Best-Buddies Similarity—Robust Template Matching Using Mutual Nearest Neighbors

Best-Buddies Similarity for Robust Template Matching

Tali Dekel[*1] Shaul Oron[*2] Michael Rubinstein[†3] Shai Avidan[2] William T. Freeman[1]

[1] MIT CSAIL [2] Tel Aviv University [3] Google Research

{talidek,billf}@mit.edu {shauloro,avidan}@eng.tau.ac.il mrub@google.com

Abstract

We propose a novel method for template matching in unconstrained environments. Its essence is the Best-Buddies Similarity (BBS), a useful, robust, and parameter-free similarity measure between two sets of points. BBS is based on counting the number of Best-Buddies Pairs (BBPs)—pairs of points in source and target sets, where each point is the nearest neighbor of the other. BBS has several key features that make it robust against complex geometric deformations and high levels of outliers, such as those arising from background clutter and occlusions. We study these properties, provide a statistical analysis that justifies them, and demonstrate the consistent success of BBS on a challenging real-world dataset.

1. Introduction

Finding a template patch in a target image is a core component in a variety of computer vision applications such as object detection, tracking, image stitching and 3D reconstruction. In many real-world scenarios, the template—a bounding box containing a region of interest in the source image —undergoes complex deformations in the target image: the background can change and the object may undergo nonrigid deformations and partial occlusions.

Template matching methods have been used with great success over the years but they still suffer from a number of drawbacks. Typically, all pixels (or features) within the template and a candidate window in the target image are taken into account when measuring their similarity. This is undesirable in some cases, for example, when the background behind the object of interest changes between the template and the target image (see Fig. 1). In such cases, the dissimilarities between pixels from different backgrounds may be arbitrary, and accounting for them may lead to false detections of the template (see Fig. 1(b)).

In addition, many template matching methods assume a specific parametric deformation model between the template and the target image (e.g., rigid, affine transformation, etc.). This limits the type of scenes that can be handled, and may require estimating a large number of parameters when complex deformations are considered.

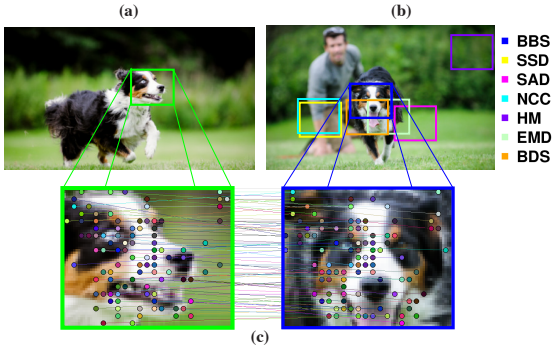


Figure 1. **Best-Buddies Similarity (BBS) for Template Matching:** (a), The template, marked in green, contains an object of interest against a background. (b), The object in the target image undergoes complex deformation (background clutter and large geometric deformation); the detection results using different similarity measures are marked on the image (see legend); our result is marked in blue. (c), The Best-Buddies Pairs (BBPs) between the template and the detected region are mostly found the object of interest and not on the background; each BBP is connected by a line and marked in a unique color.

In this paper, we propose a new method to address these problems, and show that it can be applied successfully to template matching *in the wild*. Specifically, we introduce a novel similarity measure termed *Best-Buddies Similarity (BBS)*, analyze its key features, and perform extensive evaluation of its performance compared to a number of commonly used alternatives on a challenging data set.

BBS measures the similarity between two sets of points in $\mathbb{R}^d$. A key feature of this measure is that it relies only on a subset (usually small) of pairs of points – the *Best-Buddies Pairs (BBPs)*. A pair of points is considered a BBP if each point is the nearest neighbor of the other in the corresponding point set. BBS is then taken to be the fraction of BBP out of all the points in the set.

[*] The first two authors contributed equally to this work

[†] Part of this work was done while the author was at Microsoft Research

Albeit simple, this measure turns out to have important and nontrivial properties. Because BBS counts only the pairs of points that are best buddies, it is robust to significant amounts of outliers. Another, less obvious property is that the BBS between two point sets is maximal when the points are drawn from the same distribution, and drops sharply as the distance between the distributions increases. In other words, if two points are BBP, they were likely drawn from the same distribution. We provide a statistical formulation of this observation, and analyze it numerically in the 1D case for point sets drawn from distinct Gaussian distributions (often used as a simplified model for natural images). The ability of BBS to reliably match features coming from the same distribution, in the presence of outliers, makes it highly attractive for robust template matching under visual changes and geometric deformations.

We apply the BBS measure for template matching by representing both the template and each of the candidate image regions as point sets in a joint $xyRGB$ space. BBS is used to measure the similarity between the two sets of points in this location-appearance space. The aforementioned properties of BBS now readily apply to template matching. That is, pixels on the object of interest in both the template and the candidate patch can be thought of as originating from the same underlying distribution. These pixels in the template are likely to find best buddies in the candidate patch, and hence would be considered as inliers. In contrast, pixels that come from different distributions, e.g., pixels from different backgrounds, are less likely to find best buddies, and hence would be considered outliers (see Fig. 1(c)). Given this important property, BBS bypasses the need to explicitly model the underlying object appearance and deformation.

To summarize, the main contributions of this paper are: (a) introducing BBS – a useful, robust, parameter-free measure for template matching in unconstrained environments, (b) analysis providing theoretical justification of its key features, and (c) extensive evaluation on challenging real data and comparison to a number of commonly used template matching methods.

