TEC2011-25995 EventVideo (2012-2014)

Strategies for Object Segmentation, Detection and Tracking in Complex Environments for Event Detection in Video Surveillance and Monitoring

D4.2 VISUAL ATTENTION-DRIVEN TRACKING

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

AUTHOR LIST

Miguel Ángel García miguelangel.garcia@uam.es

CHANGE LOG

Version	Data	Editor	Description
0.1	26-02-2014	Miguel Ángel García	Initial version
1.0	16-06-2014	José M. Martinez	First version

CONTENTS

1. Introduction

This document summarizes a new visual attention model based on a joint perceptual space of both color and brightness, and shows that this model significantly improves the task of video tracking by finding more discriminant visual features, especially when dealing with objects that are very similar visually. That joint color and brightness space is based on a biologically-inspired theoretical perceptual model originally proposed by Izmailov and Sokolov in the scope of psychophysics.

The present document also summarizes a computational model that allows the direct application of Izmailov and Sokolov's theoretical model to digital images, since the original model can only be applied to perceptual data directly drawn from psychophysical experiments. Experimental results with real video sequences show that the proposed visual attention model yields significantly more accurate results in the application scope of video tracking than well-known visual attention models that process color and brightness separately.

The proposed models have been developed and implemented by the Video Processing and Understanding Lab in the Escuela Politécnica Superior of the Universidad Autónoma de Madrid.

1.1. Document structure

This document contains the following chapters:

- Chapter 1: Introduction.
- Chapter 2: Summarizes the Izmailov and Sokolov's perceptual model.
- Chapter 3: Describes a new computational model that allows the application of the original Izmailov and Sokolov's theoretical model to the determination of color differences in digital images.
- Chapter 4: Describes a new visual attention model based on the aforementioned computational adaptation of Izmailov and Sokolov's perceptual model.
- Chapter 5: Proposes a simple video tracking algorithm based on the previously proposed visual attention model.
- Chapter 6: Shows experimental tracking results with the proposed visual attention model.

- Chapter 7: Gives some conclusions and future work.

2. Izmailov and Sokolov's perceptual model

The theoretical perceptual model proposed by Izmailov and Sokolov in [1] yields a metric color space in which every point represents a specific color and Euclidean distances between points are proportional to perceived color differences. This model was derived by analyzing color differences through psychophysical experiments with human subjects and multidimensional scaling analysis techniques.

In particular, a 3-D semi-spherical space with axes X_1, X_2 and X_3 was defined such that the perceptual *chromatic* difference ΔC_{ij} between two equibright colors $\left({}^{i}X_1, {}^{i}X_2, {}^{i}X_3\right)^T$ and $\left({}^{j}X_1, {}^{j}X_2, {}^{j}X_3\right)^T$ in this space can be estimated by means of the interpoint Euclidean distance, $\left(\Delta C_{ij}\right)^2 = \left(\Delta X_1\right)^2 + \left(\Delta X_2\right)^2 + \left(\Delta X_3\right)^2$, where $\Delta X_\chi = {}^{i}X_\chi - {}^{j}X_\chi$. Similarly, a 2-D space with axes Y_1 and Y_2 was defined such that the perceptual *achromatic* difference ΔW_{ij} between two luminance levels $\left({}^{i}Y_1, {}^{i}Y_2\right)^T$ and $\left({}^{j}Y_1, {}^{j}Y_2\right)^T$ can be estimated through the Euclidean distance between both points, $\left(\Delta W_{ij}\right)^2 = \left(\Delta Y_1\right)^2 + \left(\Delta Y_2\right)^2$, where $\Delta Y_\chi = {}^{i}Y_\chi - {}^{j}Y_\chi$. Finally, a 4-D hyper-spherical color/luminance space with coordinates Z_1, Z_2, Z_3 and Z_4 was defined such that the perceptual difference ΔS_{ij} between two color/luminance points $\left({}^{i}Z_1, {}^{i}Z_2, {}^{i}Z_3, {}^{i}Z_4\right)^T$ and $\left({}^{j}Z_1, {}^{j}Z_2, {}^{j}Z_3, {}^{j}Z_4\right)^T$ can be estimated through the Euclidean distance between both points, $\left(\Delta S_{ij}\right)^2 = \left(\Delta Z_1\right)^2 + \left(\Delta Z_2\right)^2 + \left(\Delta Z_3\right)^2 + \left(\Delta Z_4\right)^2$, where $\Delta Z_\chi = {}^{i}Z_\chi - {}^{j}Z_\chi$.

By analyzing the projections of all 4-D points into every dimension, the authors established the following relationship between these four dimensions and the ones obtained before: $Z_1 = X_1$, $Z_2 = X_2$, $Z_3 = Y_2 X_3$ and $Z_4 = Y_1 X_3$. Taking these relationships into account, the perceptual difference ΔS_{ij} between two color/luminance points is finally rewritten as: $\left(\Delta S_{ij}\right)^2 = \left(\Delta C_{ij}\right)^2 + {}^{i}X_3\,{}^{j}X_3\left(\Delta W_{ij}\right)^2$, where ΔC_{ij} is the chromatic difference, ΔW_{ij} the achromatic difference, and X_3 the third component of the chromatic space, which is directly related to the color's lightness as will be shown in the next section.

