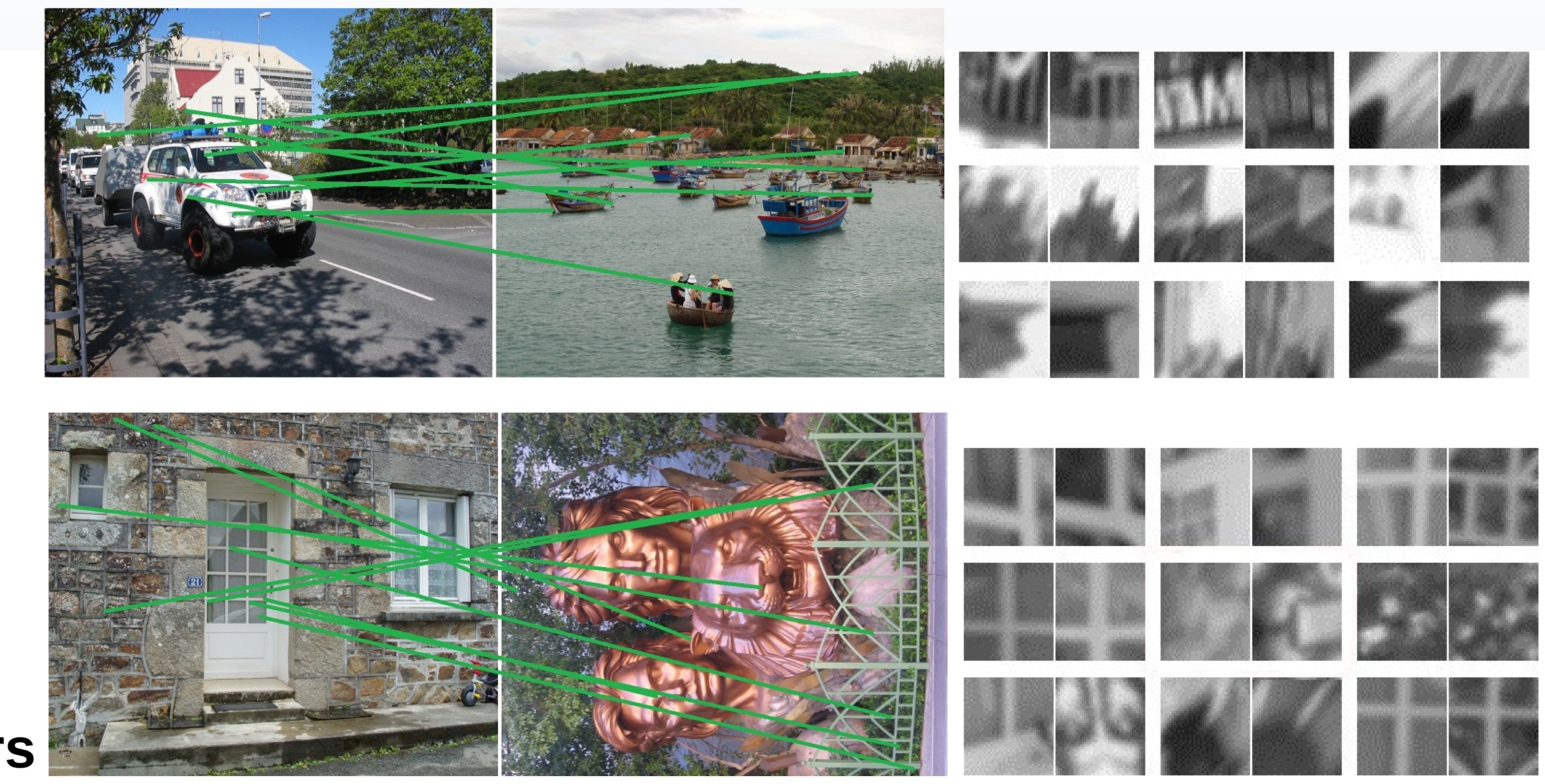
REGION MATCHING AND SIMILARITY ENHANCING FOR IMAGE RETRIEVAL

Guixuan Zhang, Zhi Zeng, Shuwu Zhang, Hu Guan, Qinzhen Guo
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Motivation and Objectives

