AN EMBEDDED 3D GEOMETRY SCORE FOR MOBILE VISUAL SEARCH

Hanwei Wu, Haopeng Li and Markus Flierl

School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm
{hanwei, haopeng, mflierl}@kth.se

ABSTRACT

The scoring function is a central component in mobile visual search. In this paper, we propose an embedded 3D geometry score for mobile 3D visual search (M3DVS). In contrast to conventional mobile visual search, M3DVS uses not only the visual appearance of query objects, but utilizes also the underlying 3D geometry. The proposed scoring function interprets visual search as a process that reduces uncertainty among candidate objects when observing a query. For M3DVS, the uncertainty is reduced by both appearance-based visual similarity and 3D geometric similarity. For the latter, we give an algorithm for estimating the query-dependent threshold for geometric similarity. In contrast to visual similarity, the threshold for geometric similarity is relative due to the constraints of image-based 3D reconstruction. The experimental results show that the embedded 3D geometry score improves the recall-datarate performance when compared to a conventional visual score or 3D geometry-based re-ranking.

1. INTRODUCTION

Mobile 3D visual search introduces the concept of 3D geometric information into the search problem [1] [2] [3]. It improves the search results by assessing the actual 3D geometry when compared to conventional appearance-based 2D image methods. Specifically, it addresses scenarios in which different real 3D objects appear similar in captured images. For example, consider the case where a poster shows a picture of a real 3D object.

Recently there are a number of works which has remarkable improvements in reducing the size of image feature data and in reducing the computation footprint in searching process for mobile visual search [4] [5] [6]. For the design of scoring function, the authors of [7] introduced a co-indexing scheme that incorporates the image similarities based on local features and semantic attributes as the ranking criteria. [8] proposed a bi-layer graph structure for querying multi-modal data. In this work, we discuss how to combine the geometric information with the visual appearance of the object. We introduce an embedded 3D geometry score that improves the recall-datarate performance of the mobile search system.

For a given query, the rank of a retrieved object can be determined by its visual appearance and geometric layout similarity. Conventional mobile visual search evaluates the geometric information of the object in the geometric consistency check (GCC) step of the retrieval pipeline [9] [10] [11] [12]. The GCC serves either as a separate re-ranking of the short list of objects obtained from visual descriptor matching or as a rejection rule for outliers. Hence, the final ranking does not reflect all information of the object that can be obtained through search. In this paper, we take a different perspective to look on this problem. Inspired by the work of [13] which gives an explanation of the relation between term-frequency and inverse document-frequency (tf-idf) and mutual information, we propose to use the mutual information between query and candidate objects to determine an embedded 3D geometry score for ranking. With that, we interpret visual search as a process that reduces the uncertainty among candidate objects when observing a query. Before observing a query, we have no prior preference over candidate objects. Hence, the candidate objects on the server are equally likely to be retrieved. After observing a query, the updated object distribution conditioned on the query is obtained. The resulting mutual information between query and candidate objects will lead to the proposed embedded 3D geometry score.

For the mobile 3D visual search system, we build on our previous work [1]. We construct scalable multi-view vocabulary trees based on multi-view image features [2] [14]. Moreover, multi-view imagery is used to obtain the 3D geometric information of an object [15].

The paper is organized as follows: Section 2 summarizes the 3D feature correspondences. Section 3 discusses the geometric similarity parameters. Section 4 introduces the visual-geometric score. Section 5 discusses our experimental results.

2. 3D FEATURE CORRESPONDENCES

Using the Bag-of-Words model, let $O = \{o_1, ..., o_K\}$ be the set of candidate objects with size $|O| = K$. Let $V = \{v_1, ..., v_M\}$ be the set of visual words created by the vocabulary tree, and let $Q = \{q_1, ..., q_N\}$ be the set of query descriptors of size $N$. Each query descriptor $q_i$ is a concatenation of an appearance-based multi-view descriptor $v_i$ and...
and a 3D location \( g_i \) such that \( q_i = \left( \frac{v_i}{g_i} \right) \), where, for example, \( v_i \in \mathbb{R}^{128} \) is a SIFT-based [16] multi-view descriptor associated with the 3D location \( g_i \in \mathbb{R}^3 \).

For correctly matched feature pairs, the 3D world coordinate of object points \( \vec{w}_o^q \) in the database can be obtained by the seven parameter Helmert transformation [17] of the 3D world coordinate of query points \( \vec{g}_q^o \) according to

\[
\vec{w}_o^q = \pi(\vec{g}_q^o) = k \Phi \vec{g}_q^o + \vec{t},
\]

where \( k \) is the scale parameter in \( \mathbb{R}^+ \), \( \Phi \) the rotation matrix in \( \mathbb{R}^3 \), and \( \vec{t} \) the translation parameter in \( \mathbb{R}^3 \). The estimation of multiple parameters is time consuming and makes real-time applications impossible. There are several proposals to accelerate the geometric consistency check step. For image-based retrieval, [18] estimates parameters such as scale or orientation of local descriptors to reduce the computation. In 3D space, however, we consider the geometric parameters and is found empirically. The subsequent matching will check the geometric misalignment with respect to each object. Objects that exceed the geometric similarity threshold will be rejected. Note that due to the threshold \( J \), not all the candidate objects will have sufficient geometric similarity. Note that descriptors which indicate visual similarity without corresponding geometric similarity are less discriminative.

