Top-Down Visual Attention with Complex Templates

Jan Tünnermann, Christian Born and Bärbel Mertsching
GET Lab, University of Paderborn, Pohlweg 47-49, 33098 Paderborn, Germany

Keywords: Visual Attention, Saliency, Top-Down Control, Visual Search, Object Detection.

Abstract: Visual attention can support autonomous robots in visual tasks by assigning resources to relevant portions of an image. In this biologically inspired concept, conspicuous elements of the image are typically determined with regard to different features such as color, intensity or orientation. The assessment of human visual attention suggests that these bottom-up processes are complemented – and in many cases overruled – by top-down influences that modulate the attentional focus with respect to the current task or a priori knowledge. In artificial attention, one branch of research investigates visual search for a given object within a scene by the use of top-down attention. Current models require extensive training for a specific target or are limited to very simple templates. Here we propose a multi-region template model that can direct the attentional focus with respect to complex target appearances without any training. The template can be adaptively adjusted to compensate gradual changes of the object's appearance. Furthermore, the model is integrated with the framework of region-based attention and can be combined with bottom-up saliency mechanisms. Our experimental results show that the proposed method outperforms an approach that uses single-region templates and performs equally well as state-of-the-art feature fusion approaches that require extensive training.

1 INTRODUCTION

Artificial attention models serve the purpose of guiding a visual agent's high-level resources towards relevant portions of a scene. Technical models are typically grounded on the Feature-Integration Theory (Treisman and Gelade, 1980) and implemented according to the scheme described by Koch and Ullman (1985) where saliency is determined for different features (as color or orientation) separately and subsequently fused to an overall saliency map.

Bottom-up models of visual attention direct the attentional focus towards salient parts of the scene that stand out from their neighborhood. Saliency can be determined in various ways, e.g. by applying filters on image pyramids (Itti et al., 1998; Belardinelli et al., 2009), finding conspicuities in frequency domain (Hou and Zhang, 2007; Jian Li and He, 2011) or in a region-based manner (Aziz and Mertsching, 2008a; Tünnermann and Mertsching, 2012). A recent and extensive review of different artificial attention models has been published by Borji and Itti (2012).

However, the fact that the task at hand has a substantial influence on scene analysis had already been demonstrated by Yarbus (1967), who had shown that subjects produce considerably different scan paths (fixations and saccades) for the same scene but with different tasks. Also, more recent assessments of the primate vision system yield evidence that top-down influences (information fed back from higher cognitive levels) have an impact on even early pre-attentional stages, see e.g. (Li et al., 2004) and (Hilkenmeier et al., 2009).

This paper contributes to the branch of research in artificial modeling that aims to integrate such top-down influences. Top-down mechanisms have been modeled in a broad variety, reaching from the influence of scene context (or "gist") (Torralba et al., 2006; Oliva and Torralba, 2006) over knowledge and familiarity (Aziz et al., 2011) to specific search tasks (Itti and Koch, 2001; Navalpakkam and Itti, 2006; Aziz and Mertsching, 2008b; Wischnewski et al., 2010; Kouchaki and Nasrabadi, 2012). Here, we focus on the visual search task, which is known to strongly rely on attention in biological systems (Wolfe and Horowitz, 2004). The artificial models pursuing this idea can be divided into two groups: *feature map fusion* approaches and *template-based* approaches. *Feature map fusion* approaches (Itti and Koch, 2001; Navalpakkam and Itti, 2006; Kouchaki and Nasrabadi, 2012) extend the model by Itti et al. (1998) which creates an image-pyramid and filters it

Tünnermann J., Born C. and Mertsching B..
Top-Down Visual Attention with Complex Templates.
DOI: 10.5220/0004302403700377
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 370-377
ISBN: 978-989-8565-47-1

for different low-level features, such as color, orientation and intensity and calculates center-surround differences. The resulting maps from different channels and different resolutions of the pyramid are fused to form a master map of saliency which is used to obtain the focus of attention. In the classic bottom-up formulation, the fusion is a simple superposition of the individual maps. Itti and Koch (2001) have investigated different fusion strategies, one of them containing information from supervised learning of a search target that increases the likelihood that the corresponding object will become most salient in a linear combination of the maps. Navalpakkam and Itti (2006) include knowledge about target and background in the determination of the fusion weights. Recently, Kouchaki and Nasrabadi (2012) proposed nonlinear fusion using a neural network that was trained with the target object. *Template-based* approaches work with a segmentation of the scene and an abstract template that is defined by a set of region features (such as color, size, symmetry, orientation, etc.) of the target (Aziz and Mertsching, 2008b; Wischnewski et al., 2010). Therefore, these approaches do not require training and can deal with abstract information about the target ("look for a large, red object") which is advantageous in many contexts where training in advance or supervision is not an option. However, an object may be more complex than what a single set of region features can capture; additionally, it can occur in different perceptual configurations due to distance, occlusion or perspective distortion. Here, we suggest a method to consider multi-region templates that can represent more complex objects and are tolerant to changes in the perceptual appearance. Our evaluation shows that this extended *template-based* model has a performance that compares well with *feature map fusion* models (that rely on extensive training), without requiring any training.

