Pattern Recognition 60 (2016) 571–584

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Visual attention based on a joint perceptual space of color and brightness for improved video tracking

Víctor Fernández-Carbajales, Miguel Ángel García *, José M. Martínez

Video Processing and Understanding Lab, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049, Madrid, Spain

ARTICLE INFO

Article history:
Received 30 March 2015
Received in revised form
6 June 2016
Accepted 8 June 2016
Available online 16 June 2016

Keywords:
Visual attention
Color and brightness perceptual model
Saliency maps
Video tracking

ABSTRACT

This paper proposes a new visual attention model based on a joint perceptual space of both color and brightness, and shows that this model is able to extract 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 paper proposes a computational model that allows the 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 particular application scope of video tracking than well-known visual attention models that process color and brightness separately.

1. Introduction

Video tracking aims at determining the trajectory of moving objects over the frames of a video sequence [1,2]. Once one or several objects have been selected in an initial frame, the tracking process identifies those objects in subsequent frames, that is, it associates the objects identified in a given frame with the objects already identified in previous frames.

Video tracking is still an open problem due to the huge complexity of the scenarios where it must usually be applied. For instance, for video surveillance applications in complex environments, it must be able to cope with deficient illumination conditions, a large number of objects usually of a limited size due to the overhead location of the capturing devices, rather similar visually and, in addition, which interact with other objects quickly, frequently and according to complex patterns.

Therefore, one of the key aspects and main challenges in video tracking is the proper characterization of the target objects to be tracked in order to be able to identify them without ambiguity in subsequent frames provided they are still present. That characterization must be based on measurable features extracted from the objects themselves, mainly related to their motion pattern and visual appearance. Regarding the latter, it is crucial to be able to detect visual features that can be reliably matched between successive frames.

The visual features that have traditionally been utilized for tracking are corners, contours and regions. Unfortunately, their usefulness can be significantly compromised when considering the complex scenarios and limited image qualities discussed above. On the one hand, corners may be perceived with not enough contrast or definition to be reliably detected. Even if they can be detected, nothing guarantees that their surroundings be discriminant enough as to be able to visually identify them and their associated objects in successive frames. In addition, the presence of image noise may lead to the detection of false corners that end up yielding wrong matchings. In turn, contours can even be more unreliable in this context, since, in general, they do not guarantee a correct delimitation and hence identification of the tracked objects in cluttered scenes with self-occlusions and groupings. Something similar applies to regions, as their appearance (typically color or texture distributions) is usually characterized in a rather approximate and, hence, ambiguous way through a wide variety of statistical models (histograms, spatiograms, kernels, mixtures of Gaussians, etc.), and on the other hand, since regions are very sensitive to self-occlusions and groupings due to their somewhat global nature.

It is thus necessary to consider alternative approaches for extracting reliable visual features that allow a better characterization of the tracked objects based on their visual appearance. Such an alternative is visual attention [2,3]. In particular, visual attention models attempt to mimic the cognitive process that allows human observers to focus on the most relevant elements (features) of a visual scene while ignoring the others. In other words, these models aim at finding visual features whose saliency makes them outstand from the scene.

* Corresponding author.
E-mail addresses: victor.fernadez@uam.es (V. Fernández-Carbajales), miguelangel.garcia@uam.es (M.Á. García), josem.martinez@uam.es (J.M. Martínez).

http://dx.doi.org/10.1016/j.patcog.2016.06.007
0031-3203/

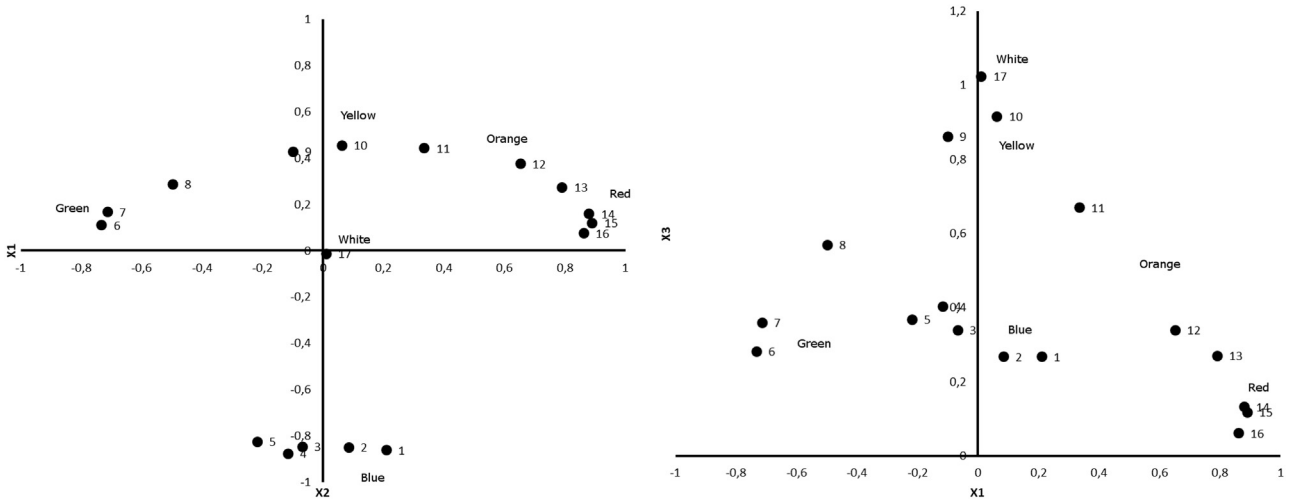

Fig. 1. Projection of 17 equibright colors on (left) the X_1X_2 plane and (right) the X_1X_3 plane.