- **Many image retrieval systems:**
  - **Adopt bag-of-visual-words model**
  - **Usually based on matching of local descriptors (SIFT)**
  - **Not distinctive enough, often lead to false matches**
- **Accurate image retrieval by seeing the big picture:**
  - **Find appropriate regions for providing contextual clues**
  - **Enhance the similarity score for true-matching SIFT pairs**



Matching Regions Estimation

Decompose the image into regions based on spatial pyramid:

Input: Image I (width W and height H)
Output: L layers of regions. In the l-th layer, there are $r_l \times r_l$ regions with size $\frac{W}{s_l} \times \frac{H}{s_l}$

We set L = 4 with

$$(r_1, r_2, r_3, r_4) = (1, 2, 3, 5)$$
$$(s_1, s_2, s_3, s_4) = (1.0, 1.5, 2.0, 3.0)$$

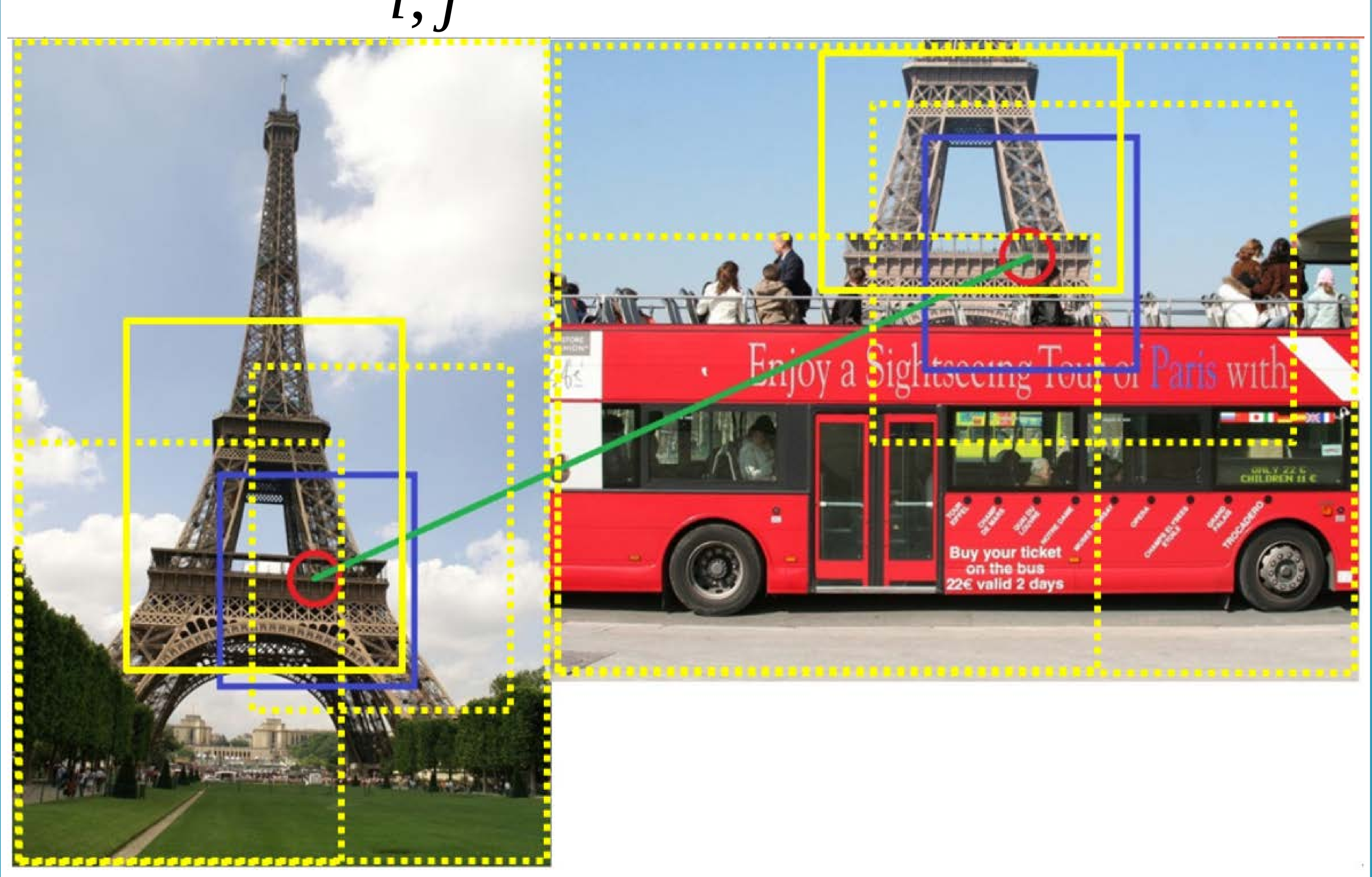
Each image I has 39 region proposals, every keypoint(SIFT) $\boldsymbol{x}$ is located in T_x regions