2 MULTI-REGION TEMPLATES

The proposed method is grounded in the concept of region-based attention, in which the scene is initially segmented into regions $\mathfrak{R} = \{R_1, R_2, \ldots, R_n\}$. The method of segmentation is not relevant here, and thus we simply assume in the following that a region R_i represents a connected area of pixels by the magnitudes ϕ_i^f of different features f that have been obtained during or after the segmentation. In our implementation, we use the features average color, symmetry, orientation and eccentricity (based on central 2D moments) and size in pixels ($f \in \{_{\text{color}}, _{\text{orientation}}, _{\text{symmetry}}, _{\text{eccentricity}}, _{\text{size}}\}$); for details regarding the creation of regions and features for attentional processes we refer to (Aziz and Mertsching, 2008a).

The idea of multi-region templates is straightforward. Instead of looping over the regions R_i of the scene segmentation and comparing their features to a single template region F_{td} (Aziz and Mertsching, 2008b), the scene segmentation is searched for a specific configuration $C^{F_{td}}$ of template regions that are expected at their relative positions with a certain position tolerance D. Note that we call the set of template features F_{td} also a region, as it can be obtained by segmenting an image of the target object. The proposed algorithm is based on the concept of fixing one (arbitrary) region of the template as the parent $\pi_j \in C^{F_{td}}$ of all other template regions $\tau_k \in C^{F_{td}}$. This parent is searched first in the scene segmentation just as described for single-region templates in (Aziz and Mertsching, 2008b) with $F_{td} = \pi_j$ (see also box **Top-Down Saliency Equations**). When the parent is found with a certain likelihood, i.e. a scene region R_i achieves a certain single-region top-down saliency (SRTD) γ_i according to (Aziz and Mertsching, 2008b), a new proto-object PO_p is created and the identified parent (R_i) is added to it. Then children are searched and when scene regions η_m are identified as children, the one (η_m^{max}) with the most SRTD saliency is added to the proto-object. When the parent was found with a size different from its size in the template, the expected sizes and distances for all children are adjusted. The overall top-down saliency Γ_p of the proto-object PO_p is calculated as the average of the η_m^{max}'s SRTD saliencies γ_m^{max} and the parent's saliency γ_i. The contributions of the children can be reduced by a penalty $\mu(d)$ for deviations d from the expected distance between parent and child and it is zero when the child is completely missing (not at a tolerable position with regard to D). This process is repeated with another template region being the parent until all template regions have been the parent once. This is required as there is always the chance that the parent region is missing due to occlusion or perspective distortion, or that a more fitting configuration can be found by adjusting the expected sizes with respect to another parent region. The focus of attention for the scene can be obtained as the PO_p with the highest saliency Γ_p. In our tests, we marked its position by drawing a bounding box that contains all the regions of the proto-object (see figure 2 (b)). The template configuration $C^{F_{td}}$ can be created from abstract knowledge or by segmenting an image of the target (see figure 1).

Algorithm 1 formally describes the overall procedure and generates a top-down saliency map by as-

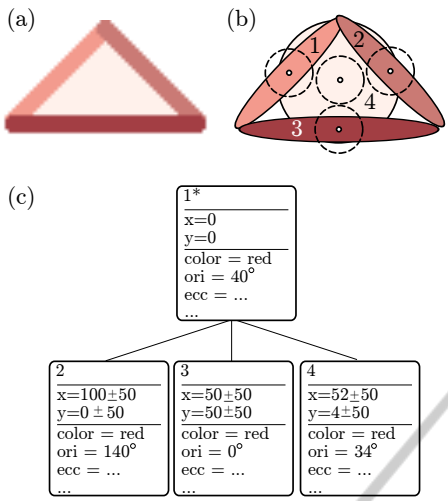

Figure 1: (a) A pixel-based input can be segmented to obtain a template. (b) Illustration of information stored for a multi-region template. The position tolerance D with respect to the region center is indicated by the dashed circles. (c) During the execution of the algorithm a region is selected as parent (here 1, marked by the *) and the other regions become children that are expected at relative positions according to the template layout.

