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 a 3D location g_i such that $q_i = \begin{pmatrix} v_i \\ g_i \end{pmatrix}$, 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 $\overrightarrow{w_o}$ in the database can be obtained by the seven parameter Helmert transformation [17] of the 3D world coordinate of query points $\overrightarrow{g_q}$ according to

$$\overrightarrow{w_o} = \pi(\overrightarrow{g_q}) = k\Phi\overrightarrow{g_q} + \overrightarrow{t}, \tag{1}$$

where k is the scale parameter in $\mathcal{R}^+$, Φ the rotation matrix in $\mathcal{R}^3$, and $\overrightarrow{t}$ the translation parameter in $\mathcal{R}^3$. 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 3D misalignment between correspondences $\|\overrightarrow{g_q} - \overrightarrow{w_o}\|_2 = ||\vec{e}(q,o)||_2 = d(q,o)$ to assess the 3D geometric similarity. Note that the 3D misalignment depends on the Helmert transformation between two 3D coordinate systems, but it is independent of the absolute location of individual points in 3D space. For details on how to determine the relative 3D world coordinates, please see our previous work [1,3].

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.

$$f_{\vec{e}} = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{e}-\vec{\mu})^T\Sigma^{-1}(\vec{e}-\vec{\mu})} \tag{2}$$

We assume that the components $e_\nu, \nu \in \{x,y,z\}$, of the 3D error $\vec{e}$ 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 $\vec{\mu}$. Hence, the distribution of the 3D error can be factorized as

$$f_{\vec{e}} = \Pi_{\nu \in x,y,z} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(e_\nu - \mu_\nu)^2}{2\sigma^2}}. \tag{3}$$

And the 3D error follows a Nakagami distribution.

Fig. 1 illustrates an example of 3D feature correspondences for a correctly matched object and the histogram of 3D misalignment. Note the distribution of the 3D misalignment.

3. GEOMETRIC SIMILARITY PARAMETERS

We exploit the observation that the 3D error of correctly matched objects follows a distribution of small variance when compared to the wide distribution of object points in the 3D coordinate space. Further, we exploit the best-feature-first policy for sending the query descriptors to the server [3]. This allows us to use the first part of query descriptor sequence as the training sequence T. This first part holds the most robust descriptors and permits us to estimate the 3D geometric parameters of the 3D misalignment robustly. Note, due to the constraints of image-based 3D reconstruction, these parameters depend on both the given query and the available objects.

In order to cope with the outliers and a small sample size problem, we use the robust statistic estimator *median absolute deviation* (MAD) [19] .

$$\text{MAD}_i(x_i) = c \cdot \text{median}_i(|\ x_i - \text{median}_j(x_j)\ |), \tag{4}$$

where c 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 $\mathcal{J}$. We use the reciprocal of the MAD and multiply with the constant C to obtain the geometric similarity threshold Θ_j for each object $j = 1, \ldots, K$. The constant factor C is used to control the precision of 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 $\mathcal{J}$, not all the candidate objects will have sufficient geometric similarity. Note that descriptors which indicate visual similarity without corresponding geometric similarity are less discriminative.

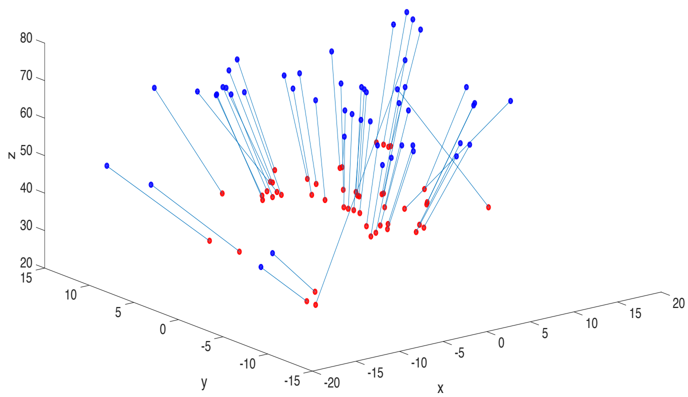
Algorithm 1 3D Geometric Error Estimation

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**
  if $|M_g(q_i \in T|o_j)| > \mathcal{J}$ **then**
    $m_j \leftarrow Median_i(d_{ij})$
    $\Theta_j \leftarrow C/MAD_i(d_{ij})$
  end if
end for
2. Output Θ_j and m_j for $j = 1, \ldots, K$.

