Detecting Irregularities in Images and in Video

Oren Boiman **Michal Irani**
Dept. of Computer Science and Applied Math
The Weizmann Institute of Science
76100 Rehovot, Israel

Abstract

We address the problem of detecting irregularities in visual data, e.g., detecting suspicious behaviors in video sequences, or identifying salient patterns in images. The term "irregular" depends on the context in which the "regular" or "valid" are defined. Yet, it is not realistic to expect explicit definition of all possible valid configurations for a given context. We pose the problem of determining the validity of visual data as a process of constructing a puzzle: We try to compose a new observed image region or a new video segment ("the query") using chunks of data ("pieces of puzzle") extracted from previous visual examples ("the database"). Regions in the observed data which can be composed using large contiguous chunks of data from the database are considered very likely, whereas regions in the observed data which cannot be composed from the database (or can be composed, but only using small fragmented pieces) are regarded as unlikely/suspicious. The problem is posed as an inference process in a probabilistic graphical model. We show applications of this approach to identifying saliency in images and video, and for suspicious behavior recognition.

1 Introduction

Detection of irregular visual patterns in images and in video sequences is useful for a variety of tasks. Detecting *suspicious behaviors* or *unusual objects* is important for surveillance and monitoring. Identifying *spatial saliency* in images is useful for quality control and automatic inspection. *Behavioral saliency* in video is useful for drawing the viewer's *attention*.

Previous approaches to recognition of suspicious behaviors or activities can broadly be classified into two classes of approaches: *rule-based methods* (e.g., [7]) and *statistical methods* without predefined rules (e.g., [10, 12]). The statistical methods are more appealing, since they do not assume a predefined set of rules for all valid configurations. Instead, they try to automatically learn the notion of regularity from the data, and thus infer about the suspicious. Nevertheless, the representations employed in previous methods have been either very restrictive (e.g., trajectories of moving objects [10]), or else too global (e.g., a single small descriptor vector for an entire frame [12]).

In this paper we formulate the problem of detecting regularities and irregularities as the problem of composing (explaining) the new observed visual data (an image or a video sequence, referred to below as "query") using spatio-temporal patches extracted from previous visual examples (the "database"). Regions in the query which can be composed using large contiguous chunks of data from the example database are considered likely. The larger those regions are, the greater the likelihood is. Regions in the query which cannot be composed from the example database (or can be composed, but only using small fragmented pieces) are regarded as unlikely/suspicious. Our approach can thus infer and generalize from just a few examples, about the validity of a much larger context of image patterns and behaviors, even if those particular configurations have never been seen before. Local descriptors are extracted from small image or video patches (composed together to large ensembles of patches), thus allowing to quickly and efficiently infer about subtle but important local changes in behavior (e.g., a man walking vs. a man walking while pointing a gun). Moreover, our approach is capable of simultaneously identifying a valid behavior in one portion of the field of view, and a suspicious behavior in a different portion the field of view, thus highlighting only the detected suspicious regions within the frame, and not the entire frame. Such examples are shown in Section 6.

Inference from image patches or fragments has been previously employed in the task of class-based object recognition (e.g. [4, 1, 3]). A small number of informative fragments have been learned and preselected for a small number of pre-defined classes of objects. However, class-based representations cannot capture the overwhelming number of possibilities of composing unknown objects or behaviors in a scene, and are therefore not suitable for our underlying task of detecting irregularities.

Our approach can also be applied for detecting *saliency* in images and in video sequences. For example, given a single image *with no prior information*, we can measure the "validity" of each image region (the "query") relative to