Those so-called saliency-based visual attention models, such as the well-known model proposed in [4] and its subsequent variations, process various visual features independently, in particular: color (chromaticity), brightness and orientations (texture). In that way, large local variations of color will yield high color saliency measures no matter the brightness associated with those regions. This behavior poses a problem in case of dark or poorly illuminated regions, in which color cannot be accurately perceived due to insufficient brightness. In those low-brightness conditions, color variations may simply be the result of image noise, in whose case, the aforementioned visual attention models will paradoxically tend to detect as reliable features image regions that are not reliable nor distinctive at all. Thus, the alleged advantage of visual attention in terms of a better visual characterization of image objects is lost. Unfortunately, those low-brightness conditions are rather typical in real applications, such as video surveillance and monitoring.

The problems associated with color perception at dark regions (also at over-illuminated regions with desaturated colors) have previously been identified in the context of computational color constancy [5], which aims at modifying the color of the pixels of a given image such that it looks the same no matter the colors of the light sources illuminating the scene. In particular, color strength [6] has been proposed as a heuristic estimate of the reliability of a pixel's hue with the rationale that the hue of dark and poorly saturated pixels is not reliable. The goal is that color constancy only takes into account reliable pixels. This problem comes from the fact that the color's hue is independent of its intensity and saturation, which applies to all color invariant models [7].

This paper proposes a new saliency-based visual attention model that jointly processes color and brightness according to a biologically inspired theoretical perceptual model originally proposed by Izmailov and Sokolov [8] in the scope of psychophysics. Since that model is only applicable to perceptual data directly drawn from psychophysical experiments, the present paper proposes a computational model that allows the direct application of Izmailov and Sokolov's theoretical model to digital images. Experimental results with real video sequences show that the proposed visual attention model significantly improves the task of video tracking by finding more discriminant visual features, especially when dealing with objects that are very similar visually, thus yielding more accurate results in the application scope of video tracking than well-known visual attention models that process color and brightness separately.

The organization of this paper is as follows. The theoretical perceptual model proposed by Izmailov and Sokolov is described in Section 2. Section 3 proposes a new computational model for applying the original Izmailov and Sokolov's theoretical model to the determination of color differences in digital images. Section 4 describes a new visual attention model based on the aforementioned computational adaptation of Izmailov and Sokolov's perceptual model. Section 5 proposes a simple video tracker based on the previously proposed visual attention model. Section 6 shows experimental tracking results with the proposed visual attention model and a comparison with alternative models. Finally, conclusions and future lines are presented in Section 7.

2. Izmailov and Sokolov's perceptual model

The theoretical perceptual model proposed by Izmailov and Sokolov in [8] yields a metric color space in which every point represents a specific color and Euclidean distances are proportional to perceived color differences. This model was derived through psychophysical experiments with human subjects and multidimensional scaling analysis techniques. Actually, three sets of experiments were carried out as described below.

In the first experiment, pairs of colors were projected simultaneously, one occupying the center of the image, referred to as the *stimulus*, and the other occupying the rest, referred to as the *background*. Both the luminance and the wavelength of the background were kept constant by considering a neutral light. In turn, the wavelength of the stimulus was varied over 16 discrete values between 425 nm and 675 nm plus the white color (i.e., 17 colors), while keeping a constant brightness for all colors. The subjects were asked to indicate their perceived color difference between the stimulus and the background in a scale from 0 (same color) to 9 (maximum color difference).

An implementation of the Shepard–Kruscal algorithm for multidimensional scaling analysis was then applied to the aforementioned subjective color differences. As a result, a 3D semi-spherical space with axes X_1, X_2 and X_3 was inferred. Fig. 1 shows the mapping of the 17 equibright colors to the X_1X_2 and X_1X_3 subspaces.

The main conclusion from these results is that the perceptual *chromatic* difference ΔC_{ij} between two equibright colors $({}^iX_1, {}^iX_2, {}^iX_3)^T$ and $({}^jX_1, {}^jX_2, {}^jX_3)^T$ in this space can be estimated by means of the interpoint Euclidean distance, where $\Delta X_\chi = {}^iX_\chi - {}^jX_\chi$:

$$(\Delta C_{ij})^2 = (\Delta X_1)^2 + (\Delta X_2)^2 + (\Delta X_3)^2. \quad (1)$$