cribing the highest saliency to each individual region that some parent-children configuration containing it yielded for it. Variable π_j loops over the template regions to select every region as parent once. The threshold $threshold_\pi$ is the minimum single-region top-down saliency required to continue. If all comparisons with the R_i are below it, this parent region is considered to be not found in the scene. The child regions in the template are denoted as τ_k and the potential child regions in the scene as η_m. Note that the function μ(d) is the penalty for variance in the expected region position which should be low close to the expected position and high further away. We used μ(d) as defined in equation 1.

$$\mu(d) = \begin{cases} (1-(\frac{d}{D})^e) & \text{if } d \leq D \\ 0 & \text{else} \end{cases} \quad (1)$$

where d is the distance between found and expected position and D the parameter that defines the maximum allowed distance. The mentioned size adjustment with respect to the current parent region π_j in the SRTD saliency computation is omitted in algorithm 1 for clarity.

The concept proposed here is independent of specific features; however, in our implementation we use the features color, symmetry, orientation, eccentricity and size of the individual regions as described by Aziz and Mertsching (2008b). The similarity measure is the single-region top-down saliency determination from (Aziz and Mertsching, 2008b), which is summarized in the box **Top-Down Saliency Equations** for the interested reader.

Top-Down Saliency Equations

(Aziz and Mertsching, 2008b)

In the following, the target is defined as F_{td} which is a collection of target features F_{td}^f. The features of scene regions R_i are denoted as ϕ_i^f. In our current implementation, the features $f \in \{\text{color}, \text{orientation}, \text{symmetry}, \text{eccentricity}, \text{size}\}$ are considered. The single-region top-down saliency for a feature (except color which is a compound feature; see below) is computed by obtaining the normalized ratio Θ^f of the feature magnitudes.

$$\Theta^f = \begin{cases} \phi_i^f / F_{td}^f & \text{for } \phi_i^f < F_{td}^f \\ F_{td}^f / \phi_i^f & \text{otherwise} \end{cases}$$

The ratio is then $0 \geq \Theta^f \geq 1$, regardless of the direction of the difference or the value range of the feature.
The single-region top-down saliency $0 \geq \gamma_i^f \geq 1$ for region R_i is then obtained as

$$\gamma_i^f = \begin{cases} \Theta^f & \text{for } \Theta^f > D^\Theta \\ 0 & \text{otherwise} \end{cases}$$

where D^Θ is a threshold below which regions are considered not similar at all (we set $D^\Theta = 0.91$ in our implementation).

The top-down color saliency (γ^f with f = color) is computed with the following formula, as color is a compound feature which consist of hue, saturation and intensity (in our HSI color space implementation). The thresholds D^h, D^s and D^I are the tolerated differences of each component and Δ_i^h, Δ_i^s and Δ_i^I the actual differences between the color components of region R_i and F_{td}^{color}. The color saliency γ_i^{color} is given by

$$\gamma_i^{color} = \begin{cases} \frac{a(D^h-\Delta_i^h)}{D^h} + \frac{b(D^s-\Delta_i^s)}{D^s} + \frac{c(D^I-\Delta_i^I)}{D^I} & \text{for } \Delta_i^h < D^h \wedge chrom \\ \frac{(a+b+c)(D^I-\Delta_i^I)}{D^I} & \text{for } \Delta_i^I < D^I \wedge achrom \\ 0 & \text{otherwise} \end{cases}$$

where *crhom* is true if (and only if) the colors of the region R_i and the template F_{td}^{color} are both chromatic. If (and only if) both are achromatic, *acrhom* is true. The weights a, b and c are used to adjust the importance of the color channels and are set to $a = 0.39$, $b = 0.22$ and $c = 0.39$ in our implementation).
The overall top-down saliency γi is computed as the (possibly) weighted average of the γ_i^f.

The result of the procedure is a region-based top-down saliency map (see figure 2 (d)) and a number of proto-objects (PO_p in algorithm 1) that contain the found configurations of regions and the group's overall saliency Γ_p. In figure 2 (b) the target has been highlighted with the bounding box of the proto-object with the highest top-down saliency (note that the saliency map in figure 2 (d) also contains contributions from other proto-objects).