The second achromatic term in the perceptual difference ΔS_{ij} introduces information about brightness differences between a region (the stimulus) and its neighborhood (the background). This is a novelty with respect to previous color difference models, which only process the color information of the compared points themselves, obviating their surroundings. This makes the

Izmailov and Sokolov model particularly attractive for visual attention, where the significance of points does not merely rely on their individual appearance, but also on the overall appearance of the regions in which they lie.

3. Computational adaptation of Izmailov and Sokolov's model

This section proposes a computational model that maps the RGB color space to the five variables that characterize the Izmailov and Sokolov perceptual model **Error! Reference source not found.** summarized in the previous section. This allows the computation of perceptual color differences directly from digital images. The following subsections describe the mapping to both the chromatic (X_1, X_2, X_3) and achromatic subspaces (Y_1, Y_2), respectively.

3.1. Computational mapping to chromatic subspace

The distribution of points in the chromatic subspace suggests the strong correlation between the first two dimensions, X_1 and X_2, and two so-called color single-opponent channels, RG and BY, corresponding to two neural pathways found in the retina and lateral geniculate nucleus (LGN) of primates. In particular, RG indicates how red (positive value) or green (negative value) a color is, whereas BY is equivalent for blue (positive value) and yellow (negative value). In turn, the third dimension X_3 is strongly correlated to the color's intensity I, which also corresponds to a third neural pathway found in the retina and LGN of primates.

Based on the existence of the aforementioned three neural pathways, a 3-D color space (RG, BY, I) can be defined according to the following mapping from the RGB color space proposed in **Error! Reference source not found.**: $RG = |r| - |g|$, $BY = |b| - |y|$, with *r*, *g*, *b*, and *y* defined as: $r = R - (G + B)/2$, $g = G - (R + B)/2$, $b = B - (R + G)/2$, $y = (R + G)/2 - |R - G|/2 - B$.

The similarity between both color spaces can be numerically assessed by determining the 3D affine transformation between the points in the (RG, BY, I) space and the ones in the (X_1, X_2, X_3) space through the 9-parameter Helmert transformation, which yields an approximation mean square error of 0.023 attributable to the fact that the color points in the original (X_1, X_2, X_3) space are the result of subjective differences perceived by human observers, whereas points in the (RG, BY, I) space are defined from exact values obtained from digital color images. Therefore, $X_1 \equiv RG$, $X_2 \equiv BY$, $X_3 \equiv I$.

3.2. Computational mapping to achromatic subspace

The achromatic subspace derived by Izmailov and Sokolov is a 2-D Euclidean space in which every combination of luminance levels for both the center of the image (stimulus) and the background is associated with a vector (Y_1, Y_2). In order to define a mapping between the RGB color space and that achromatic subspace, it is first necessary to map every normalized gray level from 0 to 1 to a luminance level in *cd/m*^{2}.

The DICOM Grayscale Standard Display Function (GSDF) defined in [3] describes the logarithmic relationship between the luminance level and the *Just-Noticeable Difference* (JND) index. One JND is the minimum variation of luminance that can be perceived by a human observer. Accordingly, the JND index, J_j, is defined in [3] as the input value to the GSDF such that the increment in one unit of J_j results in a luminance difference of one JND. Thus, the GSDF provides a straightforward mapping based on perceptual basis between luminance levels and the uniform space defined by the JND index.

The JND indices corresponding to the luminance levels utilized by Izmailov and Sokolov in their experiments can be calculated from the formulation of the inverse GSDF defined in [3]. The uniform interval defined between the minimum JND index, $J_{\min} = 22.73$, which corresponds to the minimum considered luminance of 0.2 *cd/m*^{2}, and the maximum JND index, $J_{\max} = 572.13$, associated with the maximum considered luminance of 200 *cd/m*^{2}, is normalized between zero and one: $\bar{J}_j = \left(J_j - J_{\min}\right) / \left(J_{\max} - J_{\min}\right)$, where $\bar{J}_j$ is the normalized JND index corresponding to J_j.