2. Related Work

Template matching algorithms depend heavily on the similarity measure used to match the template and a candidate window in the target image. Various similarity measures have been used for this purpose. The most popular are the Sum of Squared Differences (SSD), Sum of Absolute Differences (SAD) and Normalized Cross-Correlation (NCC), mostly due to their computational efficiency [14]. Different variants of these measures have been proposed to deal with illumination changes and noise [7, 6].

Another family of measures is composed of robust error functions such as M-estimators [2, 20] or Hamming-based distance [19, 15], which are less affected by additive noise and 'salt and paper' outliers than cross correlation related methods. However, all the methods mentioned so far assume a strict rigid geometric deformation (only translation) between the template and the target image, as they penalize pixel-wise differences at corresponding positions in the template and the query region.

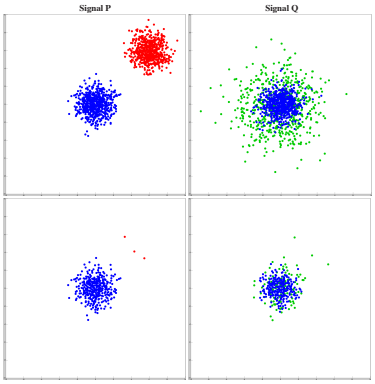


Figure 2. **Best-Buddies Pairs (BBPs) between 2D Gaussian Signals:** First row, Signal P consists of "foreground" points drawn from a normal distribution, $N(\mu_1, \sigma_1)$, marked in blue; and "background" points drawn from $N(\mu_2, \sigma_2)$, marked in red. Similarly, the points in the second signal Q are drawn from the same distribution $N(\mu_1, \sigma_1)$, and a different background distribution $N(\mu_3, \sigma_3)$. The color of points is for illustration only, i.e., BBS does not know which point belongs to which distribution. Second row, only the BBPs between the two signals which are mostly found between foreground points.

A number of methods extended template matching to deal with parametric transformations (e.g., [23, 10]). Recently, Korman *et al.* [11] introduced a template matching algorithm under 2D affine transformation that guarantees an approximation to the globally optimal solution. Likewise, Tian and Narasimhan [22] find a globally optimal estimation of nonrigid image distortions. However, these methods assume a one-to-one mapping between the template and the query region for the underlying transformation. Thus, they are prone to errors in the presence of many outliers, such as those caused by occlusions and background clutter. Furthermore, these methods assume a parametric model for the distortion geometry, which is not required in the case of BBS.

Measuring the similarity between color histograms, known as Histogram Matching (HM), offers a non-parametric technique for dealing with deformations and is commonly used in visual tracking [3, 16]. Yet, HM completely disregards geometry, which is a powerful cue. Further, all pixels are evenly treated. Other tracking methods have been proposed to deal with cluttered environments and partial occlusions [1, 9]. But unlike tracking, we are interested in detection in a single image, which lacks the redundant temporal information given in videos.

Olson [12] formulated template matching in terms of

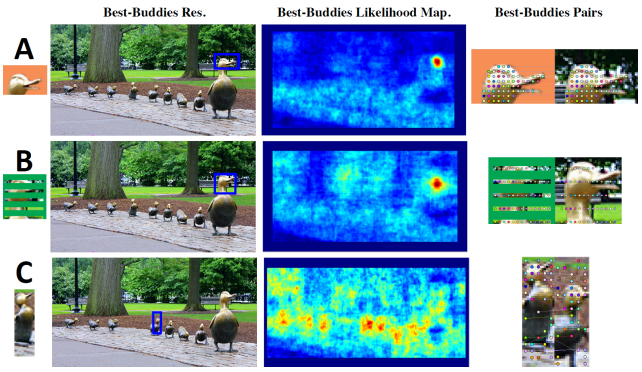


Figure 3. **BBS template matching results.** Three toys examples are shown: (A) cluttered background, (B) occlusions, (C) nonrigid deformation. The template (first column) is detected in the target image (second column) using the BBS; the results using BBS are marked in a blue. The likelihood maps (third column) show well-localized distinct modes. The BBPs are shown in last column. See text for more details.

maximum likelihood estimation, where an image is represented in a 3D location-intensity space. Taking this approach one step further, Oron *et al.*[13] use $xyRGB$ space and reduced template matching to measuring the EMD [18] between two point sets. Although BBS works in the same space, it differs from EMD, which requires $1:1$ matching and does not distinguish between inliers and outliers.

The BBS is a bi-directional measure. The importance of such two-side agreement has been demonstrated by the Bidirectional similarity (BDS) in [21] for visual summarization. Specifically, the BDS was used as a similarity measure between two images, where an image is represented by a set of patches. The BDS sums over the distances between each patch in one image to its nearest neighbor in the other image, and vice versa. In contrast, the BBS is based on a *count* of the BBPs, and makes only implicit use of their actual distance. Moreover, the BDS does not distinguish between inliers and outliers. These proprieties makes the BBS a more robust and reliable measure as demonstrated by our experiments.

In the context of image matching, another widely used measure is the Hausdorff distance [8]. To deal with occlusions or degradations, Huttenlocher *et al.* [8] proposed a fractional Hausdorff distance in which the K^{th} farthest point is taken instead of the most farthest one. Yet, this measure highly depends on K that needs to be tuned. Alternatively, Dubuisson and Jain [5] replace the max operator with sum.

It is worth mentioning, that the term *Best Buddies* was used by Pomeranz *et al.* [17] in the context of solving jigsaw puzzles. Specifically, they used a metric similar to ours in order to determine if a pair of pieces are compatible with each other.