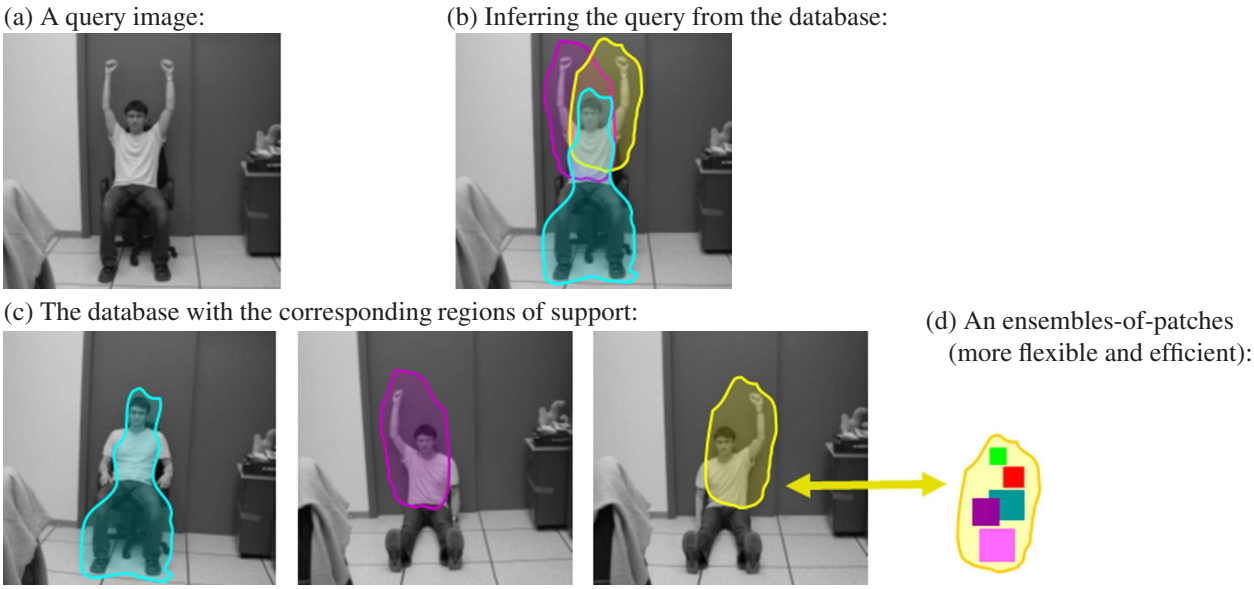


Figure 1. The basic concept – Inference by Composition. *A region in the query image is considered likely if it has a large enough contiguous region of support in the database. New valid image configurations can thus be inferred from the database, even though they have never been seen before.*

the remaining portions of the same image (the "database" used for this particular query). An image region will be detected as salient if it cannot be explained by anything similar in other portions of the image. Similarly, given a single video sequence (with no prior knowledge of what is a normal behavior), we can detect "salient behaviors" as behaviors which cannot be supported by any other dynamic phenomena occurring at the same time in the video.

Previous approaches for detecting image saliency (e.g., [6]) proposed measuring the degree of dissimilarity between an image location and its immediate surrounding region. Thus, for example, image regions which exhibit large changes in contrast are detected as salient image regions. Their definition of "visual attention" is derived from the same reasoning. Nevertheless, we believe that the notion of saliency is not necessarily determined by the immediate surrounding image regions. For example, a single yellow spot on a black paper may be salient. However, if there are many yellow spots spread all over the black paper, then a single spot will no longer draw our attention, even though it still induces a large change in contrast relative to its surrounding vicinity. Our approach therefore suggests a new and more intuitive interpretation of the term "saliency", which stems from the inner statistics of the entire image. Examples of detected spatial saliency in images and behavioral saliency in video sequences are also shown in Section 6.

Our paper therefore offers four main contributions:

1. We propose an approach for inferring and generalizing from just a few examples, about the validity of a much larger context of image patterns and behaviors, even if those particular configurations have never been seen before.
2. We present a new graph-based Bayesian inference algorithm which allows to efficiently detect *large ensembles of patches* (e.g., hundreds of patches), at multiple spatio-temporal scales. It simultaneously imposes constraints on the relative geometric arrangement of these patches in the ensemble as well as on their descriptors.
3. We propose a new interpretation to the term "saliency" and "visual attention" in images and in video sequences.
4. We present a single unified framework for treating several different problems in Computer Vision, which have been treated separately in the past. These include: attention in images, attention in video, recognition of suspicious behaviors, and recognition of unusual objects.

2 Inference by Composition

Given only a few examples, we (humans) have a notion of what is regular/valid, and what is irregular/suspicious, even when we see new configurations that we never saw before. We do not require explicit definition of all possible valid configurations for a given context. The notion of "regularity"/"validity" is learned and generalized from just a few examples of valid patterns (of behavior in video, or of appearance in images), and all other configurations are automatically inferred from those.