As described above, every vector (Y_1, Y_2) is a function of two luminance levels: one for the stimulus and another for the background. Let α and β be the normalized JND indices associated with the stimulus and background, respectively. The values of Y_1 and Y_2 obtained by Izmailov and Sokolov for all the combinations of α and β they considered show that the evolution of Y_1 for a same β is in accordance with a sum of two Gaussians. In turn, fixing β, the values of Y_2 follow a sigmoid-like function of α. In particular, Y_1 and Y_2 have been formulated as:

$$Y_1(\alpha,\beta) = \frac{s_0(\beta)\, e^{-\frac{(\alpha-\mu_0(\beta))^2}{2\sigma_0^2(\beta)}}}{\sigma_0(\beta)\sqrt{2\pi}} + \frac{s_1(\beta)\, e^{-\frac{(\alpha-\mu_1(\beta))^2}{2\sigma_1^2(\beta)}}}{\sigma_1(\beta)\sqrt{2\pi}} + d(\beta) \qquad Y_2(\alpha,\beta) = \frac{\gamma\left(\alpha-\eta(\beta)\right)}{\delta + \left|\alpha-\eta(\beta)\right|},$$

with the different coefficients having been estimated through non-linear fitting:

$$s_0(\beta) = 0.1409\beta^2 - 0.0964\beta + 0.0255 \qquad s_1(\beta) = 0.3109\beta^2 - 0.2397\beta + 0.0659$$

$$\mu_0(\beta) = -0.2295\beta^2 + 1.0105\beta - 0.1581 \qquad \mu_1(\beta) = 2.2817\beta^2 - 0.3646\beta + 0.3444$$

$$\sigma_0(\beta) = 0.2869\beta^2 - 0.1268\beta + 0.0890 \qquad \sigma_1(\beta) = 1.3638\beta^2 - 0.8679\beta + 0.2864$$

$$\delta(\beta) = -0.9670\beta^2 + 0.8441\beta + 0.2625 \qquad \eta(\beta) = 0.5495\beta + 0.1001$$

After having formulated the achromatic components Y_1 and Y_2 as bivariate functions of α and β, with the latter being normalized JND indices, it is essential to define the mapping between RGB color vectors and those normalized JND indices.

Since CMOS and CCD sensors typically mounted on video cameras have a linear response to light due to the almost linear photosensitivity of silicon, most cameras apply a non-linear correction (gamma compression) in order to emulate the logarithmic response of the human vision system and, at the same time, to increase their dynamic range. Gamma compression typically applies a $\frac{1}{2.2}$-power function to the three linear RGB channels generated by the camera's sensor, yielding the gamma-compressed RGB channels (R', G', B') associated with every pixel of a digital color image.

The ITU-R BT.601 standard (former CCIR 601) followed by the majority of standard-definition (SDTV) video cameras defines the perceived luminance or luma, Y', corresponding to a gamma-compressed RGB vector (R', G', B') by averaging the three components with the following weights: $Y' = 0.299R' + 0.587G' + 0.114B'$. Luma can be considered to be a computationally efficient approximation of the CIELAB lightness. In particular, by assuming that the three gamma-compressed components are normalized between 0 and 1, luma is directly utilized in this work to define the normalized JND index: $\bar{J}_j = Y'$, and hence the parameters α and β necessary for evaluating Y_1 and Y_2, respectively. Notice that since the maximum luminance level of the real scene depicted in a given digital image is not known in general, mapping the maximum luma to the maximum normalized JND index actually represents an implicit normalization of that unknown maximum luminance to the maximum level of 200 cd/m^2 utilized by Izmailov and Sokolov.

Thus, α has finally been defined as the luma of a given pixel, whereas β is obtained as the average luma of the pixels belonging to its neighborhood. For computational reasons, a 3x3 neighborhood has been considered in this work (i.e., the eight adjacent pixels). Further work is required to analyze the influence of the neighborhood's size and shape in the final result.

4. Visual attention model

Visual attention models usually generate a saliency map from a given digital image such that every image pixel is associated with a map element whose value is proportional to the visual attraction corresponding to that pixel with respect to some visual features (color, brightness, etc.). The saliency map is obtained by integrating partial saliency maps (conspicuity maps) generated for each of the visual features taken into account. For example, the well-known IKN model **Error! Reference source not found.** averages the conspicuity maps generated for three visual features: color, brightness and orientation (edginess). Thus, both color and brightness are independently processed and have a same weight in the final saliency.

This section proposes a visual attention model based on the architecture of IKN but that integrates color and brightness through the computational adaptation presented in section 3 of the perceptual model proposed by Izmailov and Sokolov **Error! Reference source not found.**. Such integration is perceptually founded and yields more consistent results than when both visual features are processed independently, as will be shown in the next section in the application scope of video tracking.

An image point is considered to be salient with respect to a given visual feature (e.g., brightness) if there is a significant difference between the value of the feature associated with that point (the center) and the values corresponding to the points within its neighborhood (the surround). This so-called center-surround antagonism occurs in the photoreceptor cells of the retina of primates and allows the visual cortex, for instance, to detect spatial edges. In practice, center-surround differences can be computed by subtracting a coarse-scale version of an image from a fine-scale version of the same image [2].

Let $\psi(x,y)=\psi_0(x,y)$ be an original image at scale 0. The image at scale *t*, $\psi_t(x,y)$, is obtained by applying Gaussian filtering to the image at scale $t-1$ and then by subsampling the result. Thus, the image at scale t has a reduction factor of 1: 2^t with respect to the original image. For instance, IKN applies 9 scales (i.e., scales 0 to 8). The original image and its successive coarser approximations constitute a Gaussian pyramid.