For a pre-matching pair $(\boldsymbol{x}, \boldsymbol{y})$ (with $score(\boldsymbol{x}, \boldsymbol{y})$ based on Hamming embedding [1]), the corresponding regional feature sets are:

$$\mathcal{P}_x = \{p_t^x, t = 1, \dots, T_x\} \qquad \mathcal{P}_y = \{p_t^y, t = 1, \dots, T_y\}$$

In order to find an appropriate region pair to provide discriminative contextual clues:

$$(m, n) = \arg\max_{i,j} f(p_i^x, p_j^y), p_i^x \in \mathcal{P}_x, p_j^y \in \mathcal{P}_y$$



The regions(depicted by solid yellow rectangle) are used for the next similarity enhancing step.

Binarized Fisher Vector

-- An easy way to measure region similarity

Fisher vector: A global representation of an image by aggregating SIFTs
Binary version: From Euclidian space in to Hamming space.
Then each region is described by a 128-bit signature. The function changes

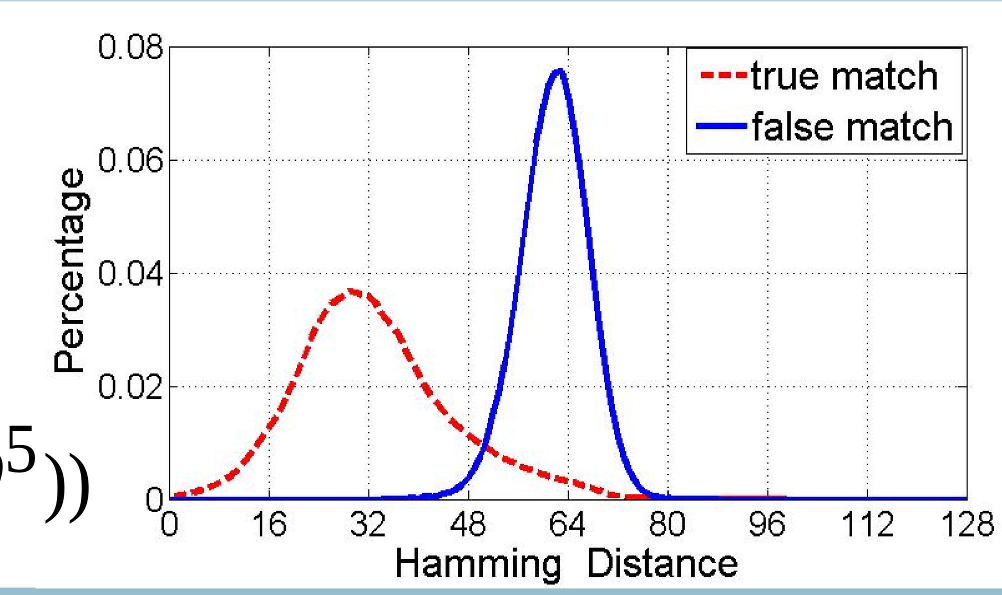
$$(m, n) = \arg\min_{i,j} h(b_f(p_i^x), b_f(p_j^y)), p_i^x \in \mathcal{P}_x, p_j^y \in \mathcal{P}_y$$