Fig. 1 illustrates the basic concept underlying this idea in the paper. Given a new image (a query – Fig. 1.a), we check whether each image region can be explained by a

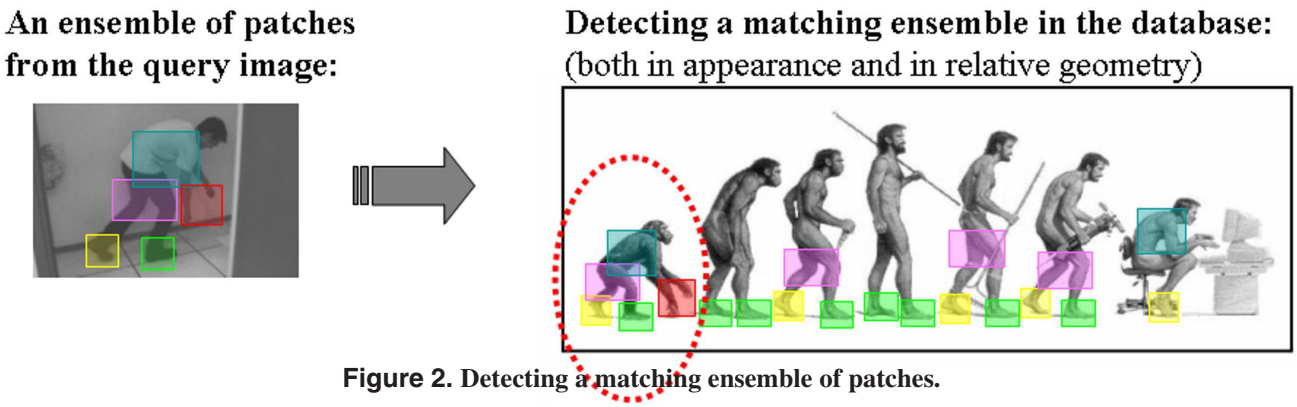


Figure 2. Detecting a matching ensemble of patches.

(a) A spatial ensemble:
(for queries on images)

(b) A space-time ensemble:
(for queries on video)

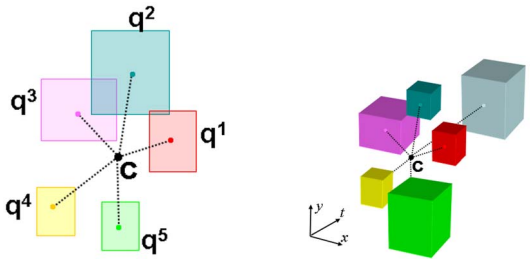


Figure 3. Ensembles of patches in images and video.

large enough contiguous region of support in the database (see Figs. 1.b and 1.c). Although we have never seen a man sitting with both arms raised, we can infer the validity of this pose from the three database images of Fig. 1.c.

Thus, regions in the new observed data/query (an image or a video sequence) which can be explained by large contiguous chunks of data from the database are considered very likely, whereas regions in the query which cannot be explained by large enough database pieces are considered unlikely or suspicious. When the visual query is an image, then those chunks of data have only a spatial extent. When the visual query is a video sequence, then those chunks of data have both a spatial and a temporal extent.

3 Ensembles of Patches

Human behaviors and natural spatial structures never repeat identically. For example, no two people walk in the same manner. One may raise his arms higher than the other, or may just walk faster.

We therefore want to allow for small non-rigid deformations (in space and in time) in our "pieces of puzzle" (chunks of data). This is particularly true for large chunks of data. To account for such local non-rigid deformations, large chunks are broken down to an *ensemble of lots of small patches* at multiple scales with their relative geometric positions. This is illustrated in Fig. 1.d. In the inference process, we search for a similar geometric configuration of patches with similar properties (of behavior, or of appearance), while allowing for small local misalignments in the relative geometric arrangement. This concept is illustrated in Fig. 2. When the visual query is an image, then an ensemble of patches is composed of spatial patches (see Fig. 3.a). When the visual query is a video sequence, then the ensemble of patches is composed of spatio-temporal patches (see Fig. 3.b), which allows to capture information about dynamic behaviors. In our current implementation, a single ensemble typically contains *hundreds* of patches, simultaneously from multiple scales (multiple spatial scales in the case of image patches, and multiple space-time scales in the case of spatio-temporal patches).

While the idea of composing new data from example patches was previously proven useful for a variety of tasks (e.g., [2, 5, 11]), these methods did not impose any geometric restriction on the example patches used for construction, i.e., their relative positions and distances in the database. This was not necessary for their purpose. It is however crucial here, for the purpose of detecting irregularities. Often, the only real cue of information for distinguishing between a likely and an unlikely phenomenon is the degree of fragmentation of its support in the database. For example, the stretched arm of a man holding a gun is similar to an instantaneous stretching of the arm while walking, but its region of support is very limited in time.