Let $F_t(x,y)$ be the values of a visual feature corresponding to the pixels of $\psi_t(x,y)$. A feature map, $F_{c,s}(x,y)$, is defined as the absolute value of the across-scale difference between the values of the visual feature at a fine scale c (center) and a coarse scale s (surround),

respectively: $F_{c,s}(x,y) = \left| F_c(x,y) \Theta F_c(x,y) \right|$, where the across-scale difference operator Θ interpolates the coarse scale to the fine scale and then applies point-by-point subtraction.

Feature maps are normalized in order to be able to combine maps corresponding to different visual features. The normalization operator N is defined as: $N\left(F_{c,s}(x,y)\right) = \left(M - \bar{m}\right)^2 F_{c,s}(x,y)$, where M is the value of the global maximum of $F_{c,s}$ and $\bar{m}$ the mean value of its local maxima. Similarly to IKN, several feature maps are determined for each visual feature such that $c_m \le c < c_M$ and $s = c + \delta$, with $\delta_m \le \delta < \delta_M$. In IKN, $c_m = 2$, $c_M = 4$, $\delta_m = 3$ and $\delta_M = 4$, which yields six feature maps spanning seven scales (2 to 8). Alternatively, the proposed model has been configured with $c_m = 4$, $c_M = 5$, $\delta_m = 1$ and $\delta_M = 3$, which also yields six maps spanning the five upper scales (4 to 8). The normalized maps corresponding to the same visual feature are combined into a single conspicuity map, $\bar{F}(x,y)$, through the across-scale addition operator $\oplus$, which interpolates the given maps to a same scale (e.g., scale 4 as proposed in IKN) and then performs a point-by-point addition: $\bar{F}(x,y) = \bigoplus_{c=c_m}^{c_M} \bigoplus_{s=c+\delta_m}^{c+\delta_M} N\left(F_{c,s}(x,y)\right)$.

In the end, the final saliency map $S(x,y)$ is obtained by averaging the conspicuity maps associated with every considered visual feature. In particular, seven visual features are independently taken into account and integrated in IKN: one achromatic feature, two chromatic features and four orientation features. The achromatic feature is the intensity component *I* defined in section 3.1. The chromatic features are approximately equivalent to the *RG* and *BY* components defined in section 3.1. In turn, the four orientation features are the outcome of respective Gabor filters with orientations 0, 45, 90 and 135 degrees, respectively.

Alternatively, the proposed visual attention model is based on two visual features: a joint chromatic-achromatic feature, $J_t(x,y)$, based on the perceptual model proposed by Izmailov and Sokolov, and an orientation feature, $O_t(x,y)$, based on the Sobel filter. In particular, the chromatic-achromatic feature corresponding to an image pixel $\psi_t(x,y)$ is defined as the 5-D vector: $J_t(x,y) = \left(X_{1,t}(x,y), X_{2,t}(x,y), X_{3,t}(x,y), Y_{1,t}(x,y), Y_{2,t}(x,y)\right)$, where $X_{i,t}(x,y)$ is the i-th chromatic component of the Izmailov and Sokolov perceptual model defined in section 3.1, whereas $Y_{1,t}(x,y)$ and $Y_{2,t}(x,y)$ are the achromatic components of the Izmailov

and Sokolov perceptual model evaluated at $\psi_t(x,y)$ as defined in section 3.2, respectively. Taking this into account, the joint chromatic-achromatic feature map is defined as: $J_{c,s}(x,y) = \left| J_c(x,y) \Theta J_s(x,y) \right|$, where the point-by-point subtraction inherent to the across-difference operator Θ is defined in this case as the perceptual difference ΔS_{ij} between the colors corresponding to both 5-D points, ${}^{i}J_c(x,y)$ and ${}^{j}J_c(x,y)$, according to the Izmailov and Sokolov model:

$$ {}^{i}J_c(x,y) - {}^{j}J_c(x,y) = \sqrt{\left(\Delta C_{ij,c}\right)^2 + {}^{i}X_{3,c}(x,y)\,{}^{j}X_{3,c}(x,y)\left(\Delta W_{ij,c}\right)^2} $$

with $\left(\Delta C_{ij,c}\right)^2$ and $\left(\Delta W_{ij,c}\right)^2$ defined in section 2. The six feature maps are finally normalized and combined with the across-addition operator, yielding the joint chromatic-achromatic conspicuity map $\bar{J}(x,y)$: $\bar{J}(x,y) = \bigoplus_{c=c_m}^{c_M} \bigoplus_{s=c+\delta_m}^{c+\delta_M} N\left(J_{c,s}(x,y)\right)$.

