Quality of experience-based image feature selection for mobile augmented reality applications

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

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

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

Publication Details
Quality of experience-based image feature selection for mobile augmented reality applications

Abstract
Mobile augmented reality applications rely on automatically recognising a visual scene through matching of derived image features. To ensure the Quality of Experience (QoE) perceived by users, such applications should achieve high matching accuracy meanwhile minimizing the waiting time to meet real-time requirement. An efficient solution is to develop an effective feature selection method to select the most robust features against distortions caused by camera capture to achieve high matching accuracy whilst transmission and matching process of the features are significant reduced. Feature selection is also beneficial to reducing the computational complexities of the matching system so that waiting time is minimized and hence user QoE is maximised. In this paper, a QoE estimation for state-of-the-art feature selection in MPEG-7 CDVS based on waiting time and matching accuracy as judged by retrieval experiments on a realistic image dataset with real-world distortions caused by image capture is analysed. The predicted QoE results suggest that feature selection can provide good QoE to users.

Keywords
selection, quality, mobile, applications, augmented, reality, image, experience, feature

Disciplines
Engineering | Science and Technology Studies

Publication Details

This conference paper is available at Research Online: http://ro.uow.edu.au/eispapers/3526
Quality of Experience-Based Image Feature Selection for Mobile Augmented Reality Applications

Yi Cao, Christian Ritz, Raad Raad
ICT Research Institute/School of Electrical Computer and Telecommunication Engineering
University of Wollongong
Wollongong, Australia
yc833@uowmail.edu.au, critz@uow.edu.au, raad@uow.edu.au

Abstract—Mobile augmented reality applications rely on automatically recognising a visual scene through matching of derived image features. To ensure the Quality of Experience (QoE) perceived by users, such applications should achieve high matching accuracy meanwhile minimizing the waiting time to meet real-time requirement. An efficient solution is to develop an effective feature selection method to select the most robust features against distortions caused by camera capture to achieve high matching accuracy whilst transmission and matching process of the features are significant reduced. Feature selection is also beneficial to reducing the computational complexities of the matching system so that waiting time is minimized and hence user QoE is maximised. In this paper, a QoE estimation for state-of-the-art feature selection in MPEG-7 CDVS based on matching accuracy as judged by retrieval experiments on a realistic image dataset with real-world distortions caused by image capture is analysed. The predicted QoE results suggest that feature selection can provide good QoE to users.

Keywords—Quality of Experience, feature selection, matching accuracy, waiting time

I. INTRODUCTION

Mobile Augmented Reality (MAR) applications are emerging to bring new augmented experiences to the users. The MAR applications targeted in this paper enhance a user’s experience by linking printed media to digital content such as a video, picture gallery or webpage [1]–[3]. A user hovers over a printed image such as an image in a newspaper with mobile device camera and the process then processes the captured scenes to recognize the image or stream content predefined in the server. Printed images that can trigger augmented content are typically pre-selected by the publisher and indicated on the page. Such applications are an alternative to using a Quick Response (QR) code, which publishers find visually intrusive if placed near a story on a newspaper or magazine article.

Figure 1 shows the workflow of the whole process. When a user scans an image using a mobile phone camera, the captured scenes are processed on a frame-by-frame basis. The image features are detected and extracted from each frame and then transmitted to a remote server. At the server side, feature matching is performed to find a corresponding image which is linked to related multimedia contents (e.g. video, webpage).

Two Key Impact Factors (KIFs) influence the user Quality of Experience (QoE) that are matching accuracy (i.e. the retrieved content is corresponding to the captured scenes) and waiting time. Generally, the capture and query process induces a user to experience a certain waiting time, which is critical to the user perceived QoE. During this procedure, the waiting time is mainly determined by several factors. Firstly, the most time-consuming part of the process is the analysis of the captured image by local feature algorithms. Here, we assume this is performed at the client side on the mobile device and the speed of this process highly depends on the employed algorithms and the computational capacity of the device. Secondly, the process of searching and matching relevant media content in the server is another time-consuming procedure depending on the retrieval method. Thirdly, the transmission capacity of the wireless network may be limited. As the accuracy of feature algorithms and matching methods increase, so too does the computational complexity, memory resources and transmission data sizes that are required [4]. As a result, the processing time and transmission time both increase. However, users are unlikely to accept too long a waiting time [5][6][7]. In addition, a longer waiting time can cause a user to become anxious and move the camera to an inappropriate position and capture irrelevant content that will not match one of the predefined images and then exacerbates the problem. It is a dilemma to achieve high matching accuracy meanwhile keeping waiting time as fast as possible in the targeted applications. Therefore, the tradeoffs of deploying feature algorithm, transmission data size and matching algorithm are required to be carefully designed to ensure that the QoE perceived by users is maximized. Thus, a guideline is needed to help application developers to consider these tradeoffs within the scope of QoE when designing and deploying MAR applications.