3. Method

Our goal is to match a template to a given image, in the presence of high levels of outliers (i.e., background clutter, occlusions) and nonrigid deformation of the object of interest. We follow the traditional sliding window approach and compute the Best-Buddies Similarly (BBS) between the template and every possible window (of the size of the template) in the image. In the following, we give a general definition of BBS and demonstrate its key features via simple intuitive toy examples. We then statistically analyze these features in Sec. 4.

General Defination: BBS measures the similarity between two sets of points $P=\{p_i\}_{i=1}^{N}$ and $Q=\{q_i\}_{i=1}^{M}$, where $p_i, q_i \in \mathbb{R}^d$. The BBS is the fraction of *Best-Buddies Pairs* (BBPs) between the two sets. Specifically, a pair of points $\{p_i \in P, q_j \in Q\}$ is a BBP if p_i is the nearest neighbor of q_j in the set Q, and vice versa. Formally,

$$bb(p_i, q_j, P, Q) = \begin{cases} 1 & \mathrm{NN}(p_i, Q) = q_j \wedge \mathrm{NN}(q_j, P) = p_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where, $\mathrm{NN}(p_i, Q) = \underset{q \in Q}{\operatorname{argmin}}\, d(p_i, q)$, and $d(p_i, q)$ is some distance measure. The BBS between the point sets P and Q is given by:

$$\mathrm{BBS}(P, Q) = \frac{1}{\min\{\mathrm{M},\mathrm{N}\}} \cdot \sum_{i=1}^{N} \sum_{j=1}^{M} bb(p_i, q_j, P, Q). \tag{2}$$

The key properties of the BBS are: 1) it relies only on a (usually small) subset of matches i.e., pairs of points that are BBPs, whereas the rest are considered as outliers. 2) BBS finds the bi-directional inliers in the data without any prior knowledge on the data or its underlying deformation. 3) BBS uses *rank*, i.e., it counts the number of BBPs, rather than using the actual distance values.

To understand why these properties are useful, let us consider a simple 2D case of two point sets P and Q. The set P consist of 2D points drawn from two different normal distributions, $N(\mu_1, \Sigma_1)$, and $N(\mu_2, \Sigma_2)$. Similarly, the points in Q are drawn from the same distribution $N(\mu_1, \Sigma_1)$, and a different distribution $N(\mu_3, \Sigma_3)$ (see first row in Fig. 2). The distribution $N(\mu_1, \Sigma_1)$ can be treated as a *foreground* model, whereas $N(\mu_2, \Sigma_2)$ and $N(\mu_3, \Sigma_3)$ are two different *background* models. As can be seen in Fig. 2, the BBPs are mostly found between the foreground points in P and Q. For set P, where the foreground and background points are well separated, 95% of the BBPs are foreground points. For set Q, despite the significant overlap between foreground and background, 60% of the BBPs are foreground points.

This example demonstrates the robustness of BBS to high level of outliers in the data. BBS captures the foreground points and does not force the background points to match. By doing so, BBS sidesteps the need to model

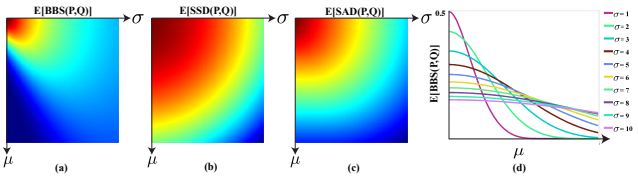


Figure 4. **The expectation of BBS in the 1D Gaussian case:** Two point sets, P and Q, are generated by sampling points from $N(0,1)$, and $N(\mu,\sigma)$, respectively. (a), the approximated expectation of BBS(P,Q) as a function of σ (x-axis), and μ (y-axis).(b)-(c), the expectation of SSD(P,Q), and SAD(P,Q), respectively. (d), the expectation of BBS as a function of μ plotted for different σ.

the background/foreground parametrically or have a prior knowledge about their underlying distributions. Furthermore, it shows that a pair of points $\{p,q\}$ is more likely to be BBP if p and q are drawn from the same distribution. We formally prove this general argument for the 1D case in Sec. 3.1. With this observations in hand, we continue with the use of BBS for template matching.

BBS for Template Matching: To apply BBS to template matching, one needs to convert each image patch to a point set in $\mathbb{R}^d$. To this end, we represent an image window in a spatial-appearance space. That is, we break the region into $k \times k$ distinct patches. Each $k \times k$ patch is represented by a k^2 vector of its RGB values and xy location of the central pixel, relative to the patch coordinate system (see Sec. 3.2 for more details). However, our method is not restricted to this particular representation and others can be used.

Following the intuition presented in the 2D Gaussian example (see Fig. 2), the use of BBS for template matching allows us to overcome several significant challenges such as background clutter, occlusions, and nonrigid deformation of the object. This is demonstrated in three synthetic examples shown in Fig. 3. The templates A and B include the object of interest in a cluttered background, and under occlusions, respectively. In both cases the templates are successfully matched to the image despite the high level of outliers. As can be seen, the BBPs are found only on the object of interest, and the BBS likelihood maps have a distinct mode around the true location of the template. In the third example, the template C is taken to be a bounding box around the forth duck in the original image, which is removed from the searched image using inpating techniques. In this case, BBS matches the template to the fifth duck, which can be seen as a nonrigid deformed version of the template. Note that the BBS does not aim to solve the pixel correspondence. In fact, the BBPs are not necessarily semantically correct (see third row in Fig. 3), but rather pairs of points that likely originated from the same distribution. This property, which we next formally analyze, helps us deal with complex visual and geometric deformations in the presence of outliers.