In turn, the orientation conspicuity map in this work is computed by means of the Sobel operator instead of the four Gabor filters proposed in IKN. The Sobel edge detector has already been shown to be a computationally efficient alternative to the four Gabor filters applied in IKN. In particular, let $I_t(x,y)$ be the gray-scale image corresponding to $\psi_t(x,y)$, with the brightness feature I being computed as defined in section 3.1. The image orientation feature is defined as: $O_t(x,y) = \sqrt{\left(I_t(x,y) * S_x\right)^2 + \left(I_t(x,y) * S_y\right)^2}$, where S_x and S_y are the 3x3 horizontal and vertical kernels of the Sobel operator, respectively, and * denotes the convolution operator. The orientation conspicuity map $\bar{O}(x,y)$ is then defined as: $\bar{O}(x,y) = \bigoplus_{c=c_m}^{c_M} \bigoplus_{s=c+\delta_m}^{c+\delta_M} N\left(O_{c,s}(x,y)\right)$, with $O_{c,s}(x,y) = \left| O_c(x,y) \Theta O_s(x,y) \right|$.

The final saliency map $S(x,y)$ is defined as a weighted average of the two conspicuity maps defined above: $S(x,y) = w_J \bar{J}(x,y) + w_O \bar{O}(x,y)$. Both conspicuity maps are given the same weight, $w_J = w_O = 0.5$, similarly to IKN. According to this formulation, saliencies are normalized between zero and one, with a value close to one representing a large visual attraction.

5. Video tracking based on visual attention

The visual attention model described in the previous section is advantageous for detecting salient visual features, which can help improve other higher-level computer vision tasks. In particular, this section describes a simple video tracker based on the straightforward matching of image blocks extracted from those features. The proposed tracking algorithm assumes a video sequence acquired with a stationary camera. The moving objects in the scene are extracted by applying the efficient background subtraction algorithm proposed in [4], which segments the current image into both background and foreground pixels. A connected-component labeling algorithm applied to the foreground pixels determines isolated regions hereafter referred to as blobs. Blobs with an area below a predefined small threshold are discarded in being considered to be due to noise.

Given a set of separate blobs extracted from the current image, the goal of a multi-object video tracking algorithm is to associate (match) every new blob with another candidate blob extracted from previous images, provided both blobs are visually similar. In particular, the proposed algorithm determines the visual similarity between every blob from the current image and all the blobs extracted from the last *M* images (*M* has been set to 25 in this work). A historic list of active blobs is thus kept, which contains blobs that have been successfully matched and blobs that have not. The latter blobs have an associated life counter initially set to zero that is incremented after every new image is processed. If the counter reaches *M*, the associated unmatched blob is removed from the list. If a new blob can be matched with a previous blob, the previous blob in the list is substituted for the new blob. Otherwise, the new blob is appended to the list. In both cases, the life counter of the new blob is reset.

In addition to geometrical information such as area, width, height and centroid coordinates, the algorithm keeps for every blob the portions of both the original image and the saliency map that intersect with the blob's binary mask. In order to reduce the saliency of visual features belonging to occlusion edges, that is, image contours due to the frontier between the foreground object and the background of the scene, which do not thus characterize the interior of the tracked object and hence the object itself, the blob's saliency map is modulated by multiplying it by the morphological distance transform of the blob.

The visual similarity between a new blob B_{new} and a previous blob B_{old} is determined as follows. The (x, y) coordinates corresponding to the maximum value of the aforementioned modulated saliency map associated with B_{new} are determined. This corresponds to the location

of the most visually salient feature in the interior of the blob. A rectangular image block of size w×h centered at that location is considered in the blob's image, with w and h being a fraction of the blob's width and height, respectively (in this work, the block's size is one fifth of the blob's size). This image block extracted from B_{new} is searched over the whole image associated with B_{old} through naive block matching, using the sum of absolute differences (SAD) as the dissimilarity measure between two blocks. Once the first block has been processed and its smallest dissimilarity measure with B_{old} recorded, a new block centered at the second most salient visual feature in B_{new} is extracted.

In order to avoid a concentration of blocks around the maximum saliency, all saliency values in a neighborhood of $2w \times 2h$ centered at the maximum saliency found before are reset. This corresponds to the concept of local inhibition or "inhibition of return" typically found in visual attention (e.g., [2]). Once the second image block is found, its minimum dissimilarity with B_{old} is obtained again. The process is iterated until a maximum of N image blocks have been extracted from B_{new} (N has experimentally been set to 5 in this work), or the modulated saliency map becomes null due to the local inhibition. For the sake of completeness, the same block extraction process is applied to B_{old} and the minimum dissimilarities of those image blocks with B_{new} are computed.

The visual similarity between the two blobs is finally obtained as the inverse of the average of dissimilarities estimated as described above. The visual similarities obtained in that way and their associated blob pairings are then sorted in descending order of similarity, excluding those pairings whose similarity is below a minimum threshold. At this point, the algorithm chooses the pairing with the maximum similarity. Its associated new blob is paired with its corresponding old blob. Then, the next pairing is considered, and the new blob is paired to the old blob, except if the old blob has already been paired before. This greedy procedure is applied until no more pairings are available. The new blobs that have not been able to be paired after the whole process are considered to be new objects and are appended to the historic list as described above.