To solve this fundamental problem, one efficient solution is to select the most robust features as few as possible. Such features should be robust against complex capture distortions and can achieve high matching accuracy in the matching system. In the other hand, reducing the amount of features can directly alleviate the transmission load and the computational complexities of the matching system (i.e. reduce transmission time $T_{\text{trans}}$ and matching processing time $T_{\text{match}}$). The feature selection time $T_{\text{sel}}$ should also be as fast as possible.

Existing research into feature selection have been done in
an on-going MPEG-7 standardization known as Compact descriptor for Visual Search (CDVS) which aims to develop a high matching accuracy and low transmission solution for visual search applications [8]–[10]. But, how the feature selection influences the QoE is still a question. The selection criterion is crucial and must be well designed under the principle of maximizing QoE. An inappropriate selection criterion will degrade the matching accuracy while keeping more features than necessary will cause undesired waiting time and resulting in significant QoE decline [11]. Therefore, to evaluate the QoE for feature selection is desired.

In this paper, A QoE estimation considering matching accuracy and waiting time is presented based on the assumption of a Bernoulli process. The rest of the paper is organized as follows: Section II describes the state-of-arts feature selection method in MPEG-7 CDVS and Section III shows the matching accuracy of retrieval results with feature selection. Section IV presents the QoE estimation of feature selection methods for the targeted applications and the conclusions are provided in Section V.

II. FEATURE SELECTION FOR SIFT FEATURE

In this section, the state-of-art SIFT feature selection methods in MPEG-7 CDVS are reviewed. The methods focused on measuring how well their outputs related to the most significant features which are important for matching between image pairs. Therefore, the relevance of the output parameters of the SIFT feature detector and the correctly matched SIFT features [8][9] is investigated.

The assumption of the methods is that the correctly matched feature pairs are different from wrongly matched feature pairs within an image pair. Such differences are characterized during the feature detection and implied in the output of the feature detector. The output parameters including the Difference-of-Gaussian (DOG) response $\theta_{\text{peak}}$ (denoted as peak in the following paragraphs), scale $\theta_{\text{scale}}$, orientation $\theta_{\text{orientation}}$, location $\theta_{\text{distance}}$ (the distance from the feature location to the image center) are evaluated individually to investigate the relevance score of these quantities to correctly matched pairs [12] as well as their combination [13] using the probability mass function of correctly matched features learned from dataset. Then, the features are selected on the basis of a relevance score. The peak of the output of the SIFT detector is superior for identifying the most relevant features compared to other parameters of the output of SIFT detector, including $\theta_{\text{orientation}}, \theta_{\text{scale}}, \theta_{\text{distance}}$ [12]. Thus, the feature selection method using $\theta_{\text{peak}}$ is chosen to investigate in this work.

The $\theta_{\text{peak}}$ is related to the local contrast in the image around a feature point and is used as the criterion for detection of SIFT feature in [14]. A feature point with low peak value indicates a low contrast local image region and hence discarded in the feature detection process. It is intuitively that it can be a good indicator for feature selection in terms of discrimination and stability. The procedure to learn the relevance function of $\theta_{\text{peak}}$ related to correctly feature matches is summarised as follows [12].

The learning process proceeds automatically on image pairs (i.e. both images in the pair contain the same object) by performing pairwise image matching:

1. Detect the SIFT features and extract the SIFT descriptors both from images in a pair in the database. For each feature, the peak value $\theta_{\text{peak}}$ is recorded from the output of feature detection;
2. Match of detected features within an image pair. The correctly matched features are selected for processing in step 3;
3. Perform Geometric Verification using RANSAC [15] and the remaining features are taken as true positive features (i.e. correctly matched features) and label as $c=1$. The other detected features are label as $c=0$;
4. Calculate the probabilities on the true positive features using Equation (1).

$$p(c = 1|\theta_{\text{peak}} \in G) = \frac{h(c=1|\theta_{\text{peak}} \in G)}{h(\theta_{\text{peak}} \in G)}$$  \hspace{1cm} (1)$$

where $G$ is region of $\theta_{\text{peak}}$. $h(.)$ calculates the histogram of $\theta_{\text{peak}}$ using $N$ bins ($N=25$) [12].