3.1. Analysis

So far, we have empirically demonstrated that the BBS is robust to outliers, and results in well-localized modes. Here, we give a statistical analysis that justifies these properties, and explains why using the count of the BBP is a good similarity measure.

We begin with a simple mathematical model in 1D, in which an "image" patch is modeled as a set of points drawn from a general distribution. Using this model, we derive the expectation of BBS between two sets of points, drawn from two given distributions $f_P(p)$ and $f_Q(q)$, respectively. We then analyze numerically the case in which $f_P(p)$, and $f_Q(q)$ are two different normal distributions. Finally, we relate these results to the multi-dimentional case. We show that the BBS distinctively captures points that are drawn from similar distributions. That is, we prove that the likelihood of a pair of points being BBP, and hence the expectation of the BBS, is maximal when the points in both sets are drawn from the same distribution, and drops sharply as the distance between the two normal distributions increases.

One-dimentional Case: Following Eq. 2, the expectation BBS(P,Q), over all possible samples of P and Q is given by:

$$E[\text{BBS}(P,Q)] = \frac{1}{\min\{M,N\}} \sum_{i=1}^{N} \sum_{j=1}^{M} E[bb_{i,j}(P,Q)], \tag{3}$$

where $bb_{i,j}(P,Q)$ is defined in Eq. 1. We continue with computing the expectation of a pair of points to be BBP, over all possible samples of P and Q, denoted by E_{BBP}. That is,

$$E_{\text{BBP}} = \iint_{P,Q} bb_{i,j}(P,Q) \Pr\{P\} \Pr\{Q\} dP dQ, \tag{4}$$

This is a multivariate integral over all points in P and Q. However, assuming each point is independent of the others this integral can be simplified as follows.

Claim:

$$E_{\text{BBP}} = \int_{-\infty}^{\infty}\int (F_Q(p^-)+1-F_Q(p^+))^{M-1} \cdot (F_P(q^-)+1-F_P(q^+))^{N-1} f_P(p) f_Q(q) dp dq, \tag{5}$$

where, $F_P(x)$, and $F_Q(x)$ denote the CDFs of P and Q, respectively. That is, $F_P(x) = \Pr\{p \le x\}$. And, $p^- = p - d(p,q)$, $p^+ = p + d(p,q)$, and q^+, q^- are similarly defined.

Proof: Due to the independence between the points, the integral in Eq.4 can be decoupled as follows:

$$E_{\text{BBP}} = \int_{p_1} \cdots \int_{p_N} \int_{q_1} \cdots \int_{q_M} bb_{i,j}(P,Q) \prod_{k=1}^{N} f_P(p_k) \prod_{l=1}^{M} f_Q(q_l) dP dQ \tag{6}$$

With abuse of notation, we use $dP = dp_1 \cdot dp_2 \cdots dp_N$, and $dQ = dq_1 \cdot dq_2 \cdots dq_M$. Let us consider the function

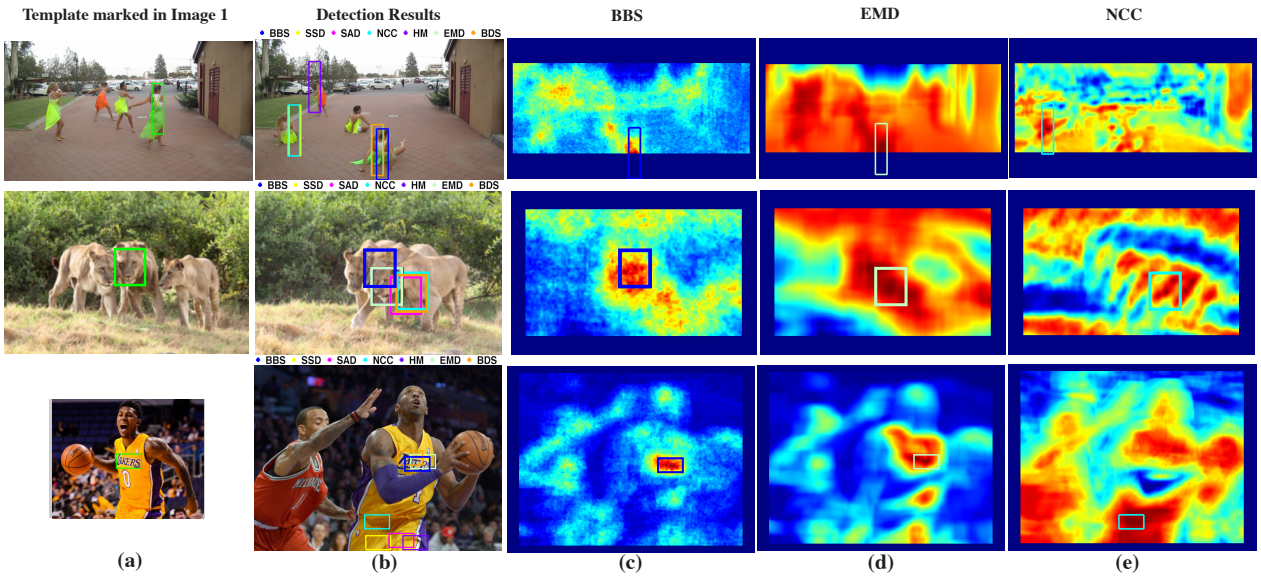


Figure 5. **BBS results on Real Data:** (a), the templates are marked in green over the input images. (b) the target images marked with the detection results of 6 different methods (see text for more details). BBS results are marked in blue. (c)-(e), the resulting likelihood maps using BBS, EMD and NCC , respectively; each map is marked with the detection result, i.e., its global maxima.