Capturing the geometric relations of patches was identified as being important for the task of class-based object recognition [1, 4, 3, 8]. Those approaches are not suitable for our objective for two reasons: (i) Their geometric configurations are restricted to a relatively small number of patches, thus cannot capture subtle differences which are crucial for detection of irregularities. (ii) Those configurations were pre-learned for a small number of pre-defined classes of objects, whereas our framework is applicable to any type of visual data. While the geometric constraints of [8] are more flexible, thus allowing to recognize new ob-

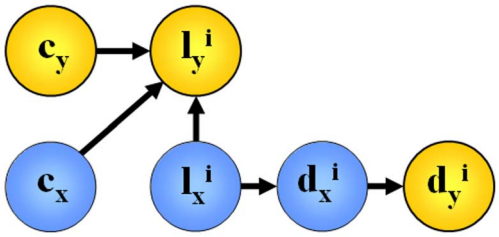


Figure 4. The probabilistic graphical model. *Observed variables are marked in "orange"; hidden database variables are marked in "blue". The directions of arrows signify Bayesian dependencies (See text for more details).*

ject configurations from just a few examples, their method is still limited to a set of predefined object classes with *predefined object centers*. This is not suitable for detecting irregularities, where there is no notion of object classes.

"Video Google" [9] imposes geometric constraints on large collections of non class-based descriptors, and searches for them very efficiently. However, those descriptors are spatial in nature and the search is restricted to individual image frames, thus not allowing to capture behaviors.

In order for the inference to be performed in reasonable times, information about the small patches and their relative arrangement must be efficiently stored in and extracted from the database. For each small patch extracted from the examples, a descriptor vector is computed and stored (see below), along with the *absolute* coordinates of the patch (spatial or spatio-temporal coordinates). Thus, the relative arrangement of all patches in the image/video database is implicitly available. Later, our inference algorithm takes an ensemble of patches from the visual query and searches the database for a similar configuration of patches (both in the descriptors and in their relative geometric arrangement). To allow for fast search and retrieval, those patches are stored in a multi-scale data structure. Using a probabilistic graphical model (Section 4), we present an efficient inference algorithm (Section 5) for the ensemble search problem.

Patch descriptors are generated for each query patch and for each database patch. The descriptors capture local information about appearance/behavior. Our current implementation uses very simple descriptors, which could easily be replaced by more sophisticated descriptors (e.g., "SIFT"):
The Spatial Image Descriptor of a small (e.g., 7×7) spatial patch is constructed as follows: The spatial gradient magnitude is computed for each pixel in the patch. These values are then stacked in a vector, which is normalized to a unit length. Such descriptors are densely extracted for each point in the image. This descriptor extraction process is repeated in several spatial scales of the spatial Gaussian pyramid of the image. Thus, a 7×7 patch extracted from a coarse scale has a larger spatial support in the input image (i.e., in the fine scale).

The Spatio-Temporal Video Descriptor of a small (e.g., $7 \times 7 \times 4$) spatio-temporal video patch is constructed from the absolute values of the temporal derivatives in all pixels of the patch. These values are stacked in a vector and normalized to a unit length. This descriptor extraction process is repeated in several spatial and temporal scales of a space-time video pyramid. Thus, a $7 \times 7 \times 4$ patch extracted from a coarse scale has a larger spatial and larger temporal support in the input sequence.

4 Statistical Formulation

Given a new visual query (an image or a video sequence), we would like to estimate the likelihood of each and every point in it. This is done by checking the validity of a large region (e.g., 50×50 region in an image, and $50 \times 50 \times 50$ region in a video sequence) surrounding *every* pixel. The large surrounding region is broken into lots (hundreds) of small patches at multiple scales (spatial or spatio-temporal), and is represented by a single ensemble of patches corresponding to that particular image/video point. Let $q^1, q^2, .., q^n$ denote the patches in the ensemble (see Fig. 3.a). Each patch q^i is associated with two types of attributes: (i) its descriptor vector d^i, and (ii) its location in absolute coordinates l^i. We choose an *arbitrary* reference point c (e.g., the center of the ensemble – see Fig. 3.a), which serves as the "origin" of the local coordinate system (thus defining the relative positions of the patches within the ensemble).

Let an observed ensemble of patches within the query be denoted by y. We would like to compute the joint likelihood $P(x, y)$ that the observed ensemble y in the query is similar to some hidden ensemble x in the database (similar both in its descriptor values of the patches, as well as in their relative positions). We can factor the joint likelihood as: $P(x, y) = P(y|x)P(x)$. Our modelling of $P(y|x)$ resembles the probabilistic modelling of the "star graph" of [3]. However, in the class-based setting of [3] what is computed is $P(y; \theta)$, where θ is a pre-learned set of parameters of a given patch-constellation of an object-class. In our case, however, there is no notion of objects, i.e., there is no prior parametric modelling of the database ensemble x. Thus, θ is undefined, and $P(x)$ must be estimated non-parametrically directly from the database of examples.