Algorithm 1: Multi-Region Top-Down Saliency.

for each $\pi_j \in C^{F_{id}}$ **do** // Every template region becomes the parent once.
  for each $R_i \in \Re$ **do** // Every scene region is checked against the parent.
    if γ_i with $F_{id} = \pi_j > threshold_\pi$ **then** // SRTD saliency for region R_i with template π_j exceeds threshold.
      Append(ProtoObjects, PO_p) // Add a new proto-object to the list.
      Append(PO_p, (R_i)) // Add the identified parent to the proto-object.
      $\Gamma_p = \Gamma_p + \gamma_i$ // Begin summing up proto-object saliency.
      for each $\tau_k \neq \pi j \in C^{F_{id}}$ **do** //Search for the child regions.
        for each $\eta_m \in \Re$ **do**
          $\Delta^{expected_k}_{found_m}$ = Distance(Position(η_m),Position(R_i)+RelativePosition(τ_k) // Store the deviation from the expected position.
          if $\Delta^{expected_k}_{found_m} < D$ **then** // If not too far away, consider as child.
            if (γ_m with $(F_{id} = \tau_k)) \cdot \mu(\Delta^{expected_k}_{found_m}) > \gamma^{max}_m$ **then** // If SRTD saliency for η_m with the template τ_k higher that the current max. ...
              $\gamma^{max}_m = \gamma_m \cdot \mu(\Delta^{expected_k}_{found_m})$ // ... store maximum saliency.
              $\eta^{max}_m = \eta_m$ // ... store maximum salient region.
            end if
          end if
        end for
        Append(PO_p,η^{max}_m) // Add best fitting candidate to proto-object.
        $\Gamma_p = (\Gamma_p + \gamma^{max}_m)/size(PO_p)$ // Add contribution of child to proto-object saliency and normalize.
      end for
    end if
  end for
end for

Distance(x,y): Returns distance between x and y; **Position(Region)**: Returns position of Region;
RelativePosition(Region): Returns position of Region relative to the parent in the template; **Append(List,Object)**: Appends Object to List.

3 ADAPTIVE TEMPLATES

The method described above is capable of compensating small deviations in the object appearance in the scene with regard to the template. The configuration can change drastically due to motion of the system or scene objects. This can be compensated to a certain degree when the current template is continuously updated. Usually, the appearance changes only slightly from frame to frame, so the template from the previous frame will still work for the current; the currently focused object (the associated proto-object) is a good template for the next frame. Furthermore, the method described in section 2 is not limited to one multi-region template. Different templates can be applied successively and the best fitting will select the focus of attention. In combination with the aforementioned adaptive updating, this means that object representations can be learned by accumulating multiple templates over time and an object that had disappeared from the scene can be rediscovered when it reappears similar to any of the stored configurations. This resembles the idea of learning objects from multiple views (Blanz et al., 1996). Interestingly, the behavior that emerges from these mechanisms is also similar to modern object tracking solutions as TLD (Kalal et al., 2009), which include adaptive updating and learning of the templates. We included a demonstration of this behavior in section 4 in addition to the evaluation in the context of visual search.

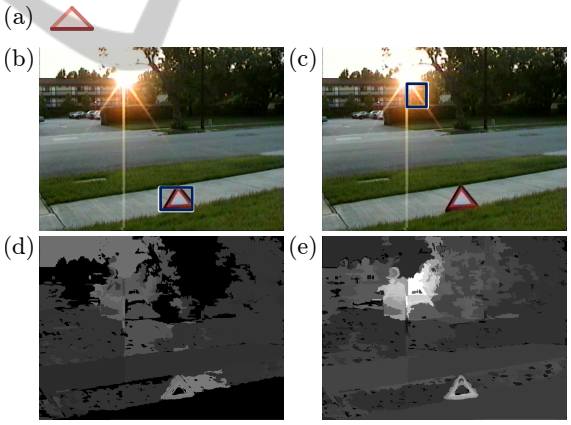

Figure 2: (a) Template used for the multi-region top-down model that finds the target in a single run, marked in (b). (c) In a bottom-up saliency calculation, the focus of attention is directed towards the salient sun. (d) A top-down saliency map obtained as described in algorithm 1. Note that saliency was assigned to regions multiple times by different parent-child configurations, whereas the final focus (shown in (b)) is obtained by finding the highest proto-object saliency (saliency summed for all regions of one configuration). The bottom-up saliency map in (e) shows peaks at the triangle but higher ones at the sun. The image is part of the *STIMTriangle* database from Itti and Koch (2001).