$bb_{i,j}(P,Q)$ for a given realization of P and Q. By definition, this indicator function equals 1 when p_i and q_j are nearest neighbors of each other, and zero otherwise. This can be expressed in terms of the distance between the points as follows:

$$bb_{i,j}(P,Q) = \prod_{k\neq i,k=1}^{N} \mathbb{I}[d(p_k,q_j) > d(p_i,q_j)] \prod_{l\neq j,l=1}^{M} \mathbb{I}[d(q_l,p_i) > d(p_i,q_j)] \tag{7}$$

where $\mathbb{I}$ is an indicator function. It follows that for a given value of p_i and q_j, the contribution of p_k to the integral in Eq. 6 can be decoupled. Specifically, we define:

$$Cp_k = \int_{-\infty}^{\infty} \mathbb{I}[d(p_k,q_j) > d(p_i,q_j)] f_P(p_k) dp_k \tag{8}$$

Assuming $d(p,q) = \sqrt{(p-q)^2} = |p-q|$, the latter can be written as:

$$Cp_k = \int_{-\infty}^{\infty} \mathbb{I}[p_k < q_j^- \vee p_k > q_j^+] f_P(p_k) dp_k \tag{9}$$

where $q_j^- = q_j - d(p_i,q_j)$, $q_j^+ = q_j + d(p_i,q_j)$. Since $q_j^- < q_j^+$, it can be easily shown that Cp_k can be expressed in terms of $F_P(x)$, the CDF of P:

$$Cp_k = F_P(q_j^-) + 1 - F_P(q_j^+) \tag{10}$$

The same derivation hold for computing Cq_l, the contribution of q_l to the integral in Eq. 6, given p_i, and q_j. That is,

$$Cq_l = F_Q(p_i^-) + 1 - F_Q(p_i^+) \tag{11}$$

where p_i^-, p_i^+ are similarly defined and $F_Q(x)$ is the CDF of Q. Note that Cp_k and Cq_l depends only on p_i and q_j and on the underlying distributions. Therefore, Eq. 6 results in:

$$\begin{aligned} E_{\text{BBP}} &= \iint_{p_i,q_j} dp_i dq_j f_P(p_i) f_Q(q_j) \prod_{k=1,k\neq i}^{N} Cp_k \prod_{l=1,l\neq j}^{N} Cq_l \\ &= \iint_{p_i,q_j} dp_i dq_j f_P(p_i) f_Q(q_j) Cp_k^{N-1} Cq_l^{M-1} \end{aligned} \tag{12}$$

Substituting the expressions for Cp_k and Cq_l in Eq. 12, and omitting the subscripts i,j for simplicity, result in Eq. 5, which completes the proof.

In general, the integral in Eq. 5 does not have a closed form solution, but it can be solved numerically for selected underlying distributions. To this end, we proceed with Gaussian distributions, which are often used as simple statistical models of image patches. We then use Monte-Carlo integration to approximate E_{BBP} for discrete choices of parameters μ and σ of Q in the range of [0, 10] while fixing the distribution of P to have $\mu = 0, \sigma = 1$. We also fixed the number of points to $N = M = 100$. The resulting approximation for E_{BBP} as a function of the parameters μ, σ is shown in Fig. 4, on the left. As can be seen, E_{BBP} is the highest at $\mu = 0, \sigma = 1$, i.e., when the points are drawn from the same distribution, and drops rapidly as the the underlying distribution of Q deviates from $N(0,1)$.

Note that E_{BBP} does not depends on p and q (because of the integration, see Eq. 5. Hence, the expected value of the BBS between the sets (Eq. 3) is given by:

$$E[\text{BBS}(P,Q)] = c \cdot E_{\text{BBP}} \tag{13}$$

where $c = \frac{MN}{\min\{M,N\}}$ is constant.

We can compare the BBS to the expectation of SSD, and SAD. The expectation of the SSD has a closed form solution given by:

$$E[\text{SSD(P,Q)}] = \int_{-\infty}^{\infty}\int (p-q)^2 f_P(p) f_Q(q|k) dp dq = 1 + \mu^2 + \sigma^2. \tag{14}$$

Replacing $(p-q)^2$ with $|p-q|$ results in the expression of the SAD. In this case, the expected value reduces to the expectation of the Half-Normal distribution and is given by:

$$E[\text{SAD(P,Q)}] = \frac{1}{\sqrt{2\pi}}\sigma_K \exp^{-\mu^2/(2\sigma^2)} + \mu(1 - 2f_P(-\mu/\sigma)) \tag{15}$$

Fig. 4(b)-(c) shows the maps of the expected values for $1-\text{SSD}_n(P,Q)$, and $1-\text{SAD}_n(P,Q)$, where SSD_n, SAD_n are the expectation of SSD and SAD, normalized to the range of [0,1]. As can be seen, the SSD and SAD results in a much wider spread around their mode. Thus, we have shown that the likelihood of a pair of points to be a BBP (and hence the expectation of the BBS) is the highest when P and Q are drawn from the same distribution and drops sharply as the distance between the distributions increases. This makes the BBS a robust and distinctive measure that results in well-localized modes.

Multi-dimensional Case: With the result of the 1D case in hand, we can bound the expectation of BBS when P and Q are sets of multi-dimensional points, i.e., $p_i, q_j \in R^d$.