Let d_y^i denote the descriptor vector of the i-th observed patch in y, and l_y^i denote its location (in absolute coordinates). Similarly, d_x^i denotes the descriptor vector of the i-th hidden (database) patch in x, and l_x^i denotes its location. Let c_y and c_x denote the "origin" points of the observed and hidden ensembles. The similarity between any such pair of ensembles y and x is captured by the following likelihood:

$$P(x, y) = P(c_x, d_x^1, .., l_x^1, .., \quad c_y, d_y^1, .., l_y^1, ..) \qquad (1)$$

(a) The database images (3 poses):

(b) Query images:

(c) Red highlights the detected "unfamiliar" image configurations (unexpected poses):

(d) Color-association of the inferred query regions with the database images (determined by MAP assignment):
(Uniform patches are assumed valid by default – for added speedup).

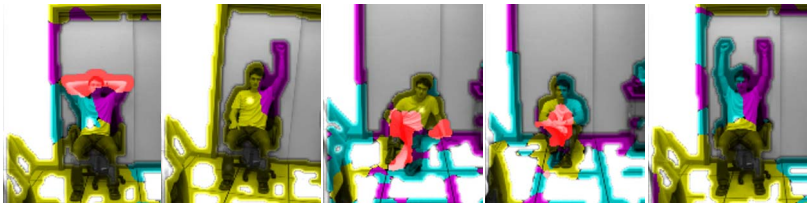


Figure 5. Detection of irregular image configurations. *New valid poses are automatically inferred from the database (e.g., a man sitting on the chair with both arms up, a man sitting on a chair with one arm up), even though they have never been seen before. New pose parts which cannot be inferred from the three database images are highlighted in red as being "unfamiliar".*

In order to make the computation of the likelihood in Eq. (1) tractable, we make some simplifying statistical assumptions. Given a hidden database patch and its descriptor d_x^i, the corresponding observed descriptor d_y^i is assumed to be independent of the other patch descriptors. (This is a standard Markovian assumption, e.g., [5], which is obviously not valid in case of overlapping patches.) We model the similarity between descriptors using a Gaussian distribution:

$$P\left(d_y^i|d_x^i\right) = \alpha_1 \exp\left(-(d_y^i - d_x^i)^T S_D^{-1}(d_y^i - d_x^i)\right) \quad (2)$$

where α_1 is a constant, and S_D is a constant covariance matrix, which determines the allowable deviation in the descriptor values. Given the *relative* location of the hidden database patch $(l_x^i - c_x)$, the relative location of the corresponding observed patch $(l_y^i - c_y)$ is assumed to be independent of all other patch locations. This assumption enables to compare the geometric arrangement of two ensembles of patches with enough flexibility to accommodate for small changes in viewing angle, scale, pose and behavior. Thus:

$$P\left(l_y^i|l_x^i, c_x, c_y\right) = \alpha_2 \cdot \exp\left(-((l_y^i - c_y) - (l_x^i - c_x))^T S_L^{-1}((l_y^i - c_y) - (l_x^i - c_x))\right) \quad (3)$$

where α_2 is a constant, and S_L is a constant covariance matrix, which captures the allowed deviations in the relative patch locations. (In this case the dependency in relative locations was modelled using a Gaussian, however the model is not restricted to that).

So far we modelled the relations between attributes *across ensembles* (descriptors: d_y^i, d_x^i, and relative locations: $l_y^i - c_y, l_x^i - c_x$). We still need to model the relations *within the hidden ensemble*, namely, the relations between a patch descriptor d_x^i to its location l_x^i. In the general case, this relation is highly non-parametric, and hence cannot be modelled analytically (in contrast to class-based approaches, e.g. [4, 3]). Therefore, we model it

(a) The input image: (b) The computed saliency map (- log likelihood): (c) The detected salient regions:

