Fast Image Matching with Visual Attention and SURF Descriptors

Vitaly Pimenov
Faculty of Applied Mathematics and Control Processes
Saint–Petersburg State University, Saint–Petersburg, Russia
vitaly.pimenov@gmail.com

Abstract

The paper describes an image matching method based on visual attention and SURF keypoints. Biologically inspired visual attention system is used to guide local interest point detection. Interest points are represented with SURF descriptors. One–to–one symmetric search is performed on descriptors to select a set of matched interest point pairs. Pairs are then weighted according to attention distribution and weights are summed up yielding a similarity score. Images are considered to be near–duplicates if similarity score exceeds a certain threshold. Experimental results illustrate high accuracy and superior computational efficiency of proposed approach in comparison with other matching techniques.

Keywords: *image matching, visual attention, local interest points, near–duplicate detection.*

1. INTRODUCTION

Image matching problem had in recent years gained remarkable attention in computer vision [1, 2], robotics [3, 4, 5] and information retrieval [6, 7, 8, 9]. A broad spectrum of practical applications such as object recognition [2, 10], content–based image retrieval [9], news retrieval [8], medical imaging [11], copyright infringement detection [7], robotic navigation [4], scene reconstruction [12] and other depends on efficient techniques for finding correspondences between images.

Image matching problem is also called *near–duplicate image identification* [6, 13, 7] or *transformed image identification* [14] in literature and consists of three stages: feature detection, feature representation and feature matching [10]. Matching problem is quite close to image registration problem [15], but, as opposed to registration, transform model estimation and image resampling are not needed for matching. The goal of matching is to detect near–duplicate images. Near duplicates are images that are equal despite the slight degree of variations caused by geometric and photometric transforms, such as lighting and viewpoint changes, differences in frame acquisition time, motion, editing etc. The complexity of matching arises from a big multitude of possible transforms. Several examples of near duplicate images are shown on Fig. 1.

In response to an increasing practical demand a variety of matching methods was developed. The performance of these methods was evaluated in diverse scenarios resulting in conclusion that local interest point matching methods are most promising [16, 10, 17, 18, 6]. Local interest points (keypoints) are salient regions detected over image scales [2]. Image matching in this case becomes a keypoint descriptor matching, where descriptor is a vector of features extracted from image patch around interest point at given scale. Different keypoint detectors and descriptors were proposed for image matching [2, 19, 10, 20, 21]. Later several studies were conducted to assess their quality (see e.g. [16, 20, 10, 18]).

Although local interest point matching remains to be the most accurate matching method its major drawback is high computational complexity: comparing two 640×480 images can take up a million of descriptor comparisons. Real–time implementation (e.g. for mobile robot vision) becomes prohibitively difficult. A variety of hardware focused approaches was already proposed for most critical tasks: VLSI–based solutions [22], FPGA architectures [23, 24], parallel [25] and DSP [26] systems.

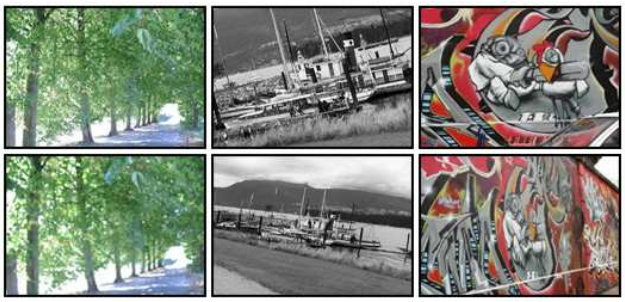

Figure 1: Examples of near duplicate image pairs.

Software alternative lies in keypoint filtering. It is known that geometric verification techniques such as RANSAC [27] require a small number of keypoint for reliable matching [28, 7, 2]. Different techniques were proposed to reduce the number of descriptor comparisons. Most of them rely on two–stage matching, where first stage is rough comparison (e.g. via descriptor discretization) to filter out non–neighbors, and second one is fine matching involving remaining descriptors.

Such approaches successfully reduce computational load, however they are still far away from real–time requirements. Best to our knowledge *matching time* estimations reported in [6] and [13] are 0.028 sec and 0.015 sec respectively. These times do not include *detection time* that is necessary to detect keypoints and compute descriptors. According to results reported in [10] detection time for widely used methods exceeds 0.45 sec per image. At the same time, mobile robot operating at moderate 3 Hz frequency has 0.33 sec for whole control system loop execution. Considering content–based retrieval where detection time is not as important as matching time, we see that 0.015 sec per pair matching results in poor 66 comparisons per second.

To overcome this difficulty on software level much simpler detection and matching methods are employed, e.g. Shi–Tomasi operator [21] or color histograms [29] that allow 10^{-5} sec matching time. However, in general, these methods cannot achieve accuracy level of their keypoint–based competitors.

Current paper proposes novel image matching method guided by a biologically motivated visual attention system. The key advantage of such system is suppression of detected keypoints number. It is achieved during detection phase performed with SURF detector [10] by filtering out points that gain little attention. As a result matching speed increases by more that an order of magnitude, detection speed increases by 2 – 4 times. Experimental results show high accuracy of matching, comparable to accuracy of descriptor matching implemented without filtration.

The remaining sections are organized as follows. Section 2 surveys related work in visual attention guided image matching and descriptor filtering. Section 3 describes proposed visual attention system. Interest point matching method is discussed in Sect. 4. Experiments and evaluation results are presented in Section 5. Finally, Sect. 6 concludes the paper and outlines directions of further research.