If the d-dimensions are uncorrelated (i.e., the covariance matrices are diagonals in the Gaussian case), a necessary (but not sufficient) condition for a pair of points to be BBP is that the point would be BBP in each of the dimensions. In this case, the analysis is done in each dimension independently according to the 1D case given earlier 5. The expectation of the BBS in the multi-dimensional case is then bounded by the product of the expectations in each of the dimensions. That is,

$$E_{\text{BBS}} \geq \prod_{i=1}^{d} E^i_{\text{BBS}}, \tag{16}$$

where E^i_{BBS} denote the expectation of BBS in the i^{th} dimension. This means that the BBS is expected to be more distinctive, i.e., to drop faster as d increases. Note that if a pair of points is not a BBP in one of the dimensions, it does not necessarily imply that the multi-dimentional pair is not BBP. Thus, this condition is necessary but not sufficient.

3.2. Implementation Details and Complexity

Computing the BBS between two point sets $P, Q \in \mathbb{R}^d$, requires computing the distance between each pair of points. That is, constructing a distance matrix D where

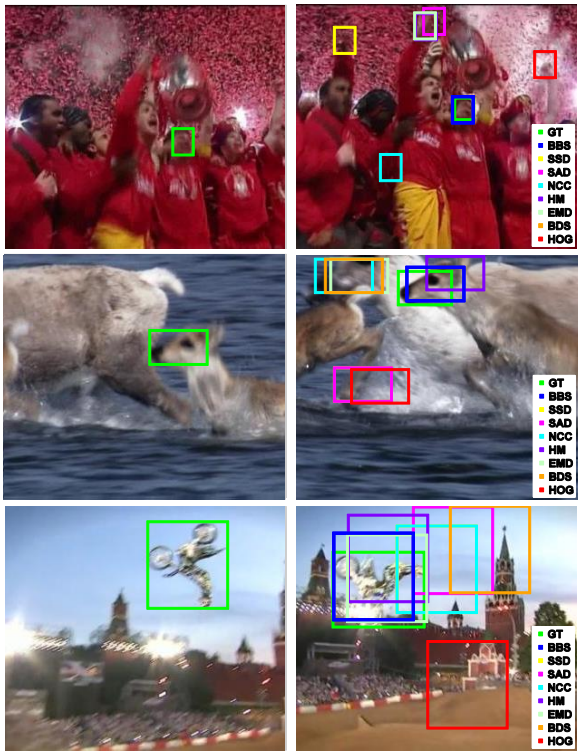


Figure 6. **Examples results with annotated data**. Left, input images with the annotated template marked in green. Right, target images and the detected bounding boxes (see legend); the ground-truth (GT) marked in green (our results in blue). BBS successfully match the template in all these examples.

$[D]_{i,j} = d(p_i, q_j)$. Given D, the nearest neighbor of $p_i \in P$, i.e. $NN(p_i, Q)$, is the minimal element in the i^{th} row of D. Similarly, $NN(q_j, P)$ is the minimal element in the j^{th} column of D. BBS is then computed by counting the number of mutual nearest neighbors (divided by a constant).

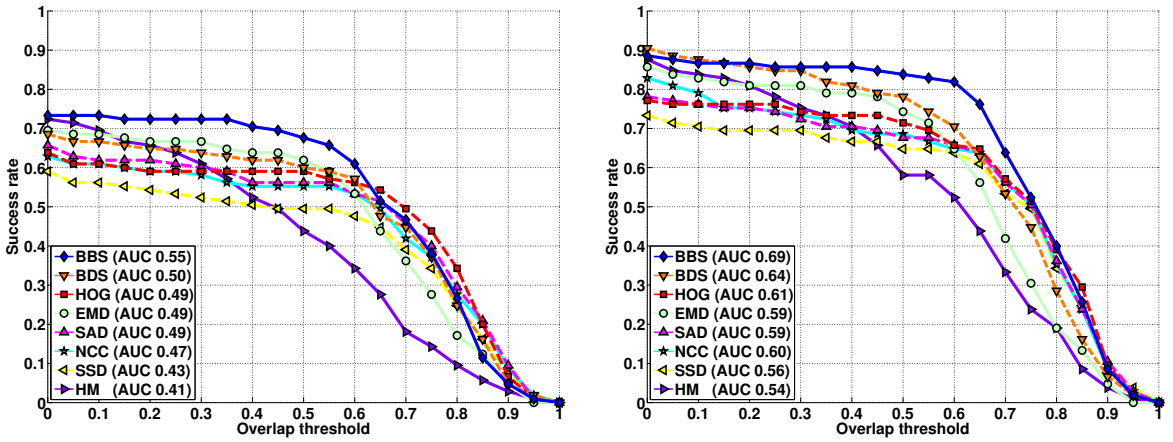
The distance measure used in our experiments is:

$$d(p_i, q_j) = ||p_i^{(A)} - q_j^{(A)}||_2^2 + \lambda||p_i^{(L)} - q_j^{(L)}||_2^2 \tag{17}$$

where superscript A denotes pixel appearance (e.g. RGB) and superscript L denotes pixel location $(x, y$ within the patch normalized to the range $[0, 1])$. $\lambda = 2$ was chosen empirically and was fixed in all of our experiments.

As previously mentioned, we break both image and template into $k \times k$ distinct patches, however for clarity we first analyze BBS complexity using all the individual pixels and only then extend it to the $k \times k$ non-overlapping patch case. Assuming $|P| = N$ and $|Q| = M$, the complexity of computing D is $O(dNM)$. Given that the image size is $|I| = L$ then constructing all L distance matrices D would require $O(dNML)$. Fortunately, computing D from scratch for each window in the image is not required as many computations can be reused.