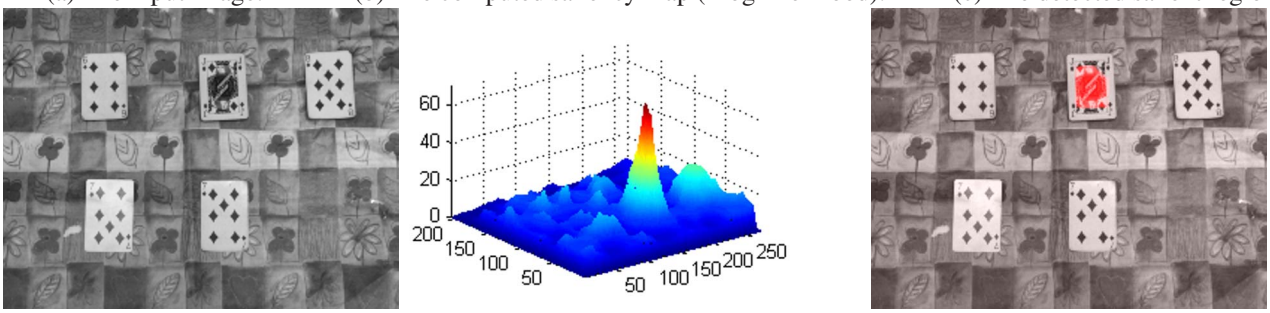


Figure 6. Identifying salient regions in a single image (no database; no prior information). *The Jack card was detected as salient. Note that even though the diamond cards are different from each other, none of them is identified as salient.*

non-parametrically using examples from the database:

$$P\left(d_x|l_x\right) = \begin{cases} 1 & (d_x, l_x) \in DB \\ 0 & otherwise \end{cases} \quad (4)$$

where d_x and l_x are an arbitrary descriptor and location.

We assume a uniform prior distribution for c_x and c_y (local origin points), i.e., no prior preference for the location of the ensemble in the database or in the query. The relation between all the above-mentioned variables is depicted in the Bayesian network in Fig. 4.

Thus, for an observed ensemble y and a hidden database ensemble x, we can factor the joint likelihood $P(x, y)$ of Eq. (1) using Eqs. (2,3,4) as follows:

$$\begin{aligned} &P(c_x, d_x^1, ..., l_x^1, ..., c_y, d_y^1, ..., l_y^1) = \\ &\quad \alpha \prod_i P(l_y^i|l_x^i, c_x, c_y) P(d_y^i|d_x^i) P(d_x^i|l_x^i) \end{aligned} \quad (5)$$

5 The Inference Algorithm

Given an observed ensemble, we seek a hidden database ensemble which maximizes its MAP (maximum a-posterior probability) assignment. This is done using the above statistical model, which has a simple and exact Viterbi algorithm. According to Eq. (5) the MAP assignment can be written as:

$$\begin{aligned} &\max_X P\left(c_x, d_x^1, ..., l_x^1, ..., c_y, d_y^1, ..., l_y^1\right) = \\ &\alpha \prod_i \max_{l_x^i} P\left(l_y^i|l_x^i, c_x, c_y\right) \max_{d_x^i} P\left(d_y^i|d_x^i\right) P\left(d_x^i|l_x^i\right) \end{aligned}$$

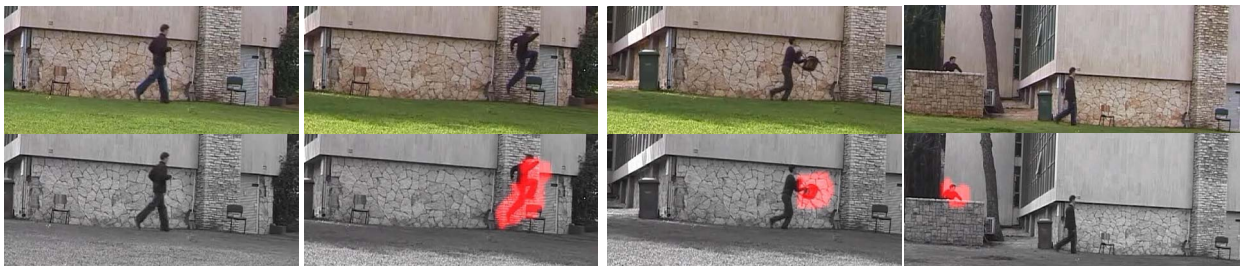
This expression can be phrased as a message passing (Belief Propagation) algorithm in the graph of Fig. 4. First we compute for each patch the message m_{dl}^i passed from node d_x^i to node l_x^i regarding its belief in the location l_x^i: $m_{dl}^i\left(l_x^i\right) = \max_{d_x^i} P\left(d_y^i|d_x^i\right) P\left(d_x^i|l_x^i\right)$. Namely, for each observed patch, compute all the candidate database locations l_x^i with high descriptor similarity. Next, for each of these candidate database locations, we pass a message about the induced possible origin locations c_x in the database: $m_{lc}^i\left(c_x\right) = \max_{l_x^i} P\left(l_y^i|l_x^i, c_x, c_y\right) m_{dl}\left(l_x^i\right)$. At this point, we have a candidate list of origins suggested by each individual patch. To compute the likelihood of an entire ensemble assignment, we multiply the beliefs from all the individual patches in the ensemble: $m_c\left(c_x\right) = \prod_i m_{lc}^i\left(c_x\right)$.