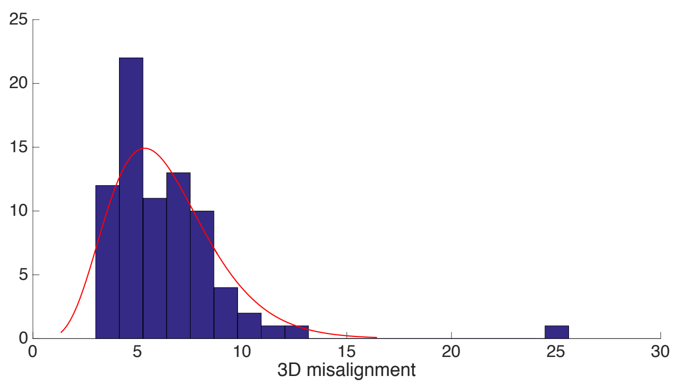
4. AN EMBEDDED 3D GEOMETRY SCORE

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.

$$I(O;V,G) = \sum_{j=1}^{K} c_j, \tag{5}$$



(a) 3D correspondences for a correctly matched object. Blue points are the 3D coordinates of the query object. Red points are the 3D coordinates of the visually matched object. The links between points show the 3D misalignment.



(b) Histogram of 3D misalignment as obtained from (a).

Fig. 1. Example of 3D misalignment.

where V is the set of visual descriptors and G the set of 3D locations of feature points. c_j is the embedded 3D geometry score of the object o_j.

Before observing any queries, we have no prior preference over candidate objects. Hence, we assume the candidates are equally likely, $P(o_j) = \frac{1}{K}$ for all $o_j \in O$. So the entropy of O is

$$\begin{aligned} H(O) &= -\sum_{o_j \in O} P(o_j) log_2 P(o_j) \\ &= -K\frac{1}{K} log_2(\frac{1}{K}) \\ &= \quad log_2 K \end{aligned} \tag{6}$$

We assume the process of quantizing one query descriptor q_i into a visual word v_i, i.e., $q_i \to v_i$ 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 $K_i = \sum_o \mathbf{1}(n_{v_i}(o) \geqslant 1)$. $n_v(o)$ is the number of visual words associated with an object, and $\mathbf{1}$ is the indicator function. The probability of candidate objects becomes $P(o_j|v_i) = \frac{1}{K_i}$. Hence, the conditional entropy of O given v_i is

$$\begin{aligned} H(O|v_i) &= -\sum_{o_j \in M_v(i)} P(o_j|v_i) log_2 P(o_j|v_i) \\ &= -K_i \frac{1}{K_i} log_2(\frac{1}{K_i}) \\ &= \quad log_2 K_i \end{aligned} \tag{7}$$

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 $m_{\mathbb{O}}$ and $\Theta_{\mathbb{O}}$

$$\|\overrightarrow{w_{\mathbb{O}}} - \overrightarrow{g_i} - m_{\mathbb{O}}\|_2 \leq \Theta_{\mathbb{O}} \tag{8}$$

where $\overrightarrow{w_{\mathbb{O}}}$ are the 3D world coordinates of the objects $\mathbb{O}$ 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 $M_g(v_i, g_i|\mathbb{O}) = M_g(i)$ 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

$$H(O|v_i, g_i) = log_2 L_i \tag{9}$$

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

$$\begin{aligned} I(O; V, G) =& H(O) - H(O|V, G) \\ =& \sum_i p(v_i, g_i)(H(O) - H(O|v_i, g_i)) \\ =& \sum_j^K \sum_i^N \frac{\tilde{f}_{ij}}{F} log_2 \frac{K}{L_i} \end{aligned} \tag{10}$$

From (11) we see that, the term $\frac{K}{L_i}$ is similar to the *inverse document frequency* term. The $\tilde{f}_{ij}$ is corresponds to the *term*

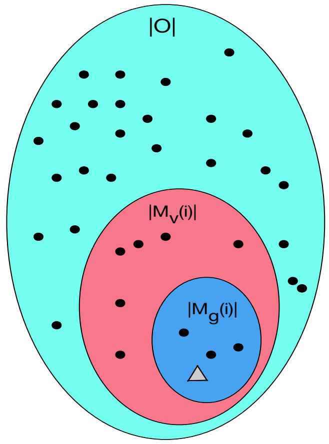


Fig. 2. The black dots represent objects in the server. The triangle represents one query descriptor. The figure illustrates that the visual and geometric constraints narrow down the objects associated with the query descriptor.

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

$$c_j = \sum_{i=1}^{N} \tilde{f}_{ij} log_2 \frac{K}{L_i}. \tag{11}$$

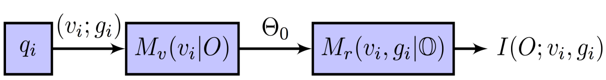


Fig. 3. The pipeline of the embedded 3D geometry score.