2. RELATED WORK

Visual attention has emerged recently as a powerful tool to make computational vision more effective, since it allows focusing analysis on restrained image areas [30, 4, 31]. This section presents a survey of modern attention–based image matching methods and descriptor filtering algorithms.

At the heart of modern computational attention theories is the concept of *saliency map*. As originally proposed in [32], saliency map refers to a *"topographically arranged map that represents visual saliency of a corresponding visual scene"* [33].

Typically, attention–based matching methods employ saliency map to detect interest points — point with high saliency. To describe interest point various approaches are used in literature.

Saliency based image matching method is presented in [34]. The foundation of this method is *Scale – Saliency* algorithm developed by Kadir and Brady [35] aimed to detect image regions salient by means of Shannon entropy measure [36]. There are two drawbacks of such approach. First is the use of normalized intensity histograms to describe interest points, since its invariance to geometric and photometric transforms is debatable. Second, *Scale – Saliency* algorithm suffers from the curse of dimensionality when it is applied to multidimensional data, e.g. RGB images. Although this issue was resolved recently [37], there is no experimental evidence proving the efficiency of multidimensional *Scale – Saliency* in image matching tasks.

The question of salient regions detection repeatability is considered in [38]. Comparison with popular Difference–of–Gaussians [2] and Harris–Laplace [19] detectors reveals the superior repeatability of detection performed by biologically inspired visual attention system. Therefore, it is concluded that filtering out descriptors corresponding to regions that cannot be detected with high repeatability can significantly reduce computational load. This result supports the motivation of current research.

Works by Stentiford and colleagues [39, 40, 17, 41] focus on constructing attention based similarity measures with applications to content–based image retrieval, motion detection and tracking. Reported results show superior performance of their approach in terms of recall–precision graph when compared to color histograms and Gabor signatures. However, these techniques had not yet been tested against local interest point detectors and no exact data concerning matching speed is available.

More sophisticated bio–inspired attention system is proposed in [9] for content–based image retrieval. In order to improve quality of detection this system combines Stentiford model of attention [41] with Itti–Koch model [30]. RGB and HMMD color intensity histograms were used as descriptors. Reported experimental result suggest high accuracy of this approach, however it was not compared to interest point based detectors.

Broad comparative study described in [18] aimed to evaluate modern near–duplicate detection methods on a large scale collection (more than one million web images) had revealed high performance of bio–inspired retina–like image matching method. The study reported poor performance of SURF [10], interest point detection method, however descriptor matching algorithm used in experiments in highly debatable, as it is mentioned by authors themselves. Furthermore, image collection was built by applying image transforms to an initial image collection, thus validity of evaluation result for natural near–duplicates detection is arguable in this case.

At the same time, question of interest point matching scalability has already been tackled from the perspective of pruning the number of descriptors used to match [28]. SIFT [2] and PCA–SIFT [7] descriptors were chosen in this research. Pruning was made on the basis of point contrast. Experimental results show reduction of execution time to $1/50$ of the original approach without any significant loss of accuracy. Nevertheless, this study focuses on image retrieval task, therefore only retrieval query speed is evaluated, while matching speed is unknown.

Summing up, attention–based approaches to image matching became quite popular along with the development of human vision models. In the light of results presented in [28], employing visual attention to filter out insignificant interest points seems to be a promising approach to improve efficiency of image matching.

3. VISUAL ATTENTION SYSTEM

The purpose of visual attention for image matching tasks if defined as following: *determine the importance of image regions in terms of attention distribution*.

Computation is performed at two stages: first, saliency map is computed from an input image; second, an absolute measure of saliency is built from the saliency map.

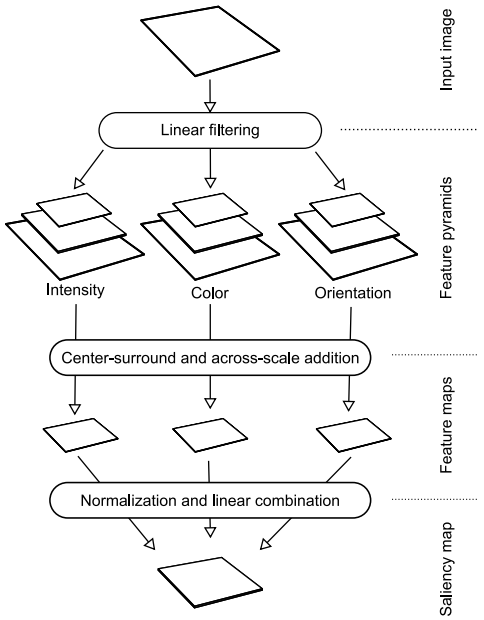

Figure 2: Overview–diagram of saliency map calculation.

In current research saliency map is implemented on pixel basis as it was proposed in [30, 4]. For each pixel saliency is computed from three feature channels mimicking human vision [42]: intensity, color and orientation. After than, per–channel maps are normalized and fused into a saliency map. Figure 2 illustrates processing flow, its details are outlined in Sect. 3.1.

Absolute measure of saliency is necessary because saliency maps contain relative values. Two distinct methods are proposed and described in Sect. 3.3 to solve this problem.

3.1 Feature Computation

Feature computation is roughly based on procedures proposed for visual attention system *VOCUS* [4]. In following subsections a brief description of procedures for each feature channel is offered.