In practice, we scan the image column by column, and



(a) Success using maximum confidence (b) Success of best out of top 7 matches

Figure 7. **Accuracy:** Success curves showing the fraction of examples with overlap $>$ TH $\in [0, 1]$. (a), using only the most likely target position (global max). (b), using the best out of top 7 modes of the confidence map. Area-under-curve (AUC) shown in the legend.

buffer all the distance matrices computed for the previous column. By doing so, all distance matrices in a new column, except the first one, require computing the distance between just *one* new pixel in Q and the template P which is done in $O(N)$ (the rest of the required distances were already computed in the previous column). Since the template is smaller than the image (typically $L >> N, M$), the complexity is dominated by $O(dNL)$ which is typically two orders of magnitude smaller than $O(dNML)$.

It is now easy to see that in the case where $k \times k$ distinct patches are used (instead of individual pixels) then $p_i^{(A)}, q_j^{(A)} \in \mathbb{R}^{(k \times k \times d)}$. In which case the dominant complexity term becomes $O(k^2 d \frac{N}{k^2} \frac{L}{k^2}) = O(\frac{dNL}{k^2})$. Using $k \times k$ patches results in a higher dimensional appearance space leading to more reliable BBPs and as can be seen by our analysis also reduces the computational complexity.

Using unoptimized Matlab code, the typical running time of our algorithm, with $k = 3$, is ~ 4 seconds for 360x480 image, and 40x30 template.

4. Results

We perform qualitative as well as extensive quantitative evaluation of our method on real world data. We compare BBS with six similarity measures commonly used for template matching. 1) Sum-of-Square-Difference (SSD), 2) Sum-of-Absolute-Difference (SAD), 3) Normalized-Cross-Correlation (NCC), 4) color Histogram Matching (HM), 5) Earth Movers Distance[18] (EMD), 6) Bidirectional Similarity [21] (BDS) computed in the same appearance-location space as BBS.

4.1. Qualitative Evaluation

Four template-image pairs taken from the Web are used for qualitative evaluation. The templates, which were manually chosen, and the target images are shown in Fig. 1(a)-(b), and in Fig. 5. In all examples, the template drastically changes its appearance due to large geometric deformation, partial occlusions, and change of background.

Detection results in Fig. 1(a)-(b), and in Fig. 5(b), show that BBS is the only method successfully matching the template in all these challenging examples. The confidence maps of BBS, presented in Fig. 5(c), show distinct and well-localized modes compared to other methods. Only EMD and NCC are shown for comparison due to space limitations[1]. The BBPs for the first example are shown in Fig. 1(c). As discussed in Sec. 3, BBS captures the bidirectional inliers, which are mostly found on the object of interest. Note that the BBPs, as discussed, are not necessarily true physical corresponding points.

4.2. Quantitative Evaluation

We now turn to the quantitative evaluation. The data for this experiment was generated from annotated video sequences previously used in Wu *et al.*[24]. The 35 color videos in this dataset capture a wide range of challenging scenes. The objects of interest are diverse and typically undergo nonrigid deformations, perform in/out-of-plane rotation and may be partially occluded.

For the purpose of template matching, 105 template-image pairs were sampled, three pairs per video. Each image pair consists of frames f and $f + 20$, where f was randomly chosen. The ground-truth annotated bounding box in frame f was used as template, and frame $f + 20$ was used as the target image. This random choice of frames creates a challenging benchmark with a wide baseline in both time and space (see examples in Fig. 6).

BBS was compared with the 6 similarity measures mentioned above. In addition, we add another similarity mea-

[1]Our data and code are publicly available at: *http://people.csail.mit.edu/talidekel/Best-BuddiesSimilarity.html*

sure that is based on SSD using dense Histogram-Of-Gradients (HOG) [4]. The ground-truth annotations were used for quantitative evaluation. Specifically, we measure the accuracy of both the top match ("accuracy") as well as the top k ranked matches ("rank-accuracy"), as follows.

Accuracy: was measured using the common bounding box overlap measure: $Acc. = \frac{\text{area}(B_e \cap B_g)}{\text{area}(B_e \cup B_g)}$ where B_e and B_g are the estimated and ground truth bounding boxes, respectively. The ROC curves show the fraction of examples with overlap larger than a threshold ($TH \in [0, 1]$), and the area-under-curve (AUC) measured quantifies overall accuracy. The success rates of all methods, based on the global maximum confidence score, are presented in Fig. 7-(a). As can be seen, BBS achieves the highest AUC score of 0.55 dominating the competing methods we have tested, with a significant margin for all threshold values ≤ 0.65. For overlap values > 0.7 the performance of all methods drops sharply. This can be attributed to the fact that overlap drops sharply for small errors and in our case using the non-overlapping patch representation generates an inherent uncertainty of 3 pixels in target localization.

We have relaxed the requirement that only the top match will be considered and tested the top 7 modes of the confidence map (instead of just the global maximum). That is, we test the 7 best matches and report the one with the highest accuracy score. See Fig. 7-(b). Again BBS outperforms competing methods reaching AUC of 0.69 and keeping a noticeable performance margin especially for threshold range $[0.3, 0.75]$. Some examples that demonstrate the power of BBS are shwon in Fig. 6.

Rank-Accuracy: For each rank k, we compute the average target position based on all top k scores, compute the accuracy, and take the median accuracy over all 105 target-image pairs. We expect methods having distinct and well-localized modes to show moderate performance decrease as more and more positions are considered for position localization. However, for methods in which modes are not well localized (i.e. where peaks are broad and the difference in confidence between the correct location and other locations is very small) we expect a more rapid drop in accuracy. The analysis was performed for all methods with k ranging from 1 to 500.