To test the effectiveness of feature selection for the mobile augmented reality applications targeted in this work, the printed media images from the MVS dataset [16] are used. This dataset contains more than 1200 camera-phone captured different types of print images including CD covers, DVD covers and book covers. These images are denoted as query images. The dataset has several key characteristics that reflect realistic situations: rigid objects, widely varying lighting conditions, perspective distortion, foreground and background clutter, and query images collected from heterogeneous low and high-end camera phones. The ground-truth reference images are also available and used for training. These ground-truth images are denoted as reference images. The learned probability of $\theta_{\text{peak}}$ related to correctly feature matches is
shown in the Figure 2. From the Figures 2, it is clear that \( \theta_{peak} \) is effective for selecting the features as it exhibits a distinctive distribution. But, the probability of \( \theta_{peak} \) of correctly matched features in the query images is different from the probability of \( \theta_{peak} \) of correctly matched features in the reference images. The correctly matched features in the query images are not correspondingly matched to features in the reference images at the same bin of \( \theta_{peak} \) due to the complicated distortion occurred in the query images of MVS dataset. Therefore, the matching accuracy of retrieval results will be influenced by feature selection and such effects on retrieval are studied in Section III.

After assigning the probability to each feature based on these distributions, the features are ranked from high probability to be matched to low probability. The feature sets can be easily filtered on the basis of ranked features using a feature number threshold according to different bitrate requirements (i.e. the available bit rate is related to the maximum number of features that can be transmitted). Figure 3 shows an example of using \( \theta_{peak} \) to select 210 SIFT features from 614 SIFT features generated for this image. The features selected by \( \theta_{peak} \) are mainly located in parts of the face and the text under the face. Many false features are filtered out.

### III. RETRIEVAL EXPERIMENTAL RESULT

In contrast to the results presented in [8][9][13] which focused on measuring the performance of image pairwise matching by using feature selection, The retrieval results presented in this section tackle the matching accuracy of retrieving corresponding content from server by using feature selection for targeted MAR applications. The image pairwise matching only verifies whether two images depict the same object or scene. The retrieval performs search and discovery of the images that depict the same object and scene as the query image within a large collection. The experimental procedure of retrieval is as follows:

1. For each query image in the dataset:
   a. Detect and extract the SIFT feature;
   b. Select the specified number of features using the proposed feature selection method. This forms the query feature set with the remaining features filtered out;
2. For the reference images in the dataset:
   a. Detect and extract the SIFT feature for each reference image;
   b. Combine the detected SIFT features of each reference image to set up the training feature set;
   c. Perform KD-tree training to obtain the reference feature search space.
3. For each query feature set:
   a. Perform the nearest neighbor search using KNN \((k=1)\) for each query feature in the trained reference feature search space;
   b. Obtain the first \(N \) (\(N=3\)) reference images with maximum feature matching pairs (Increasing \(N\) did not bring out significantly better retrieval results in the experiment);
   c. Perform cross-check KNN \((k=1)\) search within each chosen reference image to further filter the features;
   d. Apply geometric verification (RANSAC) to find the final true positive feature matching pairs.
   e. Locate the reference image on the basis of the highest number of true positive feature matching pairs;
   f. Declare a correct match using a ground truth file list.

The matching accuracy is evaluated based on the Mean Average Precision (MAP) to judge the retrieval performance [8], [18] under different bitrate (i.e. different feature number after selection):

\[
MAP = \frac{1}{Q} \sum_{q=1}^{Q} P(q)
\]

\[
P(q) = \begin{cases} 1, & \text{the matched image is correct} \\ 0, & \text{otherwise} \end{cases}
\]
For comparison, the retrieval experimental results of using Peak-based Feature Selection (PFS) in MPEG-7 CDVS and random feature selection for SIFT feature are presented in Figure 4. The random feature selection generates a random feature index list to choose features (i.e. randomly choosing a certain number of features without any criteria). Four different feature number conditions are considered in the experiment 279, 210, 114 and 50 which correspond to 2KB, 1KB, 512B and 256B compressed feature transmission sizes. The first three bit rates are standardized in the MPEG-7 CDVS [10]. The fourth bit rate is also considered in the scenario of a very poor communication condition or processing condition where a very fast transmission is desired (e.g. processing a stream of video frames to repeatedly look for a matching reference image).