The progressive elimination process: A naive implementation of the above message passing algorithm is very inefficient, since independent descriptor queries are performed for each patch in the observation ensemble, regardless of answers to previous queries performed by other patches. These patches are related by a certain geometric arrangement. We therefore use this knowledge for an efficient search by *progressive elimination* of the search space in the database: We compute the message m_{dl}^i for a small number of patches (e.g., 1). The resulting list of possible candidate origins induces a very restricted search space for the next patch. The next patch, in turn, eliminates additional origins from the already short list of candidates, etc. In order to speed-up the progressive elimination, we use *truncated* Gaussian distributions (truncated after 4σ). Thus, if n is the number of patches in the ensemble (e.g., 256), and N is the number of patches in the database (e.g., $100,000$ patches for a one-minute video database), then the search of the first patch is O(N). We keep only the best M candidate origins from the list proposed by the first patch (in our implementation, $M = 50$). The second patch is now restricted to the neighborhoods of M locations. The third will be restricted to a much smaller number of neighborhoods. Thus, in the worst case scenario, our complexity is $O(N) + O(nM) \approx O(N)$. In contrast, the complexity of the inference process in [3, 8] is $O(nN)$, while the complexity of the "constellation model" [4] is *exponential* in the number of patches. The above proposed reduction in complexity is extremely important for enabling video inference with ensembles containing hundreds of patches.

Multi-scale search: To further speedup the elimination process, we choose the first searched patches from a coarse

(a) The database sequence contains a short clip of a single person walking and jogging:

(b) Selected frames from the query sequence: (Colored frames = input; BW frames = output; Red=Suspicious)



(c) More frames from the query sequence... (Colored frames = input; BW frames = output; Red=Suspicious)

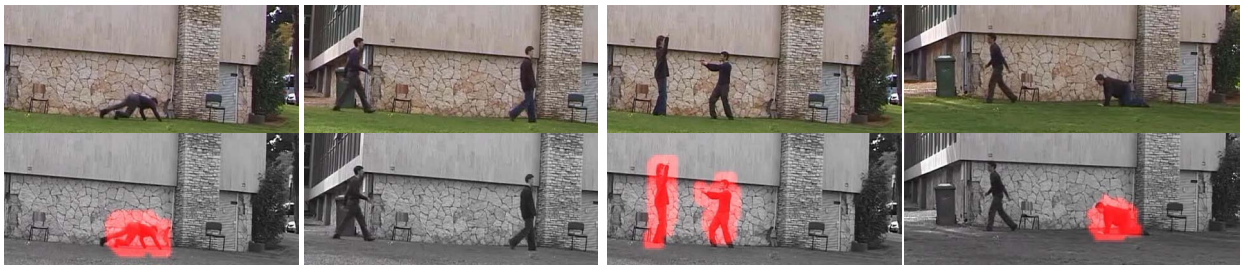


Figure 7. Detection of suspicious behaviors. *New valid behavior combinations are automatically inferred from the database (e.g., two men walking together, a different person running, etc.), even though they have never been seen before. behaviors which cannot be inferred from the database clips are highlighted in red as being "suspicious". For full videos see* `www.wisdom.weizmann.ac.il/~vision/Irregularities.html`

scale, for two reasons: (i) There is a much smaller number of coarse patches in the database than fine patches (thus decreasing the effective N in the first most intensive step), and (ii) coarse patches are more discriminative because they capture information from large regions. This eliminates candidate origins very quickly. Nevertheless, there are cases in which a valid ensemble cannot be explained in coarse scales (e.g., due to partial occlusion). In these cases (which are not very frequent), we repeat the elimination process without the coarsest scale, starting with a finer-scale patch as the first patch (but penalize the overall ensemble likelihood score). This is done in order to distinguish between these kind of ensembles and irregular (invalid) ensembles.