We assume that the correspondences follow a global transformation between two 3D coordinate systems for correctly matched objects. Further, we model the distribution of the 3D error for correct matches by the Gaussian probability density function.

\[
f_\varepsilon = \frac{1}{(2\pi)^\frac{3}{2}|\Sigma|^\frac{3}{2}} e^{-\frac{1}{2}(\varepsilon - \bar{\mu})^T \Sigma^{-1}(\varepsilon - \bar{\mu})}\]

We assume that the components \( \varepsilon_{\nu}, \nu \in \{x,y,z\} \), of the 3D error \( \varepsilon \) are i.i.d. with mean \( \mu_{\nu}, \nu \in \{x,y,z\} \). That is, the covariance matrix is diagonal \( \Sigma = \sigma^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \) and the mean 3D error is \( \bar{\mu} \). Hence, the distribution of the 3D error can be factorized as

\[
f_\varepsilon = \prod_{\nu \in \{x,y,z\}} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_{\nu} - \mu_{\nu})^2}{2\sigma^2}}.
\]

And the 3D error follows a Nakagami distribution.

**Algorithm 1** 3D Geometric Error Estimation

Initialize: Set the distance vector \( \{d_{ij}\} = 0 \) for \( i = 1, \ldots, T \) and for all objects \( j = 1, \ldots, K \).

**do** Update the distance vector by matching the incoming of descriptors of length \( T \) against the vocabulary tree;

**for all** \( o_j \in O \) **do**

\[ |M_q(q_t \in T | o_j)| > J \]

\[ m_j \leftarrow \text{median}_{i}(d_{ij}) \]

\[ \Theta_j \leftarrow \text{C/MAD}_i(d_{ij}) \]

**end if**

**end for**

2. Output \( \Theta_j \) and \( m_j \) for \( j = 1, \ldots, K \).

**4. AN EMBEDDED 3D GEOMETRY SCORE**

After obtaining 3D geometric parameters from the training sequence, we can formulate the embedded 3D geometry score as the mutual information between query and candidate objects.

\[
 I(O; V, G) = \sum_{j=1}^{K} c_j
 \]

where \( c_j \) is a constant factor that depends on the distribution of the data \( x_i \).

We estimate the median and MAD of the 3D misalignment distance \( d_{ij} := d(q_i, o_j) \) for objects whose number of visual matches exceeds a threshold \( J \). We use the reciprocal of the MAD and multiply with the constant \( C \) to obtain the geometric similarity threshold \( \Theta \) for each object \( j = 1, \ldots, K \).
We assume the process of quantizing one query descriptor \( q_i \) into a visual word \( v_i \), i.e., \( q_i \rightarrow v_i \) is equivalent to the event of random selecting a visual word from the whole set of visual words. The probability of selecting specific visual word is \( P(v_i, o_j) = \frac{f_{ij}}{F} \), where \( f_{ij} \) is the frequency of \( v_i \) associated with object \( o_j \). And \( F \) is the total frequency of all visual words in the whole set of objects. Note that the query distribution is proportional to the visual word frequency in the set of objects is very a strong assumption. However, this assumption is actually embedded in the heuristics of the tf-idf which performs well in practice.

When a \( q_i \) is quantized to a visual word \( v_i \), we can obtain the knowledge that only a subset of objects \( M_v(v_i|O) = M_v(i) \) is associated with the matched visual word \( v_i \). The number of objects in the subset is \( |M_v(i)| = K_i \), where \( K_i = \sum_o 1(n_v(o) \geq 1) \). \( n_v(o) \) is the number of visual words associated with an object, and \( I \) is the indicator function. The probability of selecting specific visual word from the whole set \( V \)

\[
H(O|v_i) = - \sum_{o_j \in M_v(i)} P(o_j|v_i) \log_2 P(o_j|v_i)
\]

\[
= -K_i \frac{1}{K_i} \log_2 \left( \frac{1}{K_i} \right) = \log_2 K_i
\]

Objects not associated with the matched visual words have zero probability. Hence, they do not contribute to the conditional entropy.

The subset of \( M_v(i) \) can be further narrowed down by considering the geometric constraint \( m_O \) and \( \Theta_O \)

\[
\|\overrightarrow{w_O} - \overrightarrow{g_i} - m_O\|_2 \leq \Theta_O
\]

where \( w_O \) are the 3D world coordinates of the objects \( O \) in the subset \( M_v(i) \). In this way, only objects that satisfy the geometric similarity threshold will remain in the set. Hence, the number of objects in the subset \( M_v(i) \) is reduced.

After using the geometric similarity threshold, the subset of objects with respect to the query \( q_i \) becomes \( M_g(v_i, g_i|O) = M_g(i) \) with size \( |M_g(i)| = L_i \); \( K_i \geq L_i \). The probability of candidate objects becomes \( P(o_j|v_i, g_i) = \frac{1}{L_i} \). Hence, the entropy of candidate objects \( O \) conditioned on both matched visual words and 3D locations is

\[
H(O|v_i, g_i) = \log_2 L_i
\]

**Fig. 2** shows the relations between sets of objects with visual and geometric constraints.

Finally, the expected mutual information between objects and both query descriptors is

\[
I(O; V, G) = H(O) - H(O|V, G)
\]

\[
= \sum_i p(v_i, g_i)(H(O) - H(O|v_i, g_i))
\]

\[
= \sum_j \sum_i \frac{f_{ij}}{F} \log_2 \frac{K}{L_i}
\]

From (11) we see that, the term \( \frac{K}{L_i} \) is similar to the inverse document frequency term. The \( \tilde{f}_{ij} \) corresponds to the term...
Fig. 2. The black dots represent objects in the server. The triangle represents one query descriptor. The figure illustrates that the visual and geometric constraints narrow down the objects associated with the query descriptor.

frequency term which is an estimation of the occurrence probability of a geometry-embedded word. The total frequency of all words $F$ is a constant factor. The embedded 3D geometry score $c_j$ for a single object $o_j$ is

$$c_j = \sum_{i=1}^{N} \tilde{f}_{ij} \log_2 \frac{K}{D_i}. \quad (11)$$

Fig. 3. The pipeline of the embedded 3D geometry score.

5. EXPERIMENTAL RESULTS

5.1. Dataset and Setup

We evaluate our embedded 3D geometry score for the multiview image dataset *Stockholm Buildings*\(^1\) which comprises 50 buildings of that city. The server holds 254 images of the 50 buildings. At least 2 views have been recorded for each building. The client may use up to 100 additional test images of the 50 buildings. We acquired server images using a Canon IXUS50 digital camera at a resolution of $2592 \times 1944$ pixels. Two sets of test images have been recorded using the Cannon camera and a SONY Xperia Z2 mobile at different viewpoints and times of a year so as to have lighting and viewpoint variations compared to the server images. An Android app can be downloaded from the project website\(^2\) for online testing.

\(^1\)http://people.kth.se/~haopeng/sthlmbuildings/

\(^2\)http://people.kth.se/~haopeng/M3DVS/index.html

Fig. 4. Comparison of the recall-datarate between different scoring schemes.

The vocabulary tree we use at the server is constructed using hierarchical multiview features. We set $D = 5$ for the number of the tree levels and $K = 8$ for the branches. the total storage of the vocabulary tree is 23.5MB compared to the 5.3GB original view by view feature database and the 400MB multi-view feature database. The recall-datarate is used to evaluate the relationship between retrieval performance and the datarate that a client sends to the server. The recall is considered successful, only if the correct object appears on the top of the ranking. The datarate is the average size of the query sent to the server.

5.2. Comparison of Scoring Functions

We test our proposed geometry-embedded score with our previous geometry-based re-reranking method and the conventional visual score. The geometry-based re-ranking method evaluates the appearance and geometric separately. We test retrieval performance on datarate that vary from 3.0 KB/query to 12.2 KB/query. The experimental results in Fig. 3 show that the recall rate increases as the data rate increases. The scores considering geometric information have better performance than the original visual score as expected. The proposed embedded 3D geometry score improves the recall-datarate in general, except at the lowest rate which is due to the short length of the used training sequence. The recall using the embedded 3D geometry score can reach over 90% at a lower datarate.

Fig. 4 shows examples of the ranking results using the embedded 3D geometry score. The image on the left is the query image, and the five images on the right show the top-
ranked objects in decreasing order from left to right. We observe that the retrieved objects share both visual and geometric similarities to the query.

6. CONCLUSIONS

We introduced an embedded 3D geometry score to reflect both visual appearance and underlying geometry of objects in the ranking score. We show that with the estimated geometric parameters, the scoring can be derived from the mutual information between query and candidate objects. The retrieval performance is improved when compared to our previous geometry-based re-ranking result. The datarate can be further reduced by applying local feature descriptor compression as standardized in MPEG-CDVS [20].

7. REFERENCES


Fig. 5. Examples of ranking results using the embedded 3D geometry score.