From Figure 4, it is evident that PFS achieves good MAP for retrieval under varying feature number. The MAP without feature selection is 95.68%. PFS achieves only a 1.6%, 3.7%, 9.6% and 22% retrieval degradation for 279, 210, 114 and 50 features, respectively. An improper random selection method degrades the matching accuracy fast with the reduction of the number of features as shown in Figure 4. For 279 features, the random method still achieves 90% because it selects on average more than 85% of the features generated by the SIFT algorithm. (the total number of detected SIFT feature is determined by the complexity of an image).

IV. QOE-DRIVEN FEATURE SELECTION

It is important to study the influence of feature selection in terms of the QoE as the amount of features directly determines the matching accuracy and waiting time which are two KIFs of QoE for targeted applications. To our best knowledge, this work is the first attempt to study the QoE in the targeted MAR applications using feature selection in MPEG-7 CDVS. In this section, the QoE in terms of waiting time is firstly presented and then the correlation of matching accuracy and waiting time which are two KIFs of QoE as the amount of features directly determines QoE perceived by users in terms of the waiting time, a subjective test was conducted on a Samsung Galaxy Tab. An application simulated the whole procedure of targeted MAR applications were developed. 51 participants were invited to attend the test. The participants were asked to operate the mobile phone camera in video mode and then ‘scan’ over a printed picture on a page to find the corresponding matching image within a database of unique images. A matched image triggered the presentation of ‘linked’ content (e.g. a web page or video). The QoE perceived by users was judged by 5-point Mean Opinion Scores (MOS) (i.e. 1-bad, 2-poor, 3-fair, 4-good, 5-excellent). The subjective test results suggested that QoE has a logarithmic function of waiting time $t$ as shown in Equation (3). The detailed information about the subjective test can be found in [11].

$$QoE(t) = -1.118 \ln(t + 1.648) + 5.864$$  \hspace{1cm} (3)

Matching accuracy is another KIF in the targeted application however the matching accuracy is an invisible KIF as the users only know that there is related content which corresponds to the capture scenes but they do not know what the content will be in terms of the matching accuracy.
Moreover, the matching accuracy is important to the publishers as they want to guarantee that the users receive the correct augmented information. Otherwise, an irrelevant feedback would degrade the QoE perceived by users significantly.

To study the correlation between matching accuracy and waiting time, a Bernoulli process is employed to model the matching processing in the targeted MAR applications because the matching accuracy is influenced by the feature selection as described in Section III. For each captured frame, after the feature extraction and feature selection, a certain number of features are transmitted to the server to perform search and matching. Each frame has a probability to be correctly matched or not. The matching result of each frame is considered to follow an identical and independent distribution. Therefore, considering the problem of finding match (i.e. success or not) for each frame as Bernoulli trials [19], it is assumed that the probability that the first occurrence of successful match requires \( M \) number of frames, each with success probability \( P \) (i.e. the MAP shown in Figure 4). Then, the probability of finding a match after processing \( M \) frames is:

\[
P_{\text{correct}} = 1 - (1 - P)^M
\]

(4)

Therefore, the frame \( M \) which makes \( P_{\text{correct}} = 1 \) (i.e. 100% match) can be calculated by the MAP of selection method for varying feature numbers \( N \) according to Figure 4. The result of frame amount \( M \) for PFS is shown in Table 1. Then, The waiting time \( t \) of targeted applications can be defined as:

\[
t = M \ast (T_{ex} + T_{sel} + T_{trans} + T_{match}) + T_{buffer}
\]

(5)