4 EVALUATION

We evaluated the top-down attention capabilities of our model in a visual search task using the popular *STIMTriangle* database from Itti and Koch (2001), where an emergency triangle is to be found in the scene (one example image from the scene is shown in figure 2). For our model we provided an abstract multi-region template (shown in figure 2 (a)) which basically consists of the three red bars and the bright center as regions of the template (as illustrated in figure 1). For the single-region top-down mechanism that is a sub-process of the proposed method and the region-based bottom-up model we set the feature weights (see Aziz and Mertsching, 2008b) to high importance for color, medium importance for size and very low importance for any other features. The parameter $threshold_{\pi}$, which mainly influences the runtime, was set to 0.6 and the position tolerance D was set to 50 pixels. The *STIMTriangle* database contains a training set and a test set (32 images each). As the region-based algorithms require no training, we evaluated them also on the training set, additionally to the test set.

Figure 3 shows a comparison of the single-region top-down model by Aziz and Mertsching (2008b) and the proposed method. The latter detects all targets in the training set in the first trial (hence, the zero mean) and requires 1.68 fixations in the average on the test set. The huge difference in success compared to the single-region method ② – it has a mean of 11.4 for the test set and 4.69 for the training set – can be explained by the fact that the single-region method relies on capturing the target's appearances with a single set of features. This means, only if the target is *perceptually* similar to the template (same projected size, same average color, etc.), then it has a high detection probability. The difference in the results of the single-region method applied to the test set and applied to the training set is a another indicator for this. The training set contains more clear appearances (always frontal, no parts occluded), so it is more likely that they perceptually resemble the template appearance. A comparison of bottom-up models, in particular *feature map fusion* models which are capable of learning various different appearances, and the proposed multi-region top-down model is depicted in figure 4.

Figure 4 ① shows results of the region-based bottom-up attention model by Aziz and Mertsching (2008a) in this task. The proposed method is marked with ② in the figure; a combination of the bottom-up computation and the proposed method is shown as ③. The bottom-up contribution was weighted five times as high as the top-down saliency; this was empirically determined and at lower or higher weights no benefit was achieved (in the future, such a combination will be automatically optimized by fusion strategies similar to those reported in (Itti and Koch, 2001)). The results marked by ④ have been reported in literature (Kouchaki and Nasrabadi, 2012) for the bottom-up model by Itti et al. (1998) for this database. Finally, the map fusion and top-down strategies by Itti and Koch (2001) and Kouchaki and Nasrabadi (2012) are labeled with ⑤.

The bottom-up models were included as the emergency-triangle often is a pop-out and they can be considered a baseline without top-down information. As the region-based models do not require any training, we run them on the full set of images from the *STIMTriangle* database. The measures we used are compatible with those used by Kouchaki and Nasrabadi (2012), but they have been normalized by the image set size (when required) to allow comparison of results from the full set and the test set. The portions of the bars in the figure that are dashed refer to the measure considering only the test set, while full bars represent the whole set. Figure 4 (a) depicts the portion of *first hit detections (FHD)*; this is the portion of images in which the target was found in a single run of the model. The bottom-up models perform 30 % to 60 % and all variants involving top-down influences at about 90 %. When a target was not hit immediately, inhibition of return (IOR) was applied, so in a consecutive run the model selects the next salient element. The *unsuccessful trials* measure (UST in figure 4 (b)) is the portion of images where the target was not found within five trials. The mean number of trials before target detection and its standard deviation is reported in figure 4 (c). The relatively high means and standard deviations, along with the quite successful FHDs, reflect that most targets were found quickly except for some images that required a large number of trials. This effect seems to be stronger for region-based methods in general and might result from segmentation problems. However, in general the performance of the proposed model compares well with *feature map fusion* models. The mean number of trials until detection is reduced when bottom-up and top-down is combined.

Figure 5 shows a demonstration of using adaptive templates as outlined in section 3. The initial template is shown in figure 5 (a) and frames of a dynamic scene in which the observer is moving is shown in figure 5 (b). Note that only every eleventh frame is shown, so the frame-to-frame difference in the evaluated scene was smaller than in the figure. When the adaptive updating of the template is not applied (figure 5 (c)) the saliency is gradually going down and

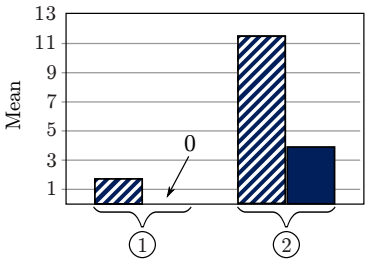

Figure 3: Mean number of fixations until the target was found. The dashed bars correspond to the evaluation of the test set and solid bars to evaluation of the training set. ① The proposed model had no wrong detections at all before the target was found on the training set, which contains images with a frontal view and the target is always fully visible. ② The results for the single-region template top-down model by Aziz and Mertsching (2008b). This model shows a large difference between the test set and training set due to the fact that it cannot compensate appearance variations that occur more pronounced in the test set. Note that in some images the single-region template model was not able to find the target within 30 trials (this happened seven times for the test set and two times for the training set). In this situation, the number of trials required was fixed at 30, so the performance of the single-region template model is overestimated and the advantage of the multi-region template model might be even larger.

when the difference of the original template and current appearance is too large, the object is no longer the most salient entity (this can be seen in the last frame of the sequence). When the adaptive updating is activated, the saliency is constantly kept at high levels, even when the resemblance of the original template and the current appearance is extremely low (as in the last frame). The test scene continued with the target completely leaving the scene (not shown), but it is rediscovered when it enters the scene again. This test was performed on data from a mobile robot simulator (Kotthäuser and Mertsching, 2010) where a very stable segmentation can be obtained. In a test on a highly dynamic and noisy real world scene, the target was fixated in only about 30 % of the frames. However, the proposed system is not intended as a tracker, but as these results show, it can be a valuable support for such systems.

5 CONCLUSIONS

The proposed top-down attention model has been evaluated in a visual search task. The comparison with other attention models that were tested in this task shows that our template-based approach performs similarly successful as models that require extensive training. The multi-region templates pre-

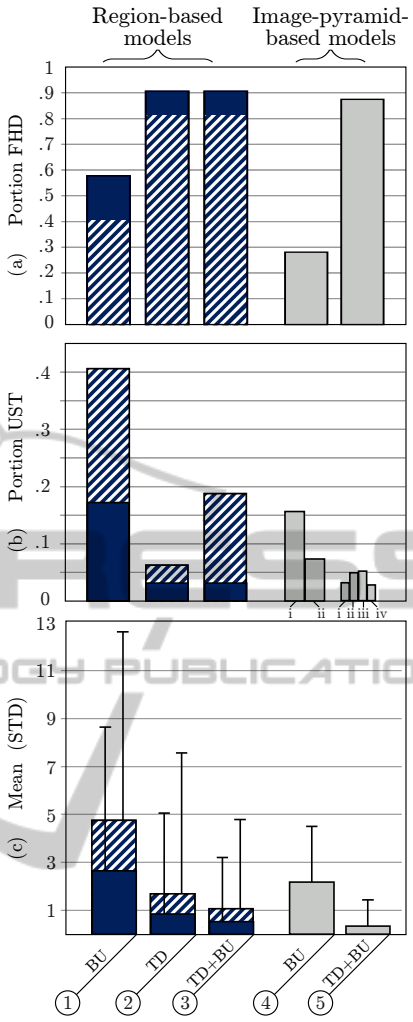

Figure 4: Different attention models evaluated on the *STIM-Triangle* database from Itti and Koch (2001). The region-based bottom-up model by Aziz and Mertsching (2008a) is marked as ①. The proposed multiregion-region top-down model is marked as ② and a combination of ① and ② as ③. When not splitted, bars of ④ refer to results of the bottom-up model by Itti et al. (1998) reported by Kouchaki and Nasrabadi (2012) and ⑤ to the nonlinear fusion suggested in (Kouchaki and Nasrabadi, 2012). (a) shows the portion of "First Hit Detections" in which the targets was selected in the first run and (b) the portion of "Unsuccessful Trials", where the target was not found in five trials. ④ i is the value for the model from Itti and Koch (2001) reported in (Kouchaki and Nasrabadi, 2012) while ④ *ii* is the value reported in (Itti and Koch, 2001) for the "Naive" feature fusion strategy. ⑤ *i* is the nonlinear fusion strategy from (Kouchaki and Nasrabadi, 2012), while ⑤ *ii,iii,iv* refer to $\mathcal{N}(.)$, "Iterative" and "Trained" from (Itti and Koch, 2001), respectively. (c) gives the mean number of trials required until target detection and the standard deviation. Dashed bars indicate that only the test set of the database was considered. As the region-based models do not require training, they were also evaluated on the full set (test + training), which is shown as solid parts of the bars.