5. EXPERIMENTAL RESULTS

5.1. Dataset and Setup

We evaluate our embedded 3D geometry score for the multi-view 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 Cannon IXUS50 digital camera at a resolution of 2592×1944 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`

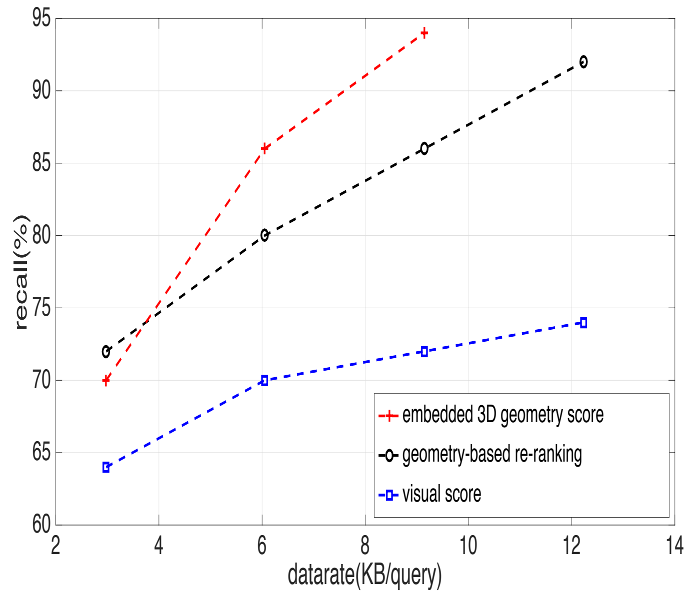


Fig. 4. Comparison of the recall-datarate between different scoring schemes.

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-reranking 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 topranked 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

[1] D. Mars, H. Wu, H. Li, and M. Flierl, "Geometry-based ranking for mobile 3D visual search using hierarchically structured multi-view features," in *Proc. of the IEEE International Conference on Image Processing*, Oct. 2015, pp. 3077 – 3081.

[2] X. Lyu, H. Li, and M. Flierl, "Hierarchically structured multi-view features for mobile visual search," in *Proc. of the IEEE Data Compression Conference*, Mar. 2014.

[3] H. Li and M. Flierl, "Mobile 3D visual search using the Helmert transformation of stereo features," in *Proc. of the IEEE International Conference on Image Processing*, Sept. 2013.

[4] D. Song, W. Liu, D. A. Meyer, D. Tao, and R. Ji, "Rank preserving hashing for rapid image search," in *Proc. of the IEEE Data Compression Conference*, Mar. 2015.

[5] D. M. Chen and B. Girod, "A hybird mobile visual search system with compact global signatures," *IEEE Trans. on Multimedia*, vol. 17, no. 7, pp. 1019 –1030, July 2015.

[6] L.-Y. Duan, Z. Lin, Z. Wang, T. Huang, and W. Gao, "Weighted component hasing of binary aggregated descriptors for fast visual search," *IEEE Trans. on Multimedia*, vol. 17, no. 6, pp. 828 –842, June 2015.

[7] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian, "Semantic-aware co-indexing for image retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2573–2587, Dec. 2015.

[8] N. Pourian, S. Karthikeyan, and B.S. Manjunath, "Search and retrieval of multi-modal data associated with image-parts," in *Proc. of the IEEE International Conference on Image Processing*, Oct. 2015, pp. 3077 – 3081.

[9] B. Girod, V. Chandrasekhar, D. M. Chen, Ngai-Man Cheung, R. Grzeszczuk, T. Reznik, G. Takacs, S.S.Tsai, and R. Vedantham, "Mobile visual search," *Signal Processing Magazine*, vol. 28, no. 4, pp. 61 –76, July 2011.

[10] S. Lepsoy, G. Francini, G. Cordara, and P.P.B. de Gusmao, "SIFT-based improvement of depth imagery," in *Proc. of the IEEE International Conference on Multimedia & Expo*, Barcelona, Spain, July 2011.

[11] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. of the International Conference on Computer Vision*, Oct. 2003.

[12] J. He, J. Feng, X. Liu, T. Cheng, T. H. Lin, H. Chung, and Shih-Fu Chang, "Mobile product search with bag of hash bits and boundary reranking.," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 3005 – 3012.

[13] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing and Management*, vol. 39, no. 1, pp. 45 –65, Jan. 2003.

[14] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2006.

[15] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd edition, 2004.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60(2), pp. 91–110, 2004.

[17] G. Watson, "Computing Helmert transformations," *Journal of Computational and Applied Mathematics*, vol. 197, no. 2, pp. 387 –394, 2006.

[18] S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar, R. Vedantham, R. Grzeszczuk, and B. Girod, "Fast geometric re-ranking for image-based retrieval," in *Proc. of the IEEE International Conference on Image Processing*, 2010.

[19] P. Huber and E. Ronchetti, *Robust Statistics*, Wiley, New York, 2nd edition, 2009.

[20] L-Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, "Overview of the MPEG-CDVS standard," *IEEE Trans. on Image Processing*, vol. 25, no. 1, pp. 179 –194, Jan. 2016.

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