, the joint distortion model introduced in Section 2 was used to generate more distorted images to study the relative performance of different feature extraction algorithms from slight distortion to severe distortion. Although each image in the data set was not taken under the same conditions, they still form a baseline set that we can adjust using the proposed distortion model. A second distorted dataset was also created using the images captured by the higher quality Canon PowerShot G11 digital camera (10 MP) to compare with Motorola Droid images.

The following distortions reflecting the joint blurring and lighting changes were added: 1) blur: images are filtered with Gaussian blur of kernel sizes ranging from 1 to 13. The interval is 2. Increasing size increases image blurring. 1 means no image blurring. If the image is blurred using kernel size more than 13, the detected keypoints do not change significantly or just disappear; 2) Illumination changes: 6 different levels of illumination changes ±50, ±100, and±128 are added to the pixel values of the images. These manipulations result in 28,518 different distorted query images.

B. Simulation result

The experiment is conducted on the aforementioned datasets by using the matching method of Section 3. Four local feature algorithms are evaluated as shown in Table 1 that indicates different combinations of keypoint-descriptor pairings labeled here as AGASTFREAK, MSERSIFT, ORBORB, and SIFTSIFT. The number of features detected from images prior to clustering is set at least 300, which has shown to ensure accurate matching results [26], [27]. Fig. 2 shows the performance of various local feature algorithms across the entire query dataset.

The matching accuracy of AGASTFREAK on query images captured by Motorola Droid and Canon have two major differences: 1) the images captured by Droid camera have better matching accuracy at the positive illumination change compared to negative change while the images captured by Canon camera have opposite trend; 2) the maximum matching accuracy reduction with illumination change for Droid images is nearly 35% while it is nearly 49% for Canon image; 3) the maximum matching accuracy reduction with blurring change for Droid images is approximately 13% while it is approximate 7% for Canon images.

The matching accuracy of MSERSIFT shows symmetrical distribution with the positive or negative shift of the illumination for Droid images while the Canon images have better matching accuracy at the negative illumination change compared to positive change. The maximum matching accuracy reduction of MSERSIFT for Droid images is nearly 25% following illumination change while it is approximate 32% for Canon images. The matching accuracy of MSERSIFT fluctuates with blurring change both for Droid images and Canon images. The maximum variation is nearly 8%. Blurring can slightly improve the matching accuracy under certain degree of blurring in different illumination.

The matching accuracy of SIFTSIFT and ORBORB is better at the positive illumination shifts compared to negative shifts for Motorola Droid while the images captured by Canon camera still show opposite trends. The maximum matching accuracy reduction of ORBORB and SIFTSIFT is about 32% and 38%, respectively, for Droid images while it is 23% and 31% respectively for Canon images. Blurring has less effect on SIFTSIFT and ORBORB compared to illumination change where the matching accuracy only varies nearly 3%.

From the curve and the value of variation of matching accuracy, the joint distortion of illumination changes and blurring all have influences on the matching accuracy for studied feature algorithms. Additionally, the illumination variation influences matching accuracy more than blurring, especially for severe distortion conditions. The severe darkness often occurs when the ambient light is too dark while the severe brightness often happens when overexposure or the ambient light is too bright. Therefore, image enhancement technologies may be employed in such situations or proper instruction (e.g. displaying color histogram) can be given to help users capture good quality images. In addition, different cameras have different optical performance which also influences the matching accuracy of feature algorithms under different joint distortion.

Results in Fig. 3 show the matching accuracy of different print media under different joint distortion. It shows that different feature algorithms have varying matching accuracy for different print media under joint distortion and different cameras. AGASTFREAK achieves better matching accuracy for book cover in the Canon images compared with Droid images. All four feature algorithms have better performance for DVD cover in Droid images compared to Canon images.
ORBORB has better matching accuracy than SIFTSIFT for museum painting in Canon images while ORBORB and SIFTSIFT have the similar performance for museum painting in Droid images. MSERSIFT has slightly better matching accuracy for book cover and DVD cover than ORBORB while MSERSIFT and ORBORB are chosen, it is better to use MSERSIFT for DVD cover and book cover while ORBORB is better for museum painting. Future work will study the issues of more comprehensive joint distortions to identify the most significant negative effects and how to overcome these effects in real time mobile applications.

V. CONCLUSION

This paper evaluates the joint effect of illumination changes and image blurring in the context of a mobile visual search application. The performance of four local feature algorithms is investigated under various joint distortions with two different cameras from the aspect of matching accuracy. Illumination changes have more influence on matching accuracy compared to image blurring for the studied local feature algorithms under tested image datasets. Different cameras also influence the performance of local feature algorithms. Thus, flexible feature selection or combinations may be required to improve the matching accuracy for a specific application within heterogeneous camera phone environment. For example, if the MSERSIFT and ORBORB are chosen, it is better to use MSERSIFT for DVD cover and book cover while ORBORB is better for museum painting. Future work will study the issues of more comprehensive joint distortions to identify the most significant negative effects and how to overcome these effects in real time mobile applications.

ACKNOWLEDGMENT

This work is supported by the Smart Services CRC, Sydney, Australia. The authors also thank the creators of Stanford Mobile Visual Search dataset that are used in this work.

REFERENCES