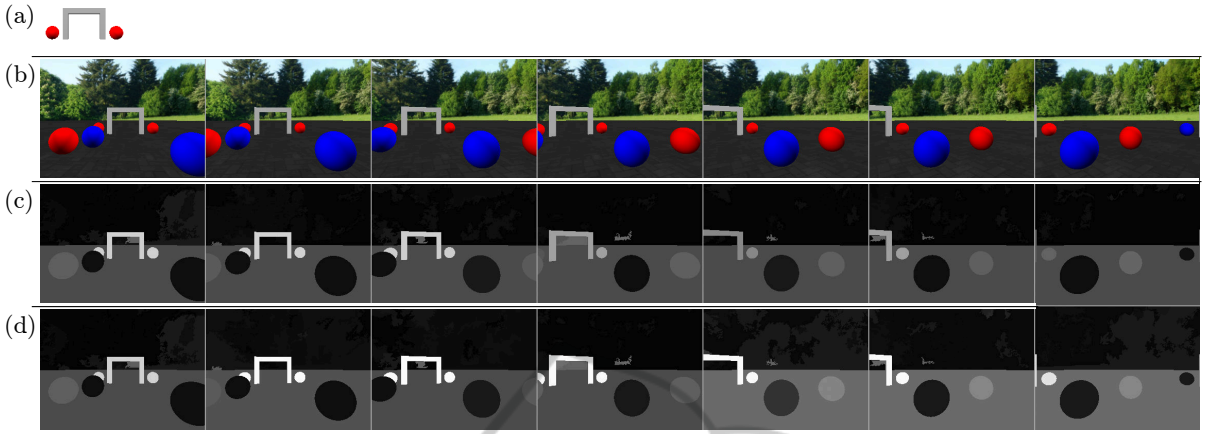

Figure 5: (a) The initial template. (b) Frames from a sequence where the perceptual image of the object is distorted and occluded by perspective changes due to motion of the system. (c) Results of the initial template being applied to each frame. (d) When the template is continuously updated with the previously focused configuration, a high saliency can be retained even when only a small part of the target is visible.

serve the advantage of region-based pre-attentional calculations. Configurations and features of the template can be altered at runtime, based on abstract information (e.g. "look for a triangle with blue border") with no need to re-learn the appearance on low levels. Furthermore, multi-region templates proved to be robust and allow heuristics to compensate perspective variations, occlusions and other alterations in the perceptual appearance of the target. We demonstrated that multi-region templates and proto-objects can be used to establish an adaptive behavior that keeps the attentional focus on the target, when it gradually changes its appearance, compensating even substantial changes compared to the original template.

The computation time of the proposed multi-region top-down model depends on the complexity of the scene and the template. For the set of images and the template from the evaluation (see section 4) the top-down saliency computation itself consumed 32 ms in average on a 3.4 GHz computer, so if a number of multi-region templates are applied to a segmentation it can be done at about 30 cycles per second. The full processing including segmentation, however, took about 330 ms at the rather high resolution of 640×480. The full processing can be sped up to a reasonable frame rate by lowering the input image resolution or adjusting segmentation parameters to yield a coarser resolution at the cost of lower detection rates. With regard to this speed-accuracy trade-off, it should be considered if a subsequent task requires a rather accurate or a rather fast suggestion from the attentional focus. For example, if a full object recognition follows, it might be sufficient to quickly generate rough candidates. If the goal is to perform some long-term behavior towards the target (such as driving towards it for a detailed analysis) it might be better to invest more effort into the attentional selection.

However, the pre-attentional processing must be as fast as possible, so recently, a GPU-based segmentation method has been developed that is well suited for region-based attention models (Backer et al., 2012). In future work on region-based top-down attention, this will replace the current sequential segmentation. Further aspects of the proposed method are subject to parallelization also. The template configuration is searched in the scene multiple times with every template regions being the parent in one search; however, these searches are completely independent from one another and can be executed at the same time. To further improve the detection rate, different relations between regions can be considered. In the current implementation we only care for the distance from parent to child regions, but other relations, such as the direction (e.g. child is left of the parent) or the appearance (child is brighter than parent) could yield useful contributions.

REFERENCES

Aziz, Z., Knopf, M., and Mertsching, B. (2011). Knowledge-Driven Saliency: Attention to the Unseen. In *ACIVS 6915*, LNCS, pages 34 – 45.

Aziz, Z. and Mertsching, B. (2008a). Fast and Robust Generation of Feature Maps for Region-Based Visual Attention. *IEEE Transactions on Image Processing*, 17, May 2008(5):633 – 644.