Enhance the Matching Score

Denote: $d_f = h(b_f(p_m^x), b_f(p_n^y))$

The similarity enhancing function:

$$score'(\boldsymbol{x}, \boldsymbol{y}) = score(\boldsymbol{x}, \boldsymbol{y}) \times (1 + \exp(-d_f^{\,5} / \theta^5))$$



Experimental Results

Table 2. Image retrieval results for different methods. We integrate all these methods and show the accuracy in the last row.

Methods	Holidays	Oxford5k	Paris	Oxford105k
HE	77.10	69.25	68.37	56.85
HE+Proposed	78.80	71.10	70.21	62.43
HE+MA+Burst	81.00	76.83	73.75	72.06
HE+MA+Burst +Proposed	**82.77**	**78.60**	**75.82**	**73.88**

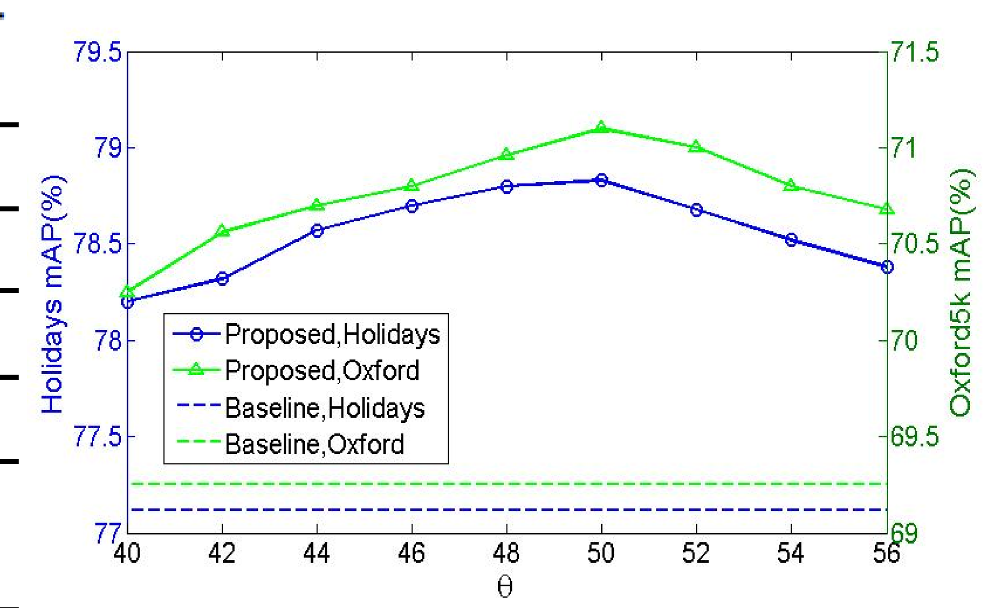


Table 3. Performance comparison with state-of-the-art methods without post-processing. * denotes the case where 128-bit SIFT binary signature is used.

Methods	Ours	Ours*	[27]	[31]	[13]	[22]	[30]	[29]	[18]	[26]*	[28]*
Holiday	**82.77**	**84.27**	82.1	81.9	81.1	82.6	81.92	78.7	-	81.0	-
Oxford5k	**78.60**	81.24	78.0	70.4	72.5	64.7	65.01	77.8	71.17	80.4	**81.3**
Paris	**75.82**	**77.78**	73.6	-	-	-	-	74.1	-	77.0	77.5
Oxford105k	**73.88**	**75.33**	72.8	-	65.2	-	-	72.9	62.34	75.0	-

Time: generating binary FV (0.05s), query the Oxford105k (0.23s)

[1] H. Jegou et al., "Improving Bag-of-Features for Large Scale Image Search", IJCV, 2010.