6. Experimental results

This section presents experimental results of the performance of the video tracker proposed in the previous section based on the visual attention model proposed in section 4, and compares it with alternative visual attention models and video tracking algorithms. Four video sequences (denoted by A, B, C and D) from the PETS dataset [5] have been segmented into fragments containing contiguous frames with multiple moving objects in order to avoid an undesired bias due to periods without moving objects. Every fragment is an independent test video sequence. Those initial fragments, as well as the video sequences obtained after applying the proposed technique, are shown in the companion website[1].

Every test frame has been annotated in order to define the detection ground-truth, which consists of a binary mask indicating the pixels that constitute the moving objects within that frame (i.e., the foreground pixels), and the tracking ground-truth, which consists of the numerical identifier, coordinates of the bottom-left corner of the bounding box and dimensions of the latter for every object within the frame. That identifier is unique for every object from the test sequence. The video tracker consists of an initial detection stage that segments every new frame into separate objects (blobs), and a tracking stage that determines the numerical identifier of every object within the frame.

The tracking performance for all the experiments has been evaluated through the following measures described in [6]: ATA and SFDA (CLEAR metrics), and NMODA, MOTA, MOTP, and NMODP (VACE metrics). All those measures are normalized between zero and one, with one representing the best performace. ATA (Average Tracking Accuracy) gives the average percentage of spatial overlap over the whole test sequence between every detected object and the ground-truth object corresponding to the numerical identifier determined by the tracking algorithm for that detected object. Therefore, it is a quality measure of both detection and tracking. SFDA (Sequence Frame Detection Accuracy) is similar to ATA but only considering the detection stage (i.e., it measures the spatial overlap between every detected object and its largest overlapping ground-truth object). NMODA (Normalized Multiple Object Detection Accuracy) estimates the average performance of the tracking stage over the whole test sequence by only considering the number of unpaired and wrongly paired objects. MOTA (Multiple Object Tracking Accuracy) is similar to NMODA but also considering mismatches between object identifiers. NMODP (Normalized Multiple Object Detection Precision) and MOTP

[1] Full resolution test video sequences, detailed experimental data and resulting video sequences can be found in the companion website: *https://sites.google.com/site/fgmtracking/*

(Multiple Object Tracking Precision) take into account the spatial overlap between detected and ground-truth objects, very similarly to SFDA and ATA, respectively.

Two groups of experiments have been performed as described in the following two sections. The first group evaluates the tracking algorithm proposed above by considering different visual attention models, showing that the proposed visual attention model is superior to the other tested alternatives in this scope. In turn, the second group compares the performance of the proposed tracking algorithm and visual attention model with well-known, publicly available tracking algorithms, showing that the proposed scheme is also advantageous.

6.1. Tracking performance for different visual attention models

The performance of the video tracking algorithm proposed in Section 5 has been evaluated with the following visual attention models: (a) The computational adaptation of the Izmailov and Sokolov model described in section 4, denoted by IS. (b) The model by Itti, Koch and Niebur [2], denoted by IKN. (c) The variation of IKN proposed by Won, Lee and Son [7], denoted by WLS. (d) The variation of IKN and WLS previously proposed by the authors [8], denoted by HYBRID. (e) The model proposed by Hou and Zhang [9], denoted by HZ. (f) The model proposed by Maruta, Sato and Isshi [10], denoted by MSI. (g) The model proposed by Judd, Ehinger, Durand and Torralba [11] based on a combination of different levels of image features, denoted by JEDT. (h) The method proposed by Avraham and Lindenbaum [12], based on a validated stochastic model that estimates the probability that an image part is of interest, referred to as ESALIENCY. (i) A simple model in which image blocks are extracted by using as a saliency measure the "cornesness" function of the Harris corner detector, denoted by HARRIS. (j) A simple model that uses as a saliency measure the gradient magnitude estimated with the Sobel operator, denoted by SOBEL. (k) A naive algorithm that randomly extracts image blocks, denoted by ALEAT.

Two variations have been tested for the methods based on saliency maps (all except ALEAT) depending on whether the image blocks are extracted according to either the local maxima of the saliency map (denoted by MAX), or the local maxima of the sum of saliencies over a local window of the block's size (denoted by SUM). The latter aims at filtering saliency noise. In turn, two additional variations have been tested for the multiscale methods IS, IKN, WLS and HYBRID depending on whether the integrated feature maps belong to the low range of scales (2 to 6), denoted by DOWN, or the high range of scales (4 to 8), denoted by UP. Finally, two variations of the proposed IS model have been considered depending on whether

the Sobel operator is used to extract the orientation maps as described in section 4 (denoted by S) or the same bank of Gabor filters utilized in IKN is applied instead (denoted by G).