Aziz, Z. and Mertsching, B. (2008b). Visual Search in Static

and Dynamic Scenes Using Fine-Grain Top-Down Visual Attention. In *ICVS 5008*, LNCS, pages 3 – 12, Santorini, Greece.

Backer, M., Tünnermann, J., and Mertsching, B. (2012). Parallel k-Means Image Segmentation Using Sort, Scan & Connected Components on a GPU. In *FTMC-III*, LNCS.

Belardinelli, A., Pirri, F., and Carbone, A. (2009). Attention in cognitive systems. chapter Motion Saliency Maps from Spatiotemporal Filtering, pages 112–123. Springer, Berlin - Heidelberg.

Blanz, V., Schölkopf, B., Bülthoff, H., Burges, C., Vapnik, V., and Vetter, T. (1996). Comparison of View-based Object Recognition Algorithms Using Realistic 3D Models. In von der Malsburg, C., von Seelen, W., Vorbrüggen, J., and Sendhoff, B., editors, *Artificial Neural Networks*, volume 1112 of *LNCS*, pages 251–256. Springer, Berlin - Heidelberg.

Borji, A. and Itti, L. (2012). State-of-the-Art in Visual Attention Modeling. *Accepted for: IEEE TPAMI*.

Hilkenmeier, F., Tünnermann, J., and Scharlau, I. (2009). Early Top-Down Influences in Control of Attention: Evidence from the Attentional Blink. In *KI 2009: Advances in Artificial Intelligence. Proceeding of the 32nd Annual Conference on Artificial Intelligence.*

Hou, X. and Zhang, L. (2007). Saliency Detection: A Spectral Residual Approach. In *IEEE CVPR*, pages 1–8.

Itti, L. and Koch, C. (2001). Feature Combination Strategies for Saliency-Based Visual Attention Systems. *Journal of Electronic Imaging*, 10(1):161–169.

Itti, L., Koch, C., and Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE TPAMI*, 20(11):1254–1259.

Jian Li, Martin Levine, X. A. and He, H. (2011). Saliency Detection Based on Frequency and Spatial Domain Analyses. In *BMVC*, pages 86.1–86.11. BMVA Press.

Kalal, Z., Matas, J., and Mikolajczyk, K. (2009). Online Learning of Robust Object Detectors During Unstable Tracking. *On-line Learning for Computer Vision Workshop*.

Koch, C. and Ullman, S. (1985). Shifts in Selective Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, 4:219–227.

Kotthäuser, T. and Mertsching, B. (2010). Validating Vision and Robotic Algorithms for Dynamic Real World Environments. In Ando, N., Balakirsky, S., Hemker, T., Reggiani, M., and Stryk, O., editors, *Simulation, Modeling, and Programming for Autonomous Robots*, volume 6472 of *LNCS*, pages 97–108. Springer, Berlin - Heidelberg.

Kouchaki, Z. and Nasrabadi, A. M. (2012). A Nonlinear Feature Fusion by Variadic Neural Network in Saliency-based Visual Attention. *VISAPP*, pages 457–461.

Li, W., Piëch, V., and Gilbert, C. D. (2004). Perceptual Learning and Top-Down Influences in Primary Visual Cortex. *Nature Neuroscience*, 7(6):651–657.

Navalpakkam, V. and Itti, L. (2006). An Integrated Model of Top-Down and Bottom-Up Attention for Optimal Object Detection. In *IEEE CVPR*, pages 2049–2056, New York, NY.

Oliva, A. and Torralba, A. (2006). Building the Gist of a Scene: The Role of Global Image Features in Recognition. In *Progress in Brain Research*, page 2006.

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual Guidance of Eye Movements and Attention in Real-world Scenes: The Role of Global Features in Object Search. *Psychological Review*, 113(4):766–786.

Treisman, A. M. and Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive psychology*, 12(1):97–136.

Tünnermann, J. and Mertsching, B. (2012). Continuous Region-Based Processing of Spatiotemporal Saliency. In *VISAPP*, pages 230 – 239.

Wischnewski, M., Belardinelli, A., Schneider, W. X., and Steil, J. J. (2010). Where to Look Next? Combining Static and Dynamic Proto-objects in a TVA-based Model of Visual Attention. *Cognitive Computation*, pages 326–343.

Wolfe, J. M. and Horowitz, T. S. (2004). What Attributes Guide the Deployment of Visual Attention and How Do They Do It? *Nature Reviews Neuroscience*, 5(6):495–501.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum., New York, NY.