Input image I is used to build monochrome and LAB Gaussian pyramids of five scales: $(s_0, \ldots, s_4)$ and $(\bar{s}_0, \ldots, \bar{s}_4)$ respectively; and Difference–of–Gaussians pyramid [2] of four scales $(DoG_1, \ldots, DoG_4)$. First two scales (s_0, s_1) are not considered in calculations to ensure robustness to noise [30].

3.1.1 Intensity

Intensity feature maps are computed by applying *center–surround mechanism* to images of Gaussian pyramid. On–center and off–center intensity [4, 42] of a pixel is computed by comparing its value with mean value of surrounding pixels. Such calculation is consistent with recent findings about surround suppression mechanisms in visual cortex [31]. Mean value is calculated by summing pixel values in 7×7 quadratic area Ω around pixel of interest:

$$I^{[s_i]}_{on-center}(x,y) = s_i(x,y) - \frac{\sum\limits_{(\bar{x},\bar{y})\in\Omega} s_i(\bar{x},\bar{y}) - s_i(x,y)}{|\Omega| - 1} \quad . \quad (1)$$

For efficiency reasons integral images [43] are used to compute a sum in equation (1).

On–center and off–center intensity maps are built for scales (s_2, s_3, s_4). Resulting maps are summed up by *across–scale addition* ($\bigoplus$) [4] — smaller maps are scaled to size of biggest one and then maps are summed up by pixels — leading final intensity maps $I_{on-center}$ and $I_{off-center}$:

$$I_{on-center} = \bigoplus_{i\in\{2,3,4\}} I^{[s_i]}_{on-center} \quad .$$

3.1.2 Color

Color feature maps are computed in terms of intensities for colors *red*, *green*, *blue* and *yellow*. Such choice of basis colors corresponds to human vision [44]. As original RGB image is converted into CIE LAB universal color space [44], Euclidean distance ρ_{LAB} between colors corresponds to human perception and can be used to calculate color intensity:

$$C^{color}_i(x,y) = \rho_{\mathrm{LAB}}(s_i(x,y), color) \quad .$$

Feature maps are computed for each basis color by applying center–surround mechanism and across–scale addition to corresponding pyramids:

$$C^{color}_{on-center} = \bigoplus_{i\in\{2,3,4\}} I^{[C^{color}_i]}_{on-center} \quad .$$

3.1.3 Orientation

Orientation feature maps highlight edges having basis orientations θ: 0°, 45°, 90° and 135°. Feature map for each orientation is computed by applying corresponding Gabor filter [44] of specified orientation to images of DoG pyramid:

$$O^\theta_i(x,y) = (G_\theta \star \mathrm{DoG}_i)(x,y) \quad .$$

Gabor filters simulate response of orientation–selective neurons in visual cortex[42]. Filtered pyramids are summed up via across–scale addition yielding four orientation maps:

$$O^\theta = \bigoplus_{i\in\{2,3,4\}} O^\theta_i \quad .$$

3.2 Saliency Map

Feature maps are then normalized and fused into a combined map. Normalization operator $\mathcal{N}(\cdot)$ is adopted from [4]:

$$\mathcal{N}(I)(x,y) = \frac{1}{\sqrt{m}} I(x,y) \quad ,$$

where m is a number of local maxima above threshold that was chosen to be 0.65% of map's global maximum. This is necessary to smooth maps having a lot of local maxima. Normalized maps are summed up with equal weights yielding a combined map:

$$M = \frac{1}{3}\mathcal{N}\left(I_{on-center} + I_{off-center}\right) + \frac{1}{3}\mathcal{N}\left(\sum_{color\in\{R,G,B,Y\}} \left(C^{color}_{on-center} + C^{color}_{off-center}\right)\right) + \frac{1}{3}\mathcal{N}\left(\sum_{\theta\in\{0^\circ,45^\circ,90^\circ,135^\circ\}} O^\theta\right) \quad .$$

Equal weights are used for simplicity reasons although it is known that variable weights are preferable [45, 46].

Example of an image and its saliency map is shown on Fig. 3.

Figure 3: Saliency map example. On the left: source image. On the right: final saliency map

3.3 Absolute Measure of Saliency

Saliency map contains relative values, but for purposes of descriptor filtering it is necessary to have absolute measure of saliency, reusable between images.

Two methods are proposed to solve this problem. First method is based on introducing a measure of attention equivalency. Second one is normalization of a saliency maps to a single scale.

3.3.1 Attention equivalency measure

To build measure of attendance it is convenient to divide an image into a set of attended points and a set of unattended points. Straightforward thresholding is ineffective because of saliency variations, even within a single object [47]. To overcome this difficulty a fuzzy set theoretic methods can be developed. Method employed for current research is a simplified variant of *fuzzy growing* [47].

Let $\Omega = \{g_k, k = \overline{0, L-1}, L = 256\}$ be a set of saliency values. Two fuzzy sets are defined: set of attended points B_A and set of unattended points B_U with membership functions (2) and (3) respectively.

$$\mu_A(g_k) = \begin{cases} 1 & g_k \geq a \\ \dfrac{g_k - u}{a - u} & u < g_k < a \\ 0 & g_k \leq u \end{cases} \quad , \quad (2)$$

$$\mu_U(g_k) = \begin{cases} 0 & g_k \geq a \\ \frac{g_k - a}{u - a} & u < g_k < a \\ 1 & g_k \leq u \end{cases} . \quad (3)$$

Parameters a and u in (2) and (3) are constants that determine optimal fuzzy 2–partition. Details on these calculations are to be found in [47]. Function $\mu_A(g_k)$ is an absolute measure of saliency.

3.3.2 Saliency Map Normalization

Less complicated way to introduce absolute saliency values it to normalize all saliency maps to a single scale. In this case saliency values can be compared. Further motivation for this kind of processing is provided by an attention conservation hypothesis, proposed in [48]. It claims that total amount of saliency is invariant: perception causes only redistribution of that amount among input stimuli (i.e. image pixels).

Neurobiological evidence is used to determine the total amount of saliency. It is known that fovea comprises about 1% of retinal size but is responsible for over 50% of information [42]. As saliency map is a $W \times H$ table with values in $[0, \ldots, 255]$, total attention amount is calculated as $I = \frac{(255 \times 1\%(W \cdot H))}{50\%}$. Value of I is then used to scale the map.

4. INTEREST POINT MATCHING

Five step procedure is performed to match a pair of images:

1. Saliency maps are computed for both images according to Sect. 3.. Either of methods described in Sect. 3.3 is used to build absolute measure of saliency.
2. Local interest points are detected, non–salient points are pruned; SURF descriptors are calculated for remaining.
3. One–to–one symmetric search is performed on descriptors to select a set of matched interest point pairs.
4. Outlying false matches are identified and filtered out.
5. Remaining pairs are weighted by their saliency. Weights are summed up yielding a similarity score. Images are considered as near–duplicates if similarity score exceeds a threshold.

4.1 Interest Point Detection and Description

SURF detector and descriptor [10] were used in current research. The motivation for such choice is two–fold. First, performance of SURF is proved to be equal or superior to performance of other methods, such as SIFT [2], PCA–SIFT [1] and GLOH [16], in independent evaluative studies [49, 10]. Second, its computational efficiency is significantly better in comparison with aforementioned methods. SURF was successfully applied in vision–based mobile robot navigation [50, 49] and handle recognition [51].

The purpose of detector is to find scale–invariant points. SURF detector is based on calculating approximate Hessian response for image points. Although calculating Gaussian response is optimal for scale–space analysis [52], it is shown in [10] that due to aliasing errors its actual performance is not as perfect as in theory. Equal performance can be achieved by approximating Gaussian with Hessian calculated with box filters (Fast–Hessian [10]). This processing can be very efficiently implemented on the basis of integral images [43].

After detection non–salient keypoints are pruned. Non–salient points are points with saliency $\bar{g}$ such that either $\mu_A(\bar{g}) = 0$ or $\bar{g} < SaliencyThreshold$ in case of normalization.

Example of pruning results is illustrated on Fig. 4. Number of points after filtration reduces to up to 1/10 of whole.

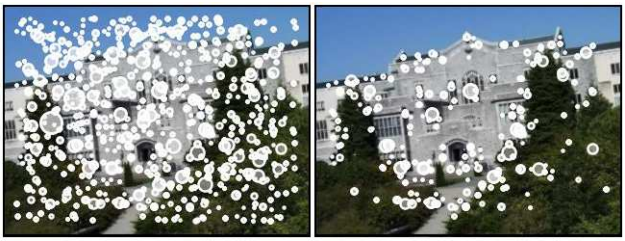

Figure 4: Attention–guided interest point filtering. On the left: image is shown with whole set of detected interest points. On the right: same image is shown with points remained after filtering.

Afterwards rotation invariant SURF descriptors (64–dimensional vectors) are computed for remaining interest points. Descriptor calculations are based on computing Haar wavelet responses in the vicinity of interest point that is implemented with integral images too. Exact details on above operations are given in [10].

4.2 Descriptor Matching

One–to–one symmetric search suggested in [6] is used in current research: given two sets of descriptors $\{P\}$ and $\{Q\}$ extracted from a pair of images (I_1, I_2), it returns pairs of closest descriptors. Algorithm 1 describes proposed search strategy.

Algorithm 1 One–to–one symmetric search

```
Given two sets of descriptors {P} and {Q}
for ∀P̄ ∈ {P} do
    Q̄ ⇐ NearestNeighbor(P̄, {Q})
    P* ⇐ NearestNeighbor(Q̄, {P})
    if P̄ == P* then
        if ρ(P̄, Q̄) < DistanceThreshold then
            pair (P̄, Q̄) is added to the result
        end if
    end if
end for
```

Function $NearestNeighbor(A, \{B\})$ in Algorithm 1 returns descriptor $\bar{B} \in \{B\}$ nearest to given A. Distance between descriptors is measured with Euclidean metric $\rho(P, Q)$. Value of $DistanceThreshold$ is experimentally chosen to be 0.2.

Straightforward implementation of $NearestNeighbor(A, \{B\})$ via exhaustive search is computationally prohibitive. To overcome this difficulty several solutions were proposed in literature: *K–d* trees [20], locality sensitive hashing [28], LIP–IS (local interest point index structure) [6], LIP–IS with inverted index [13]. For this work an extension of the latter approach is proposed.

Threshold–based matching is used instead of *nearest–neighbor–distance–ratio*, originally employed for SIFT descriptor matching [2], because explicit threshold value is necessary for chosen indexing approach. Furthermore it was found in experiments that keypoint filtering makes difference between results obtained with both algorithms almost negligible.

Basic idea behind LIP–IS is following. As soon as $\rho(P, Q)$ is most resource consuming part of matching procedure, performance can be improved if distance value will not be calculated for descriptor pairs such that $\rho(P, Q) \gg DistanceThreshold$. For this purpose

rough estimation of distance is computed. It is done with help of descriptor quantization. Original 64–bit vector of double values $P = (p_1, \ldots, p_{64})$ is transformed to $\hat{P} = (\hat{p}_1, \ldots, \hat{p}_{64})$, where $\hat{p}_i$ take on values from a discrete set $H = (h_1, \ldots, h_N)$. It is convenient to use $N = 8$, in this case distance can be estimated with fast bit operations.

Inverted index is used to further reduce sets of descriptors to compare — by indexing groups of descriptors P that have equal values in first k dimensions of corresponding $\hat{P}$ vectors. Then each descriptor is only compared to descriptors belonging to the same group in inverted index. As reported in [13] inverted index reduces matching time to 20–60% of original LIP–IS time. At the same time, it is built only once for each image and has very moderate memory footprint.

In this work an improvement to an inverted index is proposed. It was noticed during experiments that indexing first k dimensions often has a little effect because for most descriptors values corresponding to these dimension become equal after quantization. It makes inverted index useless in such cases. To get over this difficulty is was proposed to index k dimensions having maximum variance across set $\{B\}$. With this modification preliminary filtration time is steadily reduced to $20 \pm 5\%$ of original LIP–IS time. New structure was called *Maximum Variance Inverted Index (MVII)*.

4.3 False Match Identification

Due to quantization and overall descriptor design some degree of false matches can appear in nearest–neighbor search results. Outlying false matches can be identified relatively easy.

For each descriptor pair angle and distance between corresponding points are calculated. Then mean and standard deviation are computed for angles and distances across all matched pairs.

Finally, pairs such that difference between angle or distance from corresponding mean greatly exceeds respective standard deviation are considered to be false matches and are thereupon pruned. Figure 5 illustrates described procedure.

Geometrical verification techniques, such as RANSAC [27], used in several studies for same purposes (e.g. in [19]) are not employed in this research because simple procedure described above was found to be sufficient for images of test collection.

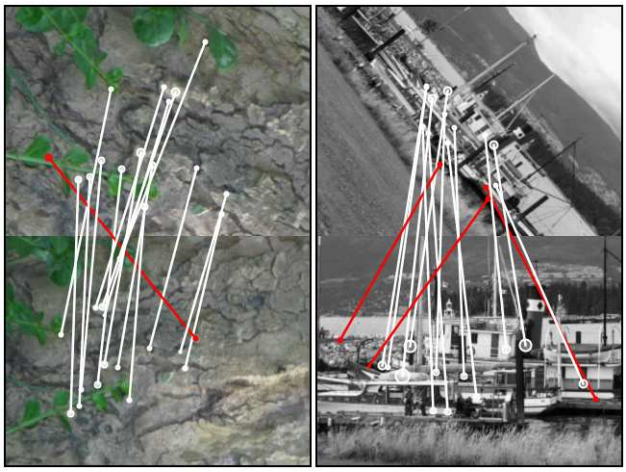

Figure 5: False match identification. White strokes depict correct matches, red strokes depict false matches.

4.4 Similarity Score

As soon as matched descriptor pairs are identified and false matches are pruned we need to make a decision of matching. For this purpose a similarity score is computed

$$S = \begin{cases} \frac{1}{255} \sum\limits_{(P,Q)\in N} (g_P + g_Q), \text{ in case of normalization,} \\ \sum\limits_{(P,Q)\in N} \left(\mu_A^{I_1}(g_P) + \mu_A^{I_2}(g_Q) \right), \text{ otherwise.} \end{cases} \quad (4)$$

In (4) N is a set of matched interest point pairs; g_P and g_Q are saliency values for points P and Q respectively; $\mu_A^{I_k}$ is membership function of attended points set computed for image I_k.

If similarity score (4) exceeds threshold $\bar{S}$ than images are said to be near duplicates. Threshold value $\bar{S}$ is a variable that determines tradeoff between recall and precision.

5. EXPERIMENTS AND RESULTS

Experiments were conducted to assess the performance of proposed image matching method against different approaches. Although various setups can be used for assessment (for instance, content–based image retrieval tasks and image collection clustering), in this research we have chosen to focus solely on image pair matching. Such decision is an attempt to reduce task specific bias, e.g. influence of clustering method on clustering results.

5.1 Data Set

Evaluation was carried out on real world images with different geometric and photometric transforms. We have adopted a data set from [16], that is widely used in comparative studies (e.g. in [10, 19, 20, 53]). Collection consists of 8 groups of near duplicates, 49 images in total, and is publicly available on the Internet[1]. Images are produced with lighting and viewpoint changes, blurring, zoom, rotation and compression. Sample images are shown on Fig. 1.

5.2 System Implementation

Software used for experiments was implemented in Java to simplify performance evaluation and analysis via code profiling. All tests were run on Intel Core 2 Duo 1.83 GHz machine with 2 Gb memory under Microsoft Windows XP operating system.

5.3 Experiment Setup

Four image matching methods based on SURF descriptors were compared in quality and efficiency tests:

1. Naive matching without any filtering.
2. Matching with contrast filtering proposed in [28]: top M keypoints with highest contrast value are selected for matching.
3. Attention–guided matching with threshold filtering: top M keypoints with highest saliency are selected for matching.
4. Attention–guided matching with similarity score (4).

Last method was tested with fuzzy measure and normalization. In case of normalization a variety of $SaliencyThreshold$ values were used to assess its influence on matching quality and speed.

Following procedure was performed to evaluate the quality of matching. Each image of test collection was compared with each of

[1] http://lear.inrialpes.fr/people/mikolajczyk/

Table 1: Performance evaluation results. N# denotes normalization with a given value of $SaliencyThreshold$.

Method	Average precision	Average Recall	Saliency map time (ms)	Detection time (ms)	Matching time (ms)
Naive matching	1.0	0.94	0	340	294
Top–contrast	0.98	0.89	0	340	24
Top–saliency	0.99	0.90	106	340	24
Similarity (Fuzzy)	0.91	0.81	320	300	210
Similarity (N30)	0.93	0.85	118	300	95
Similarity (N50)	0.91	0.85	118	160	40
Similarity (N70)	0.91	0.80	118	115	12
Similarity (N90)	0.92	0.67	118	95	3
Similarity (N110)	0.91	0.59	118	85	0.8

remaining images. Results of comparisons: true and false matches, were recorded. As soon as we know original groups of near duplicates, number of correct matches can be calculated for each group.

To assess accuracy of matching two metrics were computed for each group and in average: recall and precision:

$$recall = \frac{\text{\# correct true matches}}{\text{group size}} \ ,$$

$$precision = \frac{\text{\# correct true matches}}{\text{\# true matches}} \ .$$

Three metrics were also calculated to assess speed efficiency of each method: saliency map computation time, detection time and matching time. Saliency map computation time is spent to build saliency map and equivalency classes. Detection time is spent to detect interest points and compute their descriptors for a single image. Matching time is spent to match a pair of images: it includes time to build all relevant index structures. All metrics were calculated in average across all images and all pairs respectively.

For index efficiency test following methods were chosen: no indexing; LIP–IS; LIP–IS and inverted index; LIP–IS with MVII.

5.4 Results

Table 1 summarized method performances. Resulting recall–precision graphs are shown on Fig. 6. Average precision is almost unaffected by saliency threshold, however average recall decays as threshold value increases. Exact dependencies of accuracy and average matching time on saliency threshold are also demonstrated on Fig. 6. Analyzing these plots together we can see that variable saliency threshold allows to find a suitable tradeoff between accuracy and speed.

In comparison between fuzzy measure and normalization the latter approach is a clear winner. Experiments have shown that building optimal fuzzy partition of saliency map is computationally ineffective. At the same time performance of this approach does not exceed performance of normalization achieved with considerably lower costs.

Performances of contrast–based filtering and implemented alike attention thresholding are near equal. The reason for this is following: to reach acceptable accuracy levels we have to use top 300 interest points for both approaches. Experiments have shown that difference between set of top 300 points with highest contrast and set of top 300 points with highest saliency is negligible for images of test collection. Therefore we see nearly identical recall and precision. However attention thresholding is accompanied with additional costs because saliency map must be built. Thus top–contrast method is preferable.

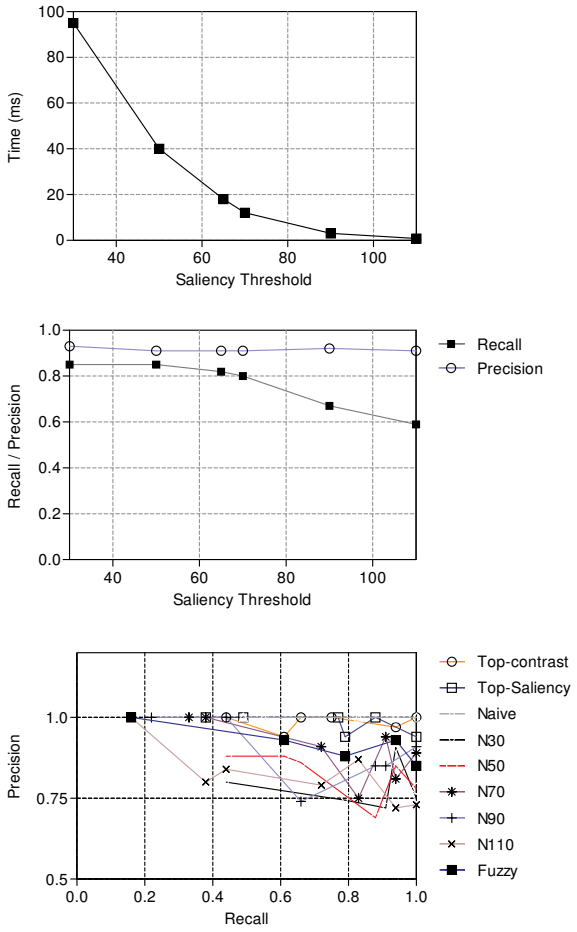

Figure 6: Evaluation results. Upper plot: dependency of average matching time on saliency threshold. Middle plot: dependency of accuracy metrics on saliency threshold. Lower plot: recall–precision graphs.

Table 2 contains obtained results that prove superior performance of LIP–IS with MVII in comparison with other techniques. We can see that original inverted index has a little effect on LIP–IS performance. We also rely on results reported in [6] indicating that LIP–IS outperforms locality sensitive hashing.

Table 2: Index structure efficiency.

Method	Matching time (ms)
No indexing	340
LIP–IS	28
LIP–IS + Inverted Index	24
LIP–IS + MVII	7

6. CONCLUSION

The paper proposed an image matching method based on visual attention and SURF keypoints. Biologically inspired visual attention system was used to guide local interest point detection and significantly reduce the number of interest points used in matching. Experimental results have shown attractive performance of new method in comparison with several different methods.

For time–critical tasks attention–guided matching based on normalization and *similarity score* is an attractive choice since it allows to increase matching speed by more that an order of magnitude (24.5 times for N70) with performance loss near 10% for average precision and average recall. At the same time, computations required to construct saliency map are a weak point of this approach. Although detection time decreases by up to 4 times, additional costs related to saliency map computation almost nullify this advantage.

In tasks where detection time is not constrained while matching time is critical, for instance, in content–based image retrieval tasks, *top–contrast* thresholding is most accurate method. But in cases where performance can be sacrificed for the sake of matching speed, *similarity score* methods can be applied as they reduce matching time by up to 30 times (for N110) from *top–contrast*.

Although saliency map computations are a weak point from speed efficiency standpoint, the use of visual attention has been proven as an effective method to achieve near real–time matching efficiency without significant loss of quality. Further research will be directed towards development of faster saliency map computation methods.

7. REFERENCES

[1] Y. Ke and R. Suthanakar, "Pca-sift: A more distinctive representation for local image descriptors," *Comput. Vis. and Pattern Recogn.*, vol. 2, pp. 506–513, 2004.

[2] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intern. J. of Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, 2007.

[4] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2006, `http://tinyurl.com/frintrop`.

[5] T. D. Barfoot, "Online visual motion estimation using fastslam with sift features," in *Proc. of IEEE/RSJ Intern. Conf. on Intelligent Robots and Systems, 2005*, 2005, pp. 579–585.

[6] W. L. Zhao, C. W. Ngo, H. K. Tan, and X. Wu, "Near-duplicate keyframe identification with interest point matching and pattern learning," *IEEE Trans. on Multimedia*, vol. 5, no. 9, pp. 1037–1048, 2007.

[7] Y. Ke, R. Suthanakar, and L. Huston, "Efficient near-duplicate detection and sub-image retrieval," in *Proc. of ACM Multimedia Conf.*, 2004, pp. 869–876.

[8] X. Wu, C.-W. Ngo, and Q. Li, "Threading and autodocumenting news videos," *Signal Processing Magazine*, vol. 23, no. 2, pp. 59–68, 2006.

[9] O. Marques, L. M. Mayron, G. B. Borba, and H. R. Gamba, "An attention-driven model for grouping similar images with image retrieval applications," *EURASIP J. of Applied Signal Processing*, vol. 2007, no. 1, pp. 116–116, 2007.

[10] H. Bay, A. Ess, T. Tuytelaars, and Van Gool L., "Surf: Speeded up robust features," *Comput. Vis. Image Underst.*, vol. 3, no. 110, pp. 346–359, 2008.

[11] X. Zheng, M. Zhou, and X. Wang, "Interest point based medical image retrieval," pp. 118–124, 2008.

[12] P. Labatut, J.-P. Pons, and R. Keriven, "Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts," in *IEEE 11th Intern. Conf. on Computer Vision. ICCV 2007*, 2007, pp. 1–8.

[13] V. Pimenov, "Near-duplicate image detection with local interest point extraction," in *Proc. of the Sixth Russian Information Retrieval Evaluation Seminar, ROMIP'2008*, 2008, pp. 145–159, Available in russian at `http://tinyurl.com/pimenov08`.

[14] M. Awrangjeb and G. Lu, "An improved curvature scale–space corner detector and a robust corner matching technique for transformed image identification," *IEEE Trans. Image Process.*, vol. 17, no. 12, pp. 2425–2441, 2008.

[15] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003.

[16] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.

[17] L. Chen and F. W. M. Stentiford, "An attention based similarity measure for colour images," in *Proc. of 16th Intern. Conf. on Artificial Neural Networks ICANN 2006*, 2006, pp. 481–487.

[18] B. Thomee, M. J. Huiskes, E. Bakker, and M. S. Lew, "Large scale image copy detection evaluation," in *MIR '08: Proc. of the 1st ACM intern. Conf. on Multimedia Information Retrieval*, 2008, pp. 59–66.

[19] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. of the 7th Eur. Conf. on Computer Vision*, 2002, pp. 128–142.

[20] K. Mikolajczyk and J. Matas, "Improving descriptors for fast tree matching by optimal linear projection," in *Proc.of IEEE Intern. Conf. on Computer Vision*, 2007, pp. 1–8.

[21] J. Shi and C. Tomasi, "Good features to track," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[22] D. Kim, K. Kim, J.-Y. Kim, S. Lee, and H.-J. Yoo, "An 81.6 gops object recognition processor based on noc and visual image processing memory," in *Proc. of IEEE Custom Integrated Circuits Conf.*, 2007, pp. 443–446.

[23] H.D. Chati, F. Muhlbauer, T. Braun, C. Bobda, and K. Berns, "Hardware/software co-design of a key point detector on fpga," in *Proc. of Intern. Symposium on Field–Programmable Custom Computing Machines*, 2007, pp. 355–356.

[24] S. Se, H. Ng, P. Jasiobedzki, and T. Moyung, "Vision based modeling and localization for planetary exploration rovers," in *Proc. of 55th Intern. Astronautical Cong.*, 2004, pp. 1–11.

[25] V. Bonato, E. Marques, and G. A. Constantinides, *Reconfigurable Computing: Architectures, Tools and Applications*, chapter A Parallel Hardware Architecture for Image Feature Detection, Springer, Heidelberg, 2008.

[26] J. Ferruz and A. Ollero, "Real-time feature matching in image sequences for non-structured environments. applications to vehicle guidance," *J. of Intelligent and Robotics Systems*, vol. 28, no. 1-2, pp. 85–123, 2000.

[27] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[28] J. J. Foo and R. Sinha, "Pruning sift for scalable near-duplicate image matching," in *ADC '07: Proc. of the 18th Conf. on Australasian Database*, 2007, pp. 63–71.

[29] W. Jia, H. Zhang, X. He, and Q. Wu, "A comparison on histogram based image matching methods," in *2006 IEEE Intern. Conf. on Advanced Video and Signal Based Surveillance*, 2006, p. 97.

[30] C. Koch, L. Itti, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, , no. 20, pp. 1254–1259, 1998.

[31] N.D.B. Bruce and J.K. Tsotsos, "Spatiotemporal saliency: Towards a hierarchical representation of visual saliency," in *5th Int. Workshop on Attention in Cognitive Systems*, 2008, pp. 98–111.

[32] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[33] E. Niebur, "Saliency map," *Scholarpedia*, vol. 2, no. 8, pp. 2675, 2007.

[34] J. S. Hare and P. H. Lewis, "Scale saliency: Applications in visual matching, tracking and view-based object recognition," in *Proc. of Distributed Multimedia Systems 2003 / Visual Information Systems 2003*, 2003, pp. 436–440.

[35] T. Kadir and M. Brady, "Saliency, scale and image description," *Intern. J. of Comp. Vision*, vol. 2, no. 45, pp. 83–105, 2001.

[36] S. Gilles, *Robust Description and Matching of Images*, Ph.D. thesis, University of Oxford, 1998.

[37] P. Suau and F. Escolano, "Multi-dimensional scale saliency feature extraction based on entropic graphs," in *Proc. of the 4th Intern. Symposium on Advances in Visual Computing*, 2008, vol. II, pp. 170–180.

[38] S. Frintrop, "The high repeatability of salient regions," in *Proc. of ECCV Workshop "Vision in Action: Efficient Strategies for Cognitive Agents in Complex Environment"*, 2008.

[39] F. W. M. Stentiford, "An attention based similarity measure with application to content-based information retrieval," in *Proc. of the Storage and Retrieval for Media Databases Conf., SPIE Electronic Imaging*, 2003, pp. 221–232.

[40] L. Chen and F. W.M. Stentiford, "Comparison of near-duplicate image matching," in *Proc. of 3rd Eur. Conf. on Visual Media Production*, 2006, pp. 38–42.

[41] F. W. M. Stentiford, "Attention-based similarity," *Pattern Recognition*, vol. 40, no. 3, pp. 771–783, 2007.

[42] S.E. Palmer, *Vision Science, Photons to Phenomenology*, MIT Press, Cambridge, 1999.

[43] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 511–518.

[44] D.A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, Berkeley, 2003.

[45] H. Hügli and Bur. A., "Adaptive visual attention model," in *Proc. of Image and Vision Computing New Zealand*, 2007, pp. 233–237.

[46] B. Rasozadeh, A. Tavakoli Targhi, and Eklundh J.-O., "An attentional system combining top-down and bottom-up influences," in *Proc. of Intern. Workshop on Attention in Cognitive Systems*, 2007, pp. 123–140.

[47] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *MULTIMEDIA '03: Proc. of the 11th ACM Intern. Conf. on Multimedia*, 2003, pp. 374–381, `http://tinyurl.com/yfma03`.

[48] A. V. Galsko, "The model of attention dynamics in perception process," *Zh. Vyssh. Nerv. Deiat.*, vol. 58, no. 6, pp. 738–754, 2008.

[49] E. Frontoni, A. Ascani, A. Mancini, and P. Zingaretti, "Performance metric for vision based robot localization," in *Robotics Science and Systems 2008, Workshop on Good Experimental Methodologies*, 2008.

[50] F. Dayoub and T. Duckett, "An adaptive appearance-based map for long-term topological localization of mobile robots," in *Proc. of the IEEE/RSJ Intern. Conf. on Intelligent Robots and Systems*, 2008, pp. 3364–3369.

[51] E. Jauregi, E. Lazkano, J. M. Martínez-Otzeta, and B. Sierra, *European Robotics Symposium 2008*, chapter Visual Approaches for Handle Recognition, Springer, Heidelberg, 2008.

[52] J. Koenderink, "The structure of images," *Biological Cybernetics*, vol. 50, pp. 363–370, 1984.

[53] G. J. Burghouts and J.-M. Geusebroek, "Performance evaluation of local colour invariants," *Comput. Vis. Image Underst.*, vol. 113, no. 1, pp. 48–62, 2009.

ABOUT THE AUTHOR

Vitaly Pimenov is a Ph.D. student at Saint–Petersburg State University, Faculty of Applied Mathematics and Control Processes.