6 Applications

The approach presented in this paper gives rise to a variety of applications which involve detection of irregularities in images and in videos:

1. Detecting Unusual Image Configurations: Given a database of example images, we can detect unusual things in a new observed image (such as objects never seen before, new image patterns, etc.) An example is shown in Fig. 5. Images of three different poses are provided as a database (Fig. 5.a). Images of other poses are provided as queries (Fig. 5.b). New valid poses (e.g., a man sitting on the chair with both arms up, a man sitting on a chair with one arm up) are automatically inferred from the database, even though they have never been seen before. New pose parts which cannot be inferred from the three database images are highlighted in red as being "unfamiliar" (Fig. 5.c). Fig. 5.d visually indicates the database image which provided most evidence for each pixel in the query images (i.e., it tells which database image contains the largest most probable region of support for that pixel. Note, however, that these are *not* the regions of support themselves). Uniform patches (with negligible image gradients) are assumed valid by default and discarded from the inference process (for added speedup).

2. Spatial Saliency in a Single Image: Given a single image (i.e., no database), salient image regions can be detected, i.e., image regions which stand out as being different than the rest of the image. This is achieved by measuring the likelihood of each image region (the "query") relative to the remaining portions of the same image (the "database" used for inferring this particular region). This process is repeated for each image region. (This process can be performed effi-

A few sample frames from an input video *(top row)*, and the corresponding detected behavioral saliency *(bottom row)*:

Figure 8. Detecting salient behaviors in a video sequence (no database and no prior information). *Saliency is measured relative to all the other behaviors observed at the same time. In this example, all the people wave their arms, and one person behaves differently. For full videos see* `www.wisdom.weizmann.ac.il/~vision/Irregularities.html`

ciently by adaptively adding and removing the appropriate descriptors from the "database" when proceeding from the analysis of one image region to the next). Such an example is shown in Fig. 6. This approach can be applied to problems in automatic visual inspection (inspection of computer chips, goods, etc.)

3. Detecting Suspicious Behaviors: Given a small database of sequences showing a few examples of valid behaviors, we can detect suspicious behaviors in a new long video sequence. This is despite the fact that we have never seen all possible combinations of valid behaviors in the past, and have no prior knowledge of what kind of suspicious behaviors may occur in the scene. These are automatically composed and inferred from space-time patches in the database sequence. An example is shown in Fig. 7, which shows a few sample frames from a 2-minute-long video clip, along with detected suspicious behaviors. For full videos see `www.wisdom.weizmann.ac.il/~vision/Irregularities.html`. The result of our algorithm is a continuous likelihood map. In our video examples, a *single* threshold was selected for an entire video sequence query. More sophisticated thresholding methods (hysteresis, adaptive threshold, etc.) can be used.

An important property of our approach is that we can incrementally and adaptively update the database when new regular/valid examples are provided, simply by appending their raw descriptors and locations to the database. No "relearning" process is needed. This is essential in the context of detecting suspicious behaviors, should a detected suspicious behavior be identified as a false alarm. In such cases, the database can be updated by appending the new example, and the process can continue.

4. Spatio-Temporal Saliency in Video: Using our approach we can identify salient behaviors within a single video sequence, without any database or prior information. For example, one person is running amongst a cheering crowd. The behavior of this person is obviously salient. In this case, saliency is measured relative to all the other behaviors observed at the same time. The "validity" of each space-time video segment (the "query") is measured relative to all the other video segments within a small window in time (the "database" for this particular video segment). This process is repeated for each video segment. Such an example is shown in fig. 8. For full videos see `www.wisdom.weizmann.ac.il/~vision/Irregularities.html`.

Video saliency can also be measured relative to other temporal windows. E.g., when the saliency is measured relative to the entire video, behaviors which occur only once will stand out. Alternatively, when the saliency is measured relative to the past (all previous frames), new behaviors which have not previously occurred will be spotted. This gives rise to a variety of applications, including video synopsis.

References

[1] E. Bart and S. Ullman. Class-based matching of object parts. In *VideoRegister04*, page 173, 2004.

[2] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, pages 1033–1038, 1999.

[3] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR03*.

[5] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *IJCV*, 40(1):25–47, October 2000.

[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 1998.

[7] Y. Ivanov and A. Bobick. Recognition of multi-agent interaction in video surveillance. In *ICCV*, 1999.

[8] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV04 Workshop on Statistical Learning in CV*.

[9] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[10] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 2000.

[11] Y. Wexler, E. Shechtman, and M. Irani. Space-time video completion. In *CVPR04*, pages I: 120–127, 2004.

[12] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR04*, pages II: 819–826, 2004.