This test relates to the claim made in Sec. 3.1, where we proved that in the 1D case, BBS drops sharply as the distance between the foreground and background distributions increases. In the context of template matching this means we expect the confidence maps of BBS to have distinct and well localized modes as is the case in the example shown in Fig. 5.

Results are presented in Fig. 8, and as expected, for methods such as SSD, SAD and NCC, which typically do not have well-localized modes, the score decreases rapidly as k increases. More robust methods such as EMD, BDS and BBS have a moderate accuracy decrease as k increases. Note that BBS which has high accuracy and distinct modes shows the best performance among the methods tested.

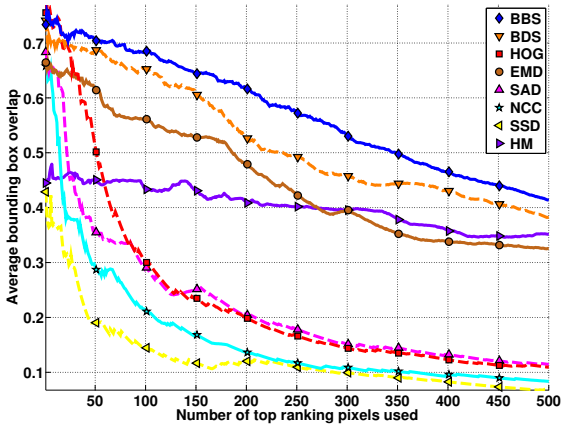


Figure 8. **Rank-accuracy:** For each rank k, we compute the average target position based on all top k scores, compute the accuracy, and take the median accuracy over all 105 template-image pairs. BBS dominates all other methods for any given value of k. Some methods show fast decay in accuracy relative to an increase in k indicating non-distinct modes in the confidence map which suggests such method are less robust.

5. Conclusions

We introduced a new measure, Best-Buddies Similarity (BBS), for template matching in the wild. We identified its key features, analyzed them and demonstrated the ability of BBS to overcome several challenges that are common in real scenes. We showed that our method outperforms a number of commonly used methods for template matching such as normalized cross correlation, histogram matching and EMD.

Our method may fail when the template is very small compared to target image, or when the outliers (occluding object or background clutter) cover most of the template.

In the scope of this paper, we worked in $xyRGB$ space, but other feature spaces may be used such as HOG features, edges or filter responses. This opens the use of BBS to other domains in computer-vision that could benefit from its properties. A natural future direction of research is to explore the use of BBS as an image similarity measure or for object localization.

Acknowledgments.

This work was supported in part by an Israel Science Foundation grant 1556/10, National Science Foundation Robust Intelligence 1212849 Reconstructive Recognition, and a grant from Shell Research.

References

[1] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. *CVPR*, 2012. 2

[2] J.-H. Chen, C.-S. Chen, and Y.-S. Chen. Fast algorithm for robust template matching with m-estimators. *Signal Processing, IEEE Transactions on*, 2003. 2

[3] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, 2000. 2

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 8

[5] M.-P. Dubuisson and A. Jain. A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568 vol.1, Oct 1994. 3

[6] E. Elboher and M. Werman. Asymmetric correlation: a noise robust similarity measure for template matching. *Image Processing, IEEE Transactions on*, 2013. 2

[7] Y. Hel-Or, H. Hel-Or, and E. David. Matching by tone mapping: Photometric invariant template matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2):317–330, 2014. 2

[8] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(9):850–863, 1993. 3

[9] X. Jia, H. Lu, and M. Yang. Visual tracking via adaptive structural local sparse appearance model. *CVPR*, 2012. 2

[10] H. Y. Kim and S. A. De Araújo. Grayscale template-matching invariant to rotation, scale, translation, brightness and contrast. In *AIVT*. Springer, 2007. 2

[11] S. Korman, D. Reichman, G. Tsur, and S. Avidan. Fast-match: Fast affine template matching. In *CVPR*, 2013. 2

[12] C. F. Olson. Maximum-likelihood image matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(6):853–857, 2002. 2

[13] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan. Locally orderless tracking. *IJCV*, 2014. 3

[14] W. Ouyang, F. Tombari, S. Mattoccia, L. Di Stefano, and W.-K. Cham. Performance evaluation of full search equivalent pattern matching algorithms. *PAMI*, 2012. 2

[15] O. Pele and M. Werman. Robust real-time pattern matching using bayesian sequential hypothesis testing. *PAMI*, 2008. 2

[16] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV 2002*. 2002. 2

[17] D. Pomeranz, M. Shemesh, and O. Ben-Shahar. A fully automated greedy square jigsaw puzzle solver. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 9–16, 2011. 3

[18] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 2000. 3, 7

[19] B. G. Shin, S.-Y. Park, and J. J. Lee. Fast and robust template matching algorithm in noisy image. In *Control, Automation and Systems, 2007. ICCAS'07. International Conference on*, 2007. 2

[20] A. Sibiryakov. Fast and high-performance template matching method. In *CVPR*, 2011. 2

[21] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *CVPR*, 2008. 3, 7

[22] Y. Tian and S. G. Narasimhan. Globally optimal estimation of nonrigid image distortion. *IJCV*, 2012. 2

[23] D.-M. Tsai and C.-H. Chiang. Rotation-invariant pattern matching using wavelet decomposition. *Pattern Recognition Letters*, 2002. 2

[24] Y. Wu, J. Lim, and M. Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 7