For each frame, the feature extraction time \( T_{ex} \) is device-dependent and mainly determined by the computational capacity of the client device and the content of captured frame; the feature selection time \( T_{sel} \) are related to feature number \( N \) and the computational capacity of the client device. The feature transmission time \( T_{trans} \) is inversely proportional to transmission bitrate \( C_{\text{link}} \) for certain transmission load \( L_{\text{bit}}(N) \) (i.e. 256B, 512B, 1KB, 2KB corresponds to 50, 114, 210, 279 features, respectively). The feature matching time \( T_{match} \) is mainly determined by the computational capacity of the server and considered insignificant compare to other times assuming a powerful server with GPU acceleration in this work. The buffer time \( T_{buffer} \) is configurable according to different players and conditions [20]. To isolate the effect of buffer time, a buffer time of 0.5s is considered for streaming related multimedia contents as previous research has indicated that such waiting time provides a satisfactory QoE [21]. The frame rate \( M \), link capacity \( C_{\text{link}} \) and the feature number \( N \) are considered as key influencing factors of waiting time \( t \) in this work. A test platform with quad-core 1.6GHz CPU and 2G RAM is used to simulate the computational capacity of a current state-of-the-art smart phone [22] to estimate \( T_{ex} \), \( T_{sel} \), \( T_{match} \) varies with different feature number but the maximum variation is only 2ms. Therefore, the average selection time \( \text{avg}T_{sel} \) is used. The maximum variation of \( T_{ex} \) across the whole dataset is only 4ms. Thus, the average \( T_{ex} = 0.138ms \) is used. In practice, a cutoff threshold of waiting time \( t \) is needed when no match can be found after a certain waiting time (i.e. multiple consecutive frames from the camera ‘scan’ do no match any image in the database). In this case, the process should be stopped and feedback is given to the users. Equation (5) can be redefined as:

\[
t = \frac{M \ast \left( \text{avg}T_{sel} + \frac{L_{\text{bit}}(N)}{C_{\text{link}}} + 0.138 \right) + 0.5 \text{, a match is found}}{2 \text{, no match is found}}
\]

(6)

2s is chosen as a cutoff threshold of waiting time to ensure satisfactory QoE [11]. \( C_{\text{link}} \) ranges from 50kbps to 4800kbps.

![Predicted QoE for PFS](image_url)

Fig. 5. The predicted QoE of PFS under 50kbps-2000kbps bitrate and varying feature number.

### Table 1. The number of frame \( M \) which makes \( P_{\text{correct}} = 1 \) using Equation (4)

<table>
<thead>
<tr>
<th></th>
<th>N=50</th>
<th>N=114</th>
<th>N=210</th>
<th>N=279</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFS</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>
as in a typical 3G/4G wireless network [23]. Substituting $t$ with Equation (6) in Equation (3) when assuming a match can be found, the predicted QoE of the feature selection related to waiting time and matching accuracy (i.e. $M$ frame to achieve 100% match) can be defined as:

$$QoE(M, N, C_{link}) = \ln \left( M \times (avgT_{select} + 0.138 + \frac{ln(38)}{C_{link}} + 2.148) + 5.864 \right)$$

(7)

The predicted QoE according to Equation (7) and Table 1 is shown in Figure 5. To illustrate the key results, transmission bitrates above 2000kps were truncated as the predicted QoE results show a flat trend beyond these transmission bitrates.

The PFS can achieve good QoE results (i.e. above 4) at all transmission conditions. The transmission bitrate only has minor effect on the predicted QoE at the low transmission bitrate for feature number 210/279 because of the increased transmission load for each frame. At the high transmission condition, the QoE is slightly decreased with the reduction of feature number as the matching accuracy is decreased. It indicates that it is better to transmit more features under high transmission condition. It is not only beneficial to ensuring matching accuracy but also make the system work under low frame condition which can reduce the computation consumption. Besides, the low frame rate system can reduce the probability that users move the camera to an inappropriate position and capture irrelevant content. Another solution to avoid this problem in high frame rate condition is to perform frame selection to determine which captured frame should be processed to find a matching image but is outside the scope of this paper.

V. CONCLUSION

A QoE estimation for feature selection in MPEG-7 CDVS is discussed in this paper. Aiming to achieve the high matching accuracy for retrieval in the targeted MAR applications to maximizing the QoE perceived by users, the feature selection based on peak value in MPEG-7 CDVS is employed for selecting a subset of goodness detected features in terms of their ability to correctly match a corresponding reference image to a query image from database. The retrieval performance of PFS is verified to be excellent on a dataset with complex realistic distortions. Using a QoE estimation based on waiting time and matching accuracy, the predicted QoE proves that PFS can provide good QoE to users under varying transmission conditions. Future work may be extended to study the QoE in terms of waiting time of PFS for different application systems, such as sending compressed and selected video frames to a server for processing or maintaining a local reference dataset on the client side and performing all processing there.

ACKNOWLEDGMENT

This work is supported by Smart Services CRC, Sydney, Australia.

REFERENCES