Since all the evaluated trackers have the same blob detection stage, exclusively differing on the applied visual attention model, only the tracking-related metrics ATA and MOTA have variations among them. In particular, Figures 1 and 2 show the complement of both metrics (i.e., 1-ATA and 1-MOTA) corresponding to the first two considered PETS sequences (A, B) for the different tested visual attention models. Every data point represents the average performance for the fragments extracted from the corresponding sequence. The figures corresponding to the other two PETS sequences and tables with the associated numerical data are provided in the companion website.

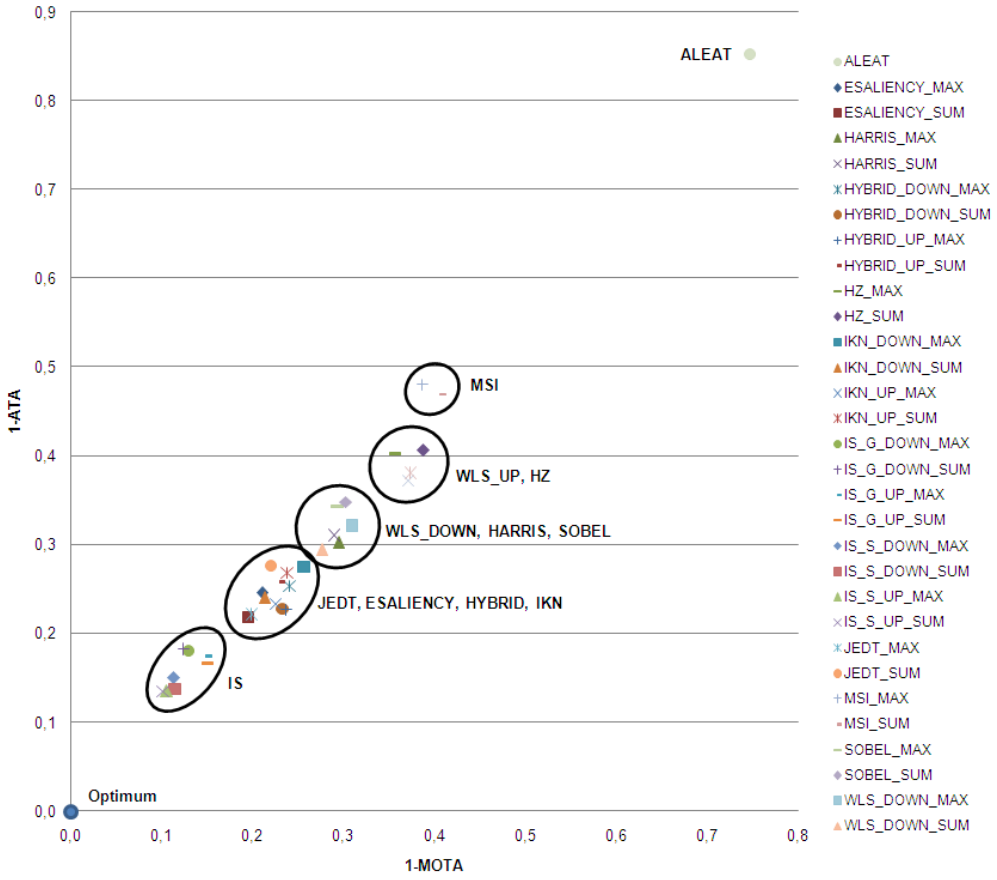

Figure 1. Tracking performance for different visual attention models (test set A).

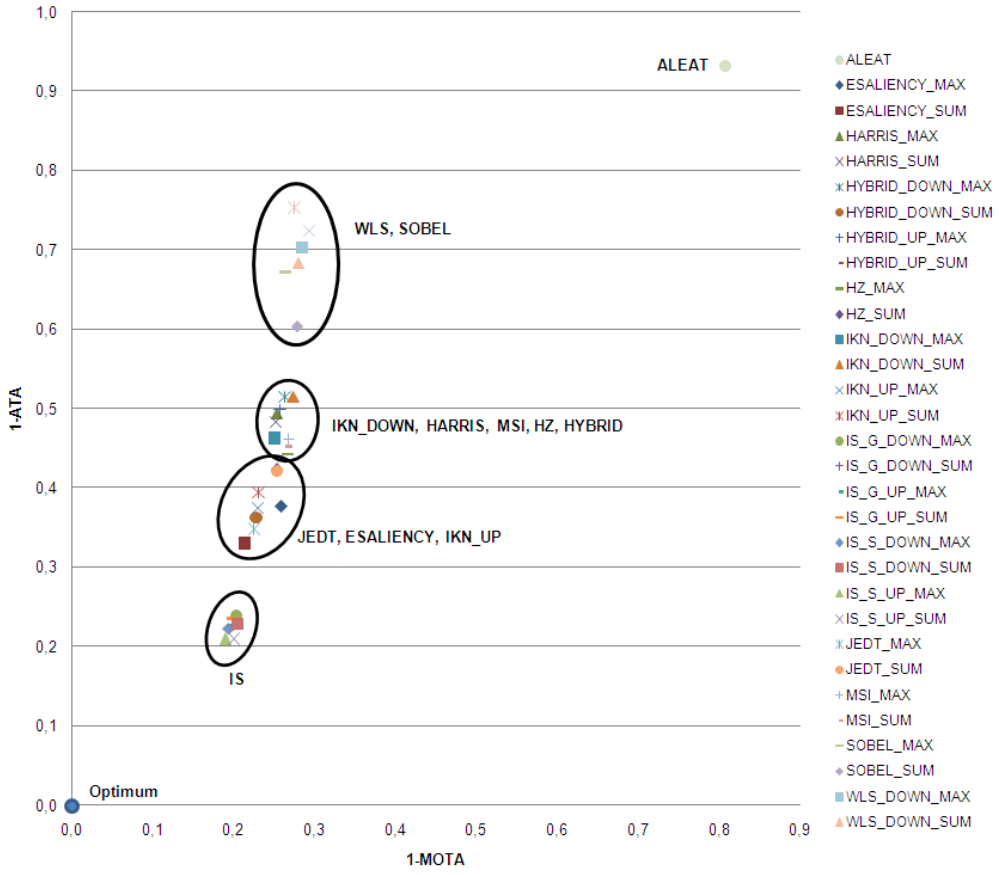

Figure 2. Tracking performance for different visual attention models (test set B).

These results show that the different variations of the proposed visual attention model yield significantly better tracking results than the other tested models. Among the latter, IKN, ESALIENCY and JEDT are the models with the closest performance. Interestingly enough, the simple and extremely efficient models based on both the Harris detector and the Sobel operator yield a very competitive tracking performance, in some cases comparable to that of far more complex, state-of-the-art visual attention models.

With respect to the variations of the proposed visual attention model, these results indicate that the Sobel version has a better tracking performance than the version based on Gabor filters, in addition to its much higher computational efficiency. With respect to SUM and MAX, both alternatives have shown a similar performance. Notwithstanding, SUM is preferred due to its noise filtering properties and since it can be efficiently computed through integral images. Finally, working on the upper range of scales (UP) generally yields a better performance than when the lower range (DOWN) is considered, in addition to its corresponding higher performance. Taking these conclusions into account, the video tracker proposed in section 5,

referred to as FGM, has been configured according to the IS_S_UP_SUM visual attention model.

7. Conclusions and future work

A computational adaptation of the theoretical perceptual model originally proposed by Izmailov and Sokolov in the scope of psychophysics has been proposed. The main advantage of that model is that it integrates the perception of both color and brightness through a formulation consistent with human perception. That computational model has then been applied to the definition of a new visual attention model by adapting the well-known IKN model, with the main advantage that the proposed model deals with both color and brightness in a seamless way, whereas IKN and other visual attention models process both visual features independently. In order to objectively assess the benefits of the proposed visual attention model, a simple video tracker has been proposed in which image blocks are extracted in regions containing salient visual features. Experimental results with real video sequences show that the developed video tracker endowed with the proposed visual attention model yields a significantly higher performance than when other visual attention models are utilized instead.

Immediate work will focus on the optimization of the proposed technique in order to make it suitable for real time video processing. This will require a careful analysis of the scales and, hence, feature maps that are integrated in the visual attention model. In addition, it will be necessary to study either simplifications or approximations of the analytical formulations that constitute the proposed computational model. Furthermore, by taking advantage of the parallel nature of the proposed technique, it will be necessary to analyze the benefits of applying multi-core parallel architectures and GPUs. In addition, future work will focus on the development of other higher-level applications of the proposed visual attention model to computer vision and robotics.

References

[1] C. Izmailov and E. Sokolov, "Spherical model of color and brightness discrimination," Psychological Science, 2(4):249–259, 1991.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1254–1259, 1998.

[3] NEMA, Digital Imaging and Communications in Medicine (DICOM): Part 14: Grayscale Standard Display Function. National Electrical Manufacturers Association, 2009.

[4] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic videos analysis for adaptive content delivery and automatic description," IEEE Transactions on Circuits and Systems of Video Technology, 10(15):1200–1209, 2005.

[5] "Performance evaluation of tracking and surveillance (PETS 2006), http://www.cvg.rdg.ac.uk/PETS2006/data.html," 2006.

[6] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text and vehicle detection and tracking in video: Data, metrics and protocol," IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(2):319–336, 2009.

[7] W. Won, M. Lee, and J. Son, "Implementation of road traffics sings detection using low-level features," in Proc. of CIT 2008.

[8] V. Fernández-Carbajales, M. A. García, and J. M. Martínez, "Improving the efficiency and accuracy of visual attention," in Proc. of AVSS 2011.

[9] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in Proc. of CVPR 2007.

[10] H. Maruta, M. Isshi, and M. Sato, "Salient region extraction based on local extrema on natural images," in Proc. of ICIP 2010.

[11] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in Proc. of ICCV 2009.

[12] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): meaningful attention using stochastic image modeling," IEEE Transactions on Pattern Analysis and Machine Intelligence, 3284):693–708, 2010.