sciendo

GESTALT THEORY, DOI 10.2478/gth-2024-0011
© 2024 (ISSN 2519-5808); Vol. 46, No. 2-3, 121–155

Original Contributions - Originalbeiträge

Luigi Burigana

Field dynamics of visual perception as framed in Markov random fields of computer vision

1. Introduction

The term “field” occurs in the languages of empirical sciences, where it may take on various meanings. One notable meaning is the following: a *field* is a system of entities that are separately located in a definite *space* and are subject to *local interactions*, in such a way that *extensive properties* shown by the system may be predicted and explained in terms of those local interactions. The adjective “local” in this statement signifies that the interactions in question only involve entities that are close or adjacent to one another in the topology of the presumed space. In turn, the adjective “extensive” signifies that the properties to be explained are attributes of more or less extended regions of the given space, that is, they may be collective properties possessed by more or less large sets of component entities, possibly collective properties revealed by the system of entities in its entirety.

In psychology it is a generally known fact that the concept of field thus characterized played a substantial role in the emergence and development of the Gestalt theory of visual perception. The well-known statement that “a theory of perception must be a field theory” by Köhler (1940, p. 55) is a spotlight on this theoretical and historical fact. For the very nature of the subject matter, any theory of visual perception is a multilevel conceptual construction, as it involves entities belonging to distinct *ontological levels*, such as the levels of distal stimuli, proximal stimuli, excitation patterns on the retinas, signals along the neural pathways, consequent projections on the visual cortex, and, especially important, the level of events subjectively experienced by an observer, that is, the level of phenomenal data or visual awareness. Within the Gestalt theory, the concept of field has been separately applied to more than one of these ontological levels; also, correspondences between fields inherent in distinct ontological levels have been considered.

In this regard, a distinction may be made between the concept of “physiological field” as termed by Koffka (1935, pp. 52-53) and the concept of “psychophysical field” as I call it in the present study. As referred to an act of vision, the former is a system of entities and interactions presumed to take place in the brain of an observer during that act; the latter is instead a system of entities and interactions that occur in the perceptual experience of the observer, and also includes relationships between those entities and the optical stimulation, whether this is described in terms of proximal or of distal stimuli[1]. The current statuses of these two specimens of the field concept of Gestalt theory are quite different. On the one side, it is a generally shared view in perceptual psychology that the concept of physiological field as featured by Gestalt theory had long ago lost its scientific credibility because it failed to pass relevant empirical tests and is not consistent with the acquisitions of contemporary visual neuroscience. This judgement especially concerns the so-called "electrical field theory" of cortical integration, which had been advanced on presuming similarities between the cortical processes in an act of vision and the dynamics of electrical charges applied to a homogeneous conducting system (Wagemans, 2014). On the other side, what I call psychophysical field theory turns out to be a paradigm of full present-day relevance and vitality for considering the kinds of problems that are currently and effectively addressed in the experimental research on visual perception. These may be problems concerning how the perceptual aspect of a part of a scene may depend on properties residing in separate places of the stimulus array, how certain perceptual characteristics may be influenced by other perceptual characteristics, or how variations caused in a certain region of a perceptual scene may induce variations in other regions of the same scene. All such research questions are linked to the general concept of field dynamics as defined in Gestalt theory and fall under the heading of "psychophysical field" as meant in this study. This is the only aspect of the Gestalt-theoretic concept of field that will be considered in our discussion. No further mention will be made of physiological fields.

By looking in another direction within vision science, a reader can see that in the area of *computer vision* a rich repertoire of models have been developed that are classified as *Markov random fields* (MRFs). These belong to the general class of probabilistic graphical models, that is, models in whose construction and analysis the mathematical theories of graphs and of probabilities come into close collaboration (Koller & Friedman, 2009). In particular, the probabilistic character of such models qualifies them as suitable guides for tasks of artificial vision when these tasks are interpreted and carried out as processes of inference under uncertainty conditions. The term "field" in the title "Markov random field" is endowed with the same general meaning described above. In other words, an MRF is itself a system of entities that are separately located in a suitable space and are subject

[1] The term "psychophysical" in the expression "psychophysical field" is in line with the concept of "outer psychophysics" as characterized by Fechner (1966, p. 9), as far as dependences between stimulus and percept properties are concerned. The meaning here assigned to "psychophysical field" is wider, however, because it includes interdependences between perceptual properties.

to local interactions or constraints in such a way that the configuration taken on by any extensive portion of the system can be predicted (with some residual uncertainty) by the aid of such local interactions or constraints.

The major aim of this article is to highlight how basic choices in the construction of MRFs in *computer vision* are consistent with the guidelines followed in the *psychological study* of human visual perception, in particular the guidelines implied by *Gestalt theory* with its focus on field dynamics and the psychophysical field paradigm as we termed it. Thus, my basic claim is that the occurrence of the term "field" in both contexts is not a mere verbal coincidence, as this term signals the existence of substantial correspondences between the two theoretical views. Conversely, the existence of such correspondences suggests that definite formal developments in the theory of MRFs may provide useful cues for formal modelling in the psychology of visual perception, thus moving back from the area of computer vision to this branch of psychological science. My analysis will include some specific suggestions in this respect.

From a historical point of view, I note that a direct evolutionary connection between the two paradigms looks improbable. Indeed, references to Gestalt theory in the computer vision literature dealing with MRFs are quite rare and concern aspects only indirectly related to field dynamics (e.g., Zhu, 1999; Zhu & Wu, 1999; Ren, Fowlkes, & Malik, 2008). Rather, the stated correspondences between the two paradigms may be valued as the evidence of a family resemblance due to the existence of common ancestors. Indeed, it is well known that in the rise of Gestalt theory (in the 1910s), especially in the theoretical investigations by Wolfgang Köhler, a significant part was played by analogies suggested by the field theory developed in the previous decades within physical science, in particular the field theory of electromagnetic phenomena as formalized by James Clerk Maxwell (1873). Now, precisely this field theory within physical science is also at the origin of probabilistic models in statistical mechanics, of which the MRFs constitute notable exemplars (Spitzer, 1971; Richey, 2010, pp. 387-389). In this sense we may state that the two paradigms under question have common ancestors. The evolutionary line that led from the field theory of late 19-th century physics to the MRFs of today's computer vision did *not* pass through the field theoretic views of Gestalt theory. Rather, it passed through the views of statistical mechanics and the definition of MRFs as a notable type of probabilistic graphical model, and then the application of such models in spatial statistics and the restoration of images corrupted by random noise. The emergence of those models as theoretical and computational tools of computer vision may reliably be dated to the mid-1980s (Cross & Jain, 1983; Geman & Geman, 1984; Besag, 1986).

This article is organized as a sequence of four waves on a fixed set of concepts. These are the concepts of graphical structure of an MRF, network of potentials constructed on that structure, probabilistic structure implied by this network, and inferences guided by this probabilistic structure. In Section 2, these concepts are defined with reference to MRFs in their generality, so that their mathematical names and meanings are made explicit. In Section 3, the special forms are defined that these concepts take on when MRFs are constructed for inference tasks in computer vision. The section ends with a table in which a representative sample of studies from the computer vision literature are listed and the inference tasks addressed in them are specified. These tasks include image segmentation, object detection, visual transparency, stereo vision, depth map construction, lightness effects, gesture recognition, and others. In Section 4, significant correspondences between MRF concepts of computer vision and salient tenets of perceptual psychology are highlighted, with special reference to the field theoretic views of Gestalt theory. In Section 5, I discuss how the flexibility of the components of MRFs may produce meaningful suggestions concerning formal modelling in perceptual psychology. Lastly, in Section 6 some remarks are added concerning concepts that lie outside the mainstream of my discussion but are significantly related to it.

2. Markov random fields in general

An MRF is a mathematical construct. Its components are termed by names and indicated by symbols whose precise meanings are to be found in the formal definition of the construct and in the mathematical disciplines that are relevant to it. The task I address in this section is the description of the basic organization of MRFs and the specification of the names and symbols by which its components are indicated. For the sake of generality and simplicity, here I refer to *generic* MRFs, that is, MRFs not necessarily associated with tasks of visual inference, as instead happens in their use within computer vision. The additional components and distinctions that become apparent within the MRF concept when this is associated with vision inference tasks constitute the topic of the next section.

Any MRF is built on a graph-theoretic platform, which is an *undirected graph* (V,E). The component V is called the *set of nodes* and amounts to a finite set of n symbols. Lower-case letters such as u and v are here used to denote individual elements of set V. The component E is called the *adjacency relation* and amounts to a set of unordered pairs of distinct nodes — formally, E is a subset of the set $V^{\{2\}}$ of $n(n\text{-}1)/2$ possible unordered pairs of distinct elements of V. In any application, each node v in V is associated as a symbol with some entity located in a definite real space, and for any two distinct nodes $u{\neq}v$ in V, the pair $\{u,v\}$ is accepted as a member of the relation E if the entities symbolized by u and v are neighbours to each other (according to a suitable criterion) in the presumed real space. Thus, the relation E describes a spatial aspect of the entities in question that is of a qualitative (topological) rather than of a quantitative (geometrical) kind. A graph (V,E), being formed of two sets, is a set-theoretic structure. However, that structure may be given a pictorial representation as a diagram on the plane, by depicting the nodes in V as points and the pairs in E as undirected lines (also called edges) between points. In Figure 1, the diagrams are shown of two graphs: on the left a graph formed of 16 nodes and 24 edges, and on the right a graph formed of 16 nodes and 42 edges. Note that such a diagram is itself a spatial structure, its reference space being the pictorial surface on which points and lines are signed. Of course, such a spatial structure must be kept conceptually apart from the structure formed of real entities located in a real space, which is presumed to be (partially) described by a graph.

In the construction of an MRF, to each node $v \in V$ in the graph a definite *variable* T_v is associated. Typically, this is a composite variable, that is, a set $T_v=(T_{v,1},\ldots,T_{v,k})$ of elemental variables, with each representing some (known or unknown) property of the real entity symbolized by the node v. To each elemental variable $T_{v,h}$ there corresponds a *range* of possible values. Here, the symbol $T_{v,h}°$ is used to denote the range of the variable $T_{v,h}$, and the corresponding lower case letter $t_{v,h}$ will denote a generic element of that range, that is, a generic possible value or state of the variable. Thus, to the composite variable T_v there corresponds the Cartesian product $T_{v,1}°\times\ldots\times T_{v,k}°$ as its range $T_v°$. For example, if $T_v=(T_{v,1},T_{v,2},T_{v,3})$, where $T_{v,1}$ and $T_{v,3}$ are binary variables with range {0,1}, while $T_{v,2}$ is a ternary variable with range {1,2,3}, then the range $T_v°$ will be the product {0,1}×{1,2,3}×{0,1}, which is a set of 12 triples. On the whole, we then have that, in constructing an MRF, a full system of variables is hypothesized:

$$T=(T_v\colon v \in V)$$

which contains a component T_v residing in any node v of the graph. The system T is called the *total variable* of the model. Its range $T°$ will be the Cartesian product of the ranges of its components, that is:

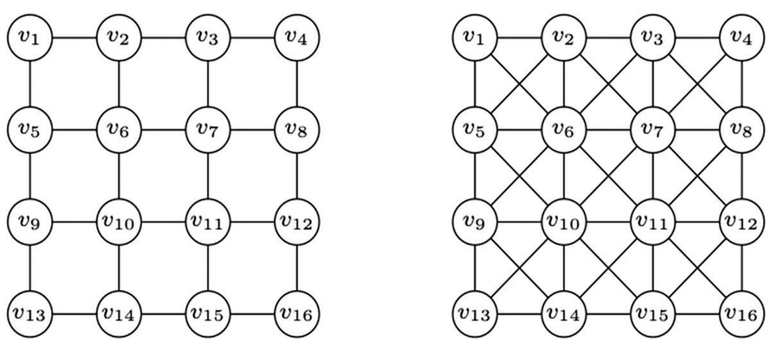

Fig. 1. Pictorial representations of two undirected graphs.

$$T^\circ = \prod_{v \in V} T_v^\circ.$$

Note that, in discussing MRFs, the term "variable" is subject to multiple uses. The term applies not only to the total variable T and to each main component T_v of it, but also to each elemental component $T_{v,h}$ of T_v, to any set of such elemental components, and, more generally, to any selection of components that may be highlighted within the total variable T.

The building blocks of an MRF are called *potentials*. Any single potential has a *scope* W, which is a selected variable in the sense now described. The potential itself is a non-negative valued function $c(W)$ whose domain is the range W° of its scope variable. Concerning the meaning, we have that a potential $c(W)$ is meant to evaluate how much each possible configuration of W (i.e., each element of the range W°) *differs* from a definite desired condition. Therefore, for arbitrary $w,w' \in W^\circ$, the event $c(w)<c(w')$ would mean that the configuration w differs less than the configuration w' from the desirable condition represented by $c(W)$, so that (in that specific regard) the assignment $W=w$ would be preferable to the assignment $W=w'$. In particular, if $c(w)=0$, then we would have that the configuration w fully satisfies the desirable condition expressed by the potential. For example, suppose that the scope of a potential $c(W)$ is a pair $W=(T_{u,h}, T_{v,h})$ of elemental real-valued variables that are similar in meaning but reside in two distinct and adjacent nodes u and v of the graph, and that the desired condition is that those variables take on equal values — this would be a "smoothness" requirement. Then the potential could be the function $c(W)=|T_{u,h}-T_{v,h}|$ whose value increases for increasing diversity between the two quantities. For another example, suppose that the scope of a potential $c(W)$ is a triple $W=(T_{v,1}, T_{v,2}, T_{v,3})$ of elemental real-valued variables residing in the same node v and that the desired condition is that the third of them equals the product of the first two. Then the potential could be the function $c(W)=(T_{v,1}\times T_{v,2}-T_{v,3})^2$, whose domain is the range $W^\circ = T_{v,1}^\circ \times T_{v,2}^\circ \times T_{v,3}^\circ$ of the scope variable.

Relative to a potential $c(W)$ thus defined, we may consider the set V_W of the nodes in the graph in which the elemental variables constituting the scope W are residing. For example, in the first of the two potentials now described, we have $V_W=\{u,v\}$ (a pair of nodes, because the variables $T_{u,h}$ and $T_{v,h}$ in that potential reside in two distinct nodes u and v), whereas in the second we have $V_W=\{v\}$ (one single node, because the variables $T_{v,1}$, $T_{v,2}$, and $T_{v,3}$ in that other potential are distinct in meaning but reside in the same node v of the graph). Any potential $c(W)$ is said to be *compatible* with the graph (V,E) underlying an MRF if the corresponding set V_W is a *clique* according to the adjacency E, that is, for any two distinct nodes $v \neq v'$ in the set V_W, the pair $\{v,v'\}$ is an edge in the graph (V,E) (in other words, all nodes in V_W are mutually neighbouring according to the adjacency relation in the graph). For example, a potential $c\left(T_{v_1}, T_{v_2}, T_{v_5}\right)$ acting on three variables located in nodes v_1, v_2, and v_5 would be incompatible with the graph in the left-hand part of Figure 1 because the nodes v_2 and v_5 are not adjacent in that graph (so, the triple of nodes $\{v_1,v_2,v_5\}$ is not a clique in that graph). Instead, the same potential is compatible with the graph in the right-hand part of the figure, because the triple of nodes $\{v_1,v_2,v_5\}$ is indeed a clique in this other graph.

For the construction of an MRF, a collection of potentials $C=\{c_1(W_1),\ldots,c_m(W_m)\}$ is designed. Such a collection is here called a *network of potentials*, and the collection $\{W_1,\ldots,W_m\}$ of the scopes of the potentials is called the *scheme* of the network. The network is said to be *compatible* with the hypothesized graph (V,E) if each of its potentials is compatible with the graph, that is, each scope W_i in the scheme corresponds to a clique in the graph. On the one side, the individual potentials in a network are soft local constraints, and each acts on a small (possibly singleton) set of elemental variables which are topologically close to each other according to the adjacency relation E. On the other side, the scheme $\{W_1,\ldots,W_m\}$ of the network typically covers the whole set of elemental variables, that is, the total variable T of the model. In fact, this is the way in which the MRF theory addresses the peculiar task of a field theory, which is to account for the emergence of complex and topologically extensive (possibly global) properties in terms of an organized plurality of formally simple and topologically local constraints.

The construction of an MRF comes to its end by applying two operations on potentials and their sum. First, the potentials in the network C are combined by summation, thus producing a function acting on the total variable T. This global function is called the *energy function* of the model and is here denoted by $e(T)$, so that

$$e(T)=c_1(W_1)+\ldots+c_m(W_m). \tag{1}$$

On considering that each single potential $c_i(W_i)$ evaluates how much the possible configurations of the local variable W_i differ from a specific desired condition, we will have that the energy function $e(T)$ evaluates how much the possible configurations of the total variable T differ from the system of desired conditions that have been taken into account in designing the network C, so that the energy $e(T)$ may be interpreted as an overall cost function[2]. The other operation is a transformation that changes the energy function $e(T)$ into a probability function $p(T)$, according to this rule:

$$p(T)=\exp(-e(T))/K \tag{2}$$

[2] As a rule, the potentials constituting a network involve parameters which contribute to the flexibility and generality of the model and enter in the construction of the energy function $e(T)$ (e.g., Geiger & Girosi, 1991, p. 406; Kohli, Ladický, & Torr, 2011, p. 314). For the sake of simplicity, here I do not make explicit the possible involvement of parameters and presume that the energy is the simple (non-weighted) sum of the potentials in a network. This does not impair the substance of our arguments.

where $\exp(-e(T))\approx 2.718^{-e(T)}$ (an exponential transformation involving the Neper constant 2.718…) and $K=\sum_{t\in T^\circ}\exp(-e(t))$ (so that $\sum_{t\in T^\circ}p(t)=1$, as generally required for a probability function). The most important detail in Equation (2) is the minus sign applied to the exponent $-e(T)$. It implies that the probability $p(T)$ is a *decreasing* monotone transform of the energy $e(T)$, so that searching for a configuration of T that maximizes the probability is tantamount to searching for a configuration of T that minimizes the energy. The probability function $p(T)$ thus constructed is called the *Gibbs distribution* (or also the Boltzmann distribution) implied by the hypothesized network of potentials (Gibbs, 1902; Kindermann & Snell, 1980; Richey, 2010).

For an example, let us consider a graph (V,E) of 9 nodes and 12 edges as represented in the lower part of Figure 2. Suppose that in each node v_i a pair (X_i,Y_i) of elemental variables is located, with ranges $X_i^\circ=\{-2,-1,1,2\}$ and $Y_i^\circ=\{-1,1\}$. Hence, $T=((X_1,Y_1),\ldots,(X_9,Y_9))$ is the total variable available, whose range may be specified as $T^\circ=(\{-2,-1,1,2\}\times\{-1,1\})^9$. Further suppose that, relative to each node $v_i\in V$ and each edge $\{v_i,v_j\}\in E$, definite potentials $c_i(X_i,Y_i)$ and $c_{i,j}(Y_i,Y_j)$ are introduced thus specified:

$c_i(x,y)=0$ or $|x|$ depending on whether $\mathrm{sign}(x)=$ or $\neq\mathrm{sign}(y)$,

for all $(x,y)\in\{-2,-1,1,2\}\times\{-1,1\}$

$c_{i,j}(y,y')=0$ or 1 depending on whether $\mathrm{sign}(y)=$ or $\neq\mathrm{sign}(y')$,

for all $(y,y')\in\{-1,1\}\times\{-1,1\}$.

The former expresses the possible disagreement between the X variable and the Y variable residing in one node, whereas the latter expresses the possible disagreement between the Y variables residing in two distinct nodes that are adjacent in the graph. The functions of the one and of the other type compose a network of 9+12=21 potentials:

$$C=\{c_i(X_i,Y_i)\colon v_i\in V\}\cup\{c_{i,j}(Y_i,Y_j)\colon \{v_i,v_j\}\in E\}.$$

This network is compatible with the graph in Figure 2, because each potential in the first group (node potentials) acts on variables that reside in one single node, and each potential in the second group (edge potentials) acts on variables that reside in adjacent nodes. Hence, the energy function implied by the network will be

$$e(T)=\sum\nolimits_{v_i\in V}c_i\left(X_i,Y_i\right)+\sum\nolimits_{\{v_i,v_j\}\in E}c_{i,j}\left(Y_i,Y_j\right)$$

and the corresponding Gibbs distribution will be

$$p(T)=\exp\left(-\sum\nolimits_{v_i\in V}c_i\left(X_i,Y_i\right)-\sum\nolimits_{\{v_i,v_j\}\in E}c_{i,j}\left(Y_i,Y_j\right)\right)$$

with $K=\sum_{t\in T^\circ}\exp(-e(t))$.

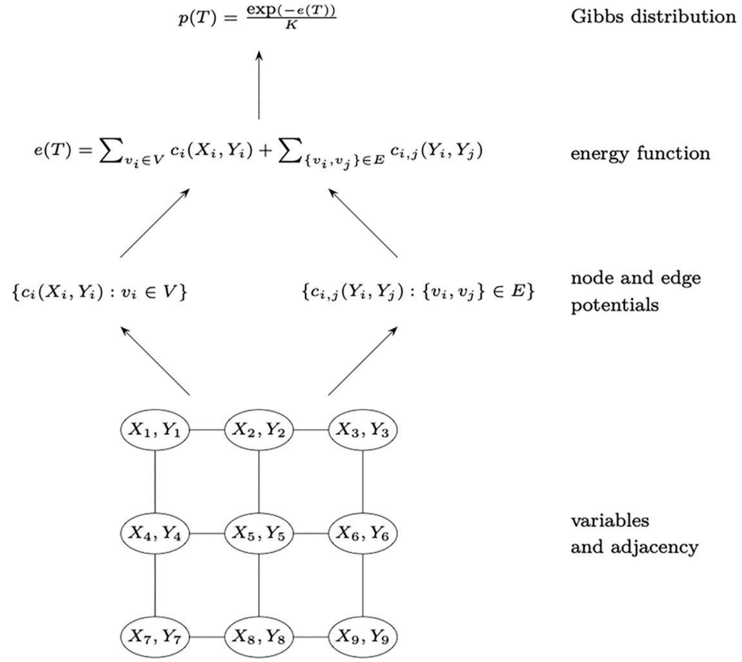

Fig. 2. A graph of 9 nodes and 12 edges, with a pair of variables (X_i, Y_i) located in each node v_i (lower part). The formation of a network of 21 potentials and of corresponding energy function and Gibbs distribution (upper part).

The diagram in the upper part of Figure 2 illustrates the combination and transformation process defined by these equations. Both $e(T)$ and $p(T)$ are non-negative valued functions acting on the range $T°$ of the total variable T. For any possible configuration t of T — that is, any point t in the space $T°$ — the corresponding energy $e(t)$ is higher the stronger the divergence of that configuration from the system of desired conditions coded in the network of potentials C. On the contrary, the resulting probability $p(t)$ is higher the weaker that divergence. Note that the toy MRF thus described is alike a small fragment of the "Ising model", which is a kind of MRF initially devised for the study of ferromagnetism, and then proved meaningful also for problems in other research areas, including questions of computer vision (Ising, 1925; Newell & Montroll, 1953; Blake & Kohli, 2011, pp. 12-13).

I conclude this preliminary section with two general comments. First, an MRF, which is constructed as a network of local potentials relative to a definite kind of real processes, may be conceived of as a system of assumptions on dependences that are at work in those processes and affect their outcomes. In this view, an MRF may serve as a guide for inferences about the prevailing states of the real processes in question. More precisely, the role of guide for inferences is played by the energy function $e(T)$ implied by the model, or indifferently by the corresponding Gibbs distribution $p(T)$. For example, once an MRF has been fully specified, so that a distribution $p(T)$ becomes determined and known, we could search for the (or some) point $\hat{t}$ in the space $T°$ at which that distribution reaches its maximum value, so that[3]

$$\hat{t}=\mathrm{argmax}_{t\in T°}p(t).$$

This result $\hat{t}$ identifies the (or some) configuration of the total variable T that proves to be maximally probable, in the light of the assumptions that concur to justify the hypothesized MRF. Because of the relationship that holds between probability and energy in Equation (2), the same result may also be expressed as

$$\hat{t}=\mathrm{argmin}_{t\in T°}e(t)$$

that is, $\hat{t}$ is a configuration of the total variable T that minimizes the energy, as this is implied by the postulates in the model. In the next section, we shall see that other inferences of this kind are also conceivable, which concern not the total variable T but selected parts of it and are supported not only by the assumptions of the model but also by information of empirical provenance.

The other general comment is about the logical role played by the graph-theoretic basis of an MRF. We stated that, once a graph (V,E) has been chosen, a network of potentials $C=\{c_1(W_1),\ldots,c_m(W_m)\}$ is judged compatible with it if, for each potential $c_i(W_i)$ in the network, the variables involved in this potential all reside in one clique of the graph, that is, they are mutual neighbours in the topology described by the adjacency E. Now, the thinner (i.e., poor in lines) an adjacency, the smaller and fewer the cliques in it tend to be. For example, the adjacency in the left-hand part of Figure 1 is thinner than that in the right-hand part, and correspondingly every clique in the former is also a clique in the latter, but not vice versa. These considerations make it clear that, on account of the request that an MRF be compatible with a preordered graph-theoretic structure, we shall have that this structure constitutes an *upper bound* for the *complexity* of admissible MRFs, so far as the complexity of a network is evaluated in term of the number of elemental variables that each of its potentials may involve: The thinner an adjacency, the smaller the cliques in it, the fewer the elemental variables that reside in any single clique, and so the less complex (in terms of number of variables) any single potential that is compatible with the given adjacency. Also this basic observation will conveniently be mentioned again in the following, when referring to the data-driven learning of the graphical structure of a model.

[3] For any real-valued function f on a domain $T°$ that, in this domain, has one point of absolute maximum, this point is symbolized as $\mathrm{argmax}_{t\in T°}f(t)$. The meaning of $\mathrm{argmin}_{t\in T°}f(t)$ is symmetrical, for a point in $T°$ (argument of f) at which the function f reaches its minimum value.

3. Markov random fields for inferences in computer vision

So far, I have outlined the MRF concept in its generality, thus building up a basic lexicon which allows us to talk about MRFs with propriety. When the concept is related to problems of vision science, as happens in computer vision, inside it further characteristics and distinctions become apparent. The specification of these characteristics and distinctions constitutes the issue of the present section. In referring to any act of (artificial) vision, I categorize this act as the performance of an *inference task*, that is, the transition from a certain set of input information, of an optical nature, to a certain set of output information, which should simulate the performance of a biological (human) act of vision. Consistent with the consolidated vocabulary of computer vision, I shall call properties of the *image* the items of input information, and properties of the *scene* the items of output information resulting from a task of vision inference. Furthermore, I shall use the term *vision episode* for any concrete trial of that task, that is, a real situation in which a suitable apparatus receives an input optical information (image) and produces a visual output information (scene) compatible with it.

The first question is about how a graph (V,E) should be determined so that it constitutes the graph-theoretic support of an MRF in computer vision. For simplicity, I will address this question only relative to vision episodes in which the optical input is a single static image, that is, episodes of static monocular artificial vision. Relative to such situations, from the available literature we may extract the following two prevailing responses to the question. One response is that the nodes in V could represent individual *pixels* (picture elements) in the image, and the edges in E could represent relationships of spatial proximity between pixels. In other words, the input optical image for a vision inference task could be filtered as a lattice of pixels, according to a preordered geometric scheme, each single pixel could be symbolized by a node in the set V, and a definite proximity relation between pixels in the image could be represented as a binary relation E on the node set V. Another conspicuous response is that the nodes in V could instead represent *regions* suitably determined within the optical image, and the component E could describe the relation of topological adjacency between those regions. In the peculiar lexicon of computer vision, single regions determined in such a way that they are optically homogeneous inside and neatly distinguishable from neighbouring regions are called "superpixels", and an exhaustive collection of such regions is called an "oversegmentation" of the image (Zhang & Ji, 2010, p. 1410). On the one side, the graphs in Figure 1, much enlarged in the number of points and lines, suggest what could be the graph-theoretic support of an MRF, when the input image is schematized as a pixel lattice. On the other side, the two parts of Figure 3 suggest how an image could be subdivided into regions (superpixels) and how the resulting subdivision could be coded as a

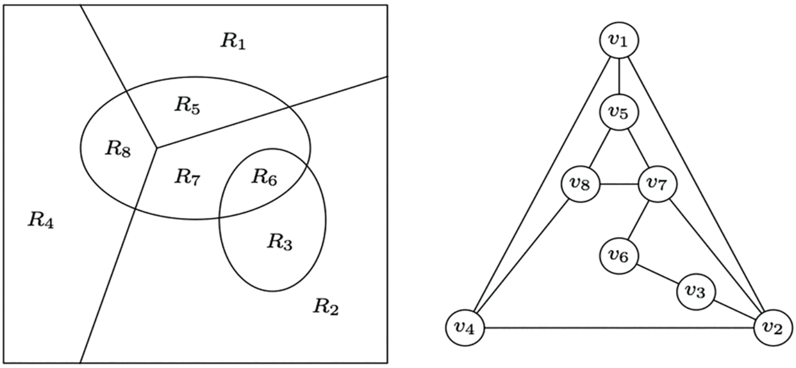

Fig. 3. Set of regions resulting from the oversegmentation of an image and corresponding planar graph, in which each node is the label of a superpixel.

planar graph. Note that, whereas a pixel lattice is determined according to a preconceived geometrical scheme (an ideal regular mask of holes) which is "blind" to the peculiar characteristics of an image, an oversegmentation into superpixels instead takes those characteristics into account, because it is required that each region be as homogeneous as possible inside and neatly distinguishable from the neighbouring regions. Thus, an oversegmentation is the result of a preliminary optical processing of a "source" image.

Regarding the variables involved in an MRF, a fundamental distinction must be made when the MRF is a model for inferences in computer vision. On the one side, there is a set X of variables each of which describes an optical property of some part of the image available in a vision episode. We refer to the elemental components of X as *input variables* and to the whole of set X as the *total input variable*. We use these terms because the optical (primitive or derivative) measures recorded in X constitute the input information for the inference to be performed in a vision episode. On the other side, there is a set of variables Y, each of which is an unknown to be solved in the inference process. In this sense, we refer to each member of Y as an *output variable* and to the whole of set Y as the *total output variable* of the MRF in question. For example, if the inference task is to decide which parts of an image should be classified as constituents of a "figure" and which parts as constituents of a "ground", then the range of any elemental variable in Y could be the pair of labels $\{f,g\}$ which represent the two alternative local choices (figure vs. ground) (Zhang & Ji, 2010, pp. 1417-1421). If instead the inference task is the construction of a "depth map" based on the image, then the range of any elemental variable in Y could be an interval $[0,d]$ of possible depth values, to be assigned to any single part or particle (superpixel or pixel) in the image (Saxena, Chung, & Ng, 2008; Hua & Tian, 2016). Thus, on the whole, the total variable of an MRF in computer vision is a system T=(X,Y)

that includes a total input variable X and a total output variable Y. For simplicity, here I suppose that the sites in which the elemental variables of the one and of the other kind are located are the same, so that the two sets may be described as $X=(X_v: v\in V)$ and $Y=(Y_v: v\in V)$, which means that to each node v in the graph an input variable X_v and an output variable Y_v are associated. In Section 5 I shall remark that other more complex organizations may be required, for the representation and guidance of vision inference tasks.

In the light of the dichotomy now defined, a distinction between two general kinds of potentials for MRFs in computer vision becomes apparent. On the one hand, we may have a potential $c(W)$ such that $W\cap X\neq\emptyset\neq W\cap Y$, that is, its scope includes both input and output variables, so that $c(W)$ can be presented as $c(W\cap X, W\cap Y)$. A potential of this kind generally imposes (in a soft manner) that the configuration to be chosen for the output variable $W\cap Y$ be in agreement with the configuration taken on by the input variable $W\cap X$ in a given image, that is, that the inference outcome be locally consistent with the available optical data. *Fidelity potentials* is among the names associated with this kind of constraints in the literature (Geman & Graffigne, 1987, p. 1505; Hochbaum, 2013, p. 170). On the other hand, we may have a potential $c(W)$ such that $W\subseteq Y$, that is, its scope only contains output variables. A potential of this kind ignores the optical information available and encourages that the values to be chosen for the output variables in W be in agreement with one another according to some desirable condition such as continuity, smoothness, connectedness, uniformity, and so on. *Interaction potentials* is among the names for this class of potentials in the computer vision literature (Wilson & Li, 2002, p. 44; Boykov & Kolmogorov, 2004, p. 1124; Kumar & Hebert, 2006, p. 185). The simplest fidelity potentials have the form $c(X_v,Y_v)$, that is, the input variable X_v and the output variable Y_v located in a single node v of the graph are considered, and a function is defined whose (non-negative) value is the larger the wider the disagreement between any value assignable to Y_v and the value taken on by X_v in an image. In turn, the simplest interaction potentials have the form $c(Y_u,Y_v)$ where $\{u,v\}\in E$, that is, the output variables Y_u and Y_v residing in two adjacent nodes of the graph are considered, and a criterion of compatibility between their possible values is formulated, regardless of the optical data available. It may be seen that the "node potentials" and the "edge potentials" constituting the toy MRF outlined in the preceding section (cf. Figure 2) are examples of the two simple types of fidelity and interaction potentials now described. Also other more complex forms are admissible, both for fidelity and for interaction potentials, granted that the general compatibility condition with the graph-theoretic basis of a network is satisfied (cf. Kohli, Ladický, & Torr, 2011; Wang, Komodakis, & Paragios, 2013, pp. 1614-1615, concerning higher-order MRF models).

On account of these distinctions, we shall have that the network of potentials for an MRF in computer vision typically is the union of a sub-network $F=\{f_1(W_1),\ldots,f_m(W_m)\}$ of fidelity potentials and a sub-network $R=\{r_1(Z_1),\ldots,r_n(Z_n)\}$ of interaction potentials, so that

$$C=F\cup R=\{f_1(W_1),\ldots,f_m(W_m),r_1(Z_1),\ldots,r_n(Z_n)\},$$

where $W_i\subseteq X\cup Y$ with $W_i\cap X\neq\emptyset\neq W_i\cap Y$ for $i=1,\ldots,m$, and $Z_j\subseteq Y$ for $j=1,\ldots,n$. According to Equation (1), the sum of all these potentials is the *energy function* $e(X,Y)$ of the MRF, so that

$$e(X,Y)=\sum_{1\leq i\leq m} f_i(W_i) + \sum_{1\leq j\leq n} r_j(Z_j).$$

This is a function on the space $X^\circ\times Y^\circ$ of the total variable (X,Y). To each point (x,y) in this space it associates a value $e(x,y)$, which measures how much the configuration (x,y) of the variable (X,Y) altogether diverges from the system of requirements (of fidelity and of interaction type) that have been postulated in designing the network $F\cup R$. Furthermore, according to Equation (2), the following exponential transform of the energy function is the *Gibbs distribution* implied by the network of potentials

$$p(X,Y)=\exp(-e(X,Y))/K.$$

This is a probability function on the space $X^\circ\times Y^\circ$. Using values in the interval [0,1], it quantifies how much any possible configuration of the total variable (X,Y) agrees with the postulated system of requirements.

The function $p(X,Y)$ is the global probability distribution of an MRF in the hypothesized context. By applying basic rules of probability theory, from it various specific functions can be derived, which are directly applicable for inferences. For example, from $p(X,Y)$, through summation, the marginal distribution $p(X)$ of the total input variable X can be deduced, so that

$$p(X)=\sum_{y\in Y^\circ} p(X,y).$$

After that, through division, from $p(X,Y)$ and $p(X)$ the family $p(Y|X)=\{p(Y|x): x\in X^\circ\}$ of conditional distributions of the total output variable Y can be deduced, so that

$$p(Y|X)=p(X,Y)/p(X)=\{p(x,Y)/p(x): x\in X^\circ\}.$$

More generally, if we are interested in the portion $Y_U\subseteq Y$ of the output variable that is located in definite part $U\subseteq V$ of the node set and the corresponding portion $X_U\subseteq X$ of the input variable, then from $p(X,Y)$ we can first derive the marginal distributions $p(X_U,Y_U)$ and $p(X_U)$ by summation, and then construct the function $p(Y_U|X_U)$ by division, so that

$$p(Y_U|X_U)=p(X_U,Y_U)/p(X_U).$$

In the theory of MRFs, other methods are also defined for obtaining such focused probability functions. These are methods that use *selected parts* of the information in a network $F\cup R$ in order to spare the ascent to the global distribution $p(X,Y)$, which is computationally expensive. So-called "conditional MRFs" are notable examples of such alternative ways (Lafferty, McCallum, & Pereira, 2001; Sutton & McCallum, 2012). In a conditional MRF it happens, for example, that relative to any configuration $x\in X^{\circ}$ of the total input variable an energy function $e(Y|x)$ is constructed (based on the network) only acting on the total output variable. The conditional distribution $p(Y|x)$ to be used in an inference is then obtained as the exponential transform of that specific energy function, in accordance with Equation (2).

Functions of this general kind constitute natural guides for inferences in computer vision. For example, suppose that, based on a definite network of potentials, the function $p(Y|X)$ has been determined, which amounts to a family $\{p(Y|x): x\in X^{\circ}\}$ of probability distributions on the range Y° of the total output variable. Then suppose that the model is applied to a particular vision episode, so that one point x' in the range X° of the input variable X is singled out, which is the configuration of X on the image available in that episode. As a consequence, from the family $p(Y|X)$ the member function $p(Y|x')$ can be extracted and, through computation, the point $\hat{y}$ in the search space Y° can be found that maximizes that function, that is

$$\hat{y}=\mathrm{argmax}_{y\in Y^{\circ}}p(y|x').$$

This would be a plausible result of the inference process. It would describe the configuration of the output variable Y that is the most probable, in the light of the optical evidence $X=x'$ and on account of the "a priori" information recorded in the network $F\cup R$. As noted, for a conditional MRF we have that the probability function $p(Y|x')$ is defined as the exponential transform of a corresponding energy function $e(Y|x')$. In this case, the inference result may alternatively be reached as the point that minimizes this energy function, which is

$$\hat{y}=\mathrm{argmin}_{y\in Y^{\circ}}e(y|x').$$

These principles and formulas may easily be generalized to situations in which the inference task concerns not the total input variable X and the total output variable Y, but a definite part of X as the source variable and a definite part of Y as the target variable. We shall return to these concepts in Section 5 for remarking on their consistency with probabilistic inferences in Bayesian statistics and, through this, with Bayesian modelling in contemporary perceptual psychology.

In Table 1, I list a sample of 30 articles in which MRF models are constructed and tested relative to inference tasks of computer vision. They cover the 40 years elapsed since the explicit appearance of this paradigm in vision science. The bibliographical details are given in the general list of references of this paper.

Table 1. Authors, publication years, and inference tasks of 30 articles from the computer vision literature with MRF models.

Authors	Year	Task
Cross & Jain	1983	texture synthesis
Geman & Geman	1984	image restoration
Geman & Graffigne	1987	texture analysis
Geiger & Girosi	1991	surface reconstruction
Kersten	1991	visual transparency
Modestino & Zhang	1992	image interpretation
Kim & Yang	1994	image segmentation
Malfait & Roose	1997	image restoration
Zhu, Wu, & Mumford	1998	texture synthesis
Wilson & Bobick	1999	gesture recognition
Zhu	1999	visual grouping
Boykov, Veksler, & Zabih	2001	stereo vision
Dass, Jain, & Lu	2002	human face detection
Wilson & Li	2002	image segmentation
Kolmogorov & Zabih	2004	multicamera vision
Kumar & Hebert	2006	image classification
Awasthi, Gagrani, & Ravindran	2007	object detection
Saxena, Chung, & Ng	2008	depth estimation
Köster, Lindgren, & Hyvärinen	2009	model learning
Nowozin & Lampert	2009	image segmentation
Ming & Hu	2010	stereo vision
Freeman & Liu	2011	texture synthesis
Kohli, Ladický, & Torr	2011	image segmentation
Winn & Shotton	2011	object detection
Panagopoulos et al.	2013	shape and illumination
Sun & Tappen	2013	image restoration
Zhou, Zhong, & Wang	2014	image segmentation
Hua & Tian	2016	depth estimation
Murray	2020	lightness effects
Ouali et al.	2024	image segmentation

4. Links between terms of MRF models and tenets of field-theoretic views of perception

In this section, I will comment on meaningful correspondences that can be recognized between the basic concepts of MRF theory as applied in computer vision and general ideas in the psychology of visual perception, with special reference to Gestalt theory within this branch of psychology. The key concepts of the MRF approach in computer vision will be considered in the same order in which they were described in the preceding section.

The construction of an MRF begins, we said, with the definition of a graph (V,E) such that the nodes in V are associated as symbols with *basic units* that are determined within an optical image (for static monocular vision). We noted that two prevailing (but not exhaustive) methodological options in the computer vision literature are those in which the basic units are pixels regularly distributed on the image plane, so that the graph symbolizes a pixel lattice, or are superpixels, that is, maximal homogeneous regions, so that the graph symbolizes an oversegmentation of the image. The first option is consonant with the cases in which, within perceptual psychology, the stimulus is described as a mosaic of optical particles (a discrete configuration), or as an image intensity function (possibly a continuous configuration), or also as a dot lattice (e.g., in experiments on perceptual grouping: Wertheimer, 2012; Bozzi, 2019; Kubovy & van den Berg, 2008). The other option agrees with situations in which the stimulus array for an experiment is constructed as a picture composed of chromatically homogeneous regions put side by side to one another. Such pictorial compositions are used, for example, in experiments on the perception of polyhedral shapes, perceptual occlusion and amodal completion, apparent transparency, or other "double representation" phenomena (Adelson, 2000; Kanizsa & Gerbino, 1982; Metelli, 1974). Pictorial compositions of this sort are here termed "map-like pictorial stimuli" due to their structural similarity with geographical maps when these are formed of several regions each of which is chromatically uniform inside and different from neighbouring regions.

In turn, the relation E between the nodes in the graph describes a *topological relation* of proximity or adjacency between the basic units symbolized by them. This topological relation is the only or the most important piece of spatial information that is recorded in the construction of an MRF in computer vision. The methodological prominence attributed to it corresponds to the importance possessed by analogous relations in well-known principles of perceptual psychology, such as the "proximity principle" as a factor of perceptual grouping, the "adjacency principle" in the study of induction effects, and "optical contact" as a factor in the spatial organization of three-dimensional perceptual scenes (Wertheimer, 2012, pp. 130-135; Gogel, 1978; Gibson, 1950, p. 178; Sedgwick, 2001, p. 137). The spatial information recorded in the relation E is of a qualitative or non-metricized character because for any two nodes in the graph we only consider whether the corresponding units in the image satisfy or do not satisfy a definite criterion of neighbourhood; we do not measure the geometric distance between the units or between salient points in them. However, a distance function becomes implicit in the graph. This is the graphical metric, according to which the distance between the units symbolized by two nodes is the length of the shortest paths in the graph that connect these nodes.

The distinction between input variables (in the system X) and output variables (in the system Y) corresponds to the distinction between properties of the *stimulus* and properties of the *percept*, which occurs as a premise in any psychological theory of visual perception. The specification of the variables both in the one and in the other system is of a relative character, as it depends on the vision inference task for which an MRF model is constructed. I shall return to this important aspect in the next section, in discussing the flexibility of the MRF paradigm in computer vision. Note the diverse epistemological statuses of the variables in the two systems. A member (or subset) of X can properly be termed a "variable" only if reference is presumed to a plurality of images, that is, a number of possible vision episodes. For any single image, and a corresponding vision episode, such a member is a specific optical datum that is available in the image, so that it is a constant. Otherwise, a member (or subset) of Y is a variable as an "unknown" also within any single vision episode. At the beginning of the episode, such a member is indeterminate within a range of possible values, and the inference task is the getting over that indeterminacy by searching for a value in the range that best satisfies a definite set of constraints. Computer vision methods use the possible configurations of the total output variable Y (or of specific parts of it) to express the results of such inference processes. For example, if an inference asks to subdivide the image into blocks corresponding to k different kinds of objects, which are labelled by symbols $o_1,\ldots,o_k$, then the set of these symbols could be adopted as the range of each of the elemental variables in Y, and the result of an inference could be expressed as an assignment of a symbol to each node in the graph, that is, to the corresponding unit in the image. Such an assignment would be a configuration of the total output variable Y (Shotton, Winn, Rother, & Criminisi, 2009). Conversely, if an inference asks to reconstruct the 3D shapes of objects, and $U \subseteq V$ is a definite connected subset of nodes in the graph corresponding to a region in the image, then the 3D shape inferred for an object could be expressed as a configuration of the variable Y_U, which is the set of elemental variables residing in the nodes of U. In this case, the range of each elemental variable could be a definite interval $[0,d]$ of possible depth values (Geiger & Girosi, 1991).

Any psychological theory of visual perception posits that the stimulus input in a vision episode has a role in the formation of the perceptual outcome in that episode. In a theory, this role could be expressed in terms of *psychophysical dependences*, that is, dependences between optical properties of the stimulus and phenomenal properties of the percept. With some exceptions (e.g., the "ecological approach" to visual perception by James J. Gibson, 1979), psychological theories also presume that the stimulus input could be insufficient to explain the uniqueness and conclusiveness of the perceptual outcome in a vision episode, so that additional factors are hypothesized for this purpose. These factors could be hard or soft constraints of compatibility or complementarity between the phenomenal properties of the percept themselves, that is, relations expressible as *intra-perceptual dependences*. The distinction stated in the preceding section between "fidelity potentials" and "interaction potentials" within MRF models of computer vision corresponds to the psychological distinction now indicated. Indeed, on considering the kinds of variables involved, a fidelity potential may be likened to a psychophysical dependence, and an interaction potential is analogous to an intra-perceptual dependence[4]. The concept of potential, as understood in the general theory of MRFs, greatly contributes to the ductility of this approach in designing models of computer vision. The approach has the task and merit of determining explicit algebraic forms for the potentials accepted in a network. Any potential is a non-negative valued function whose form depends on both the kinds of the variables involved (e.g., binary, classificatory, numerical, vector-valued, etc. variables) and the meaning of the constraint that the potential is intended to formalize. In particular, on considering the "interaction potentials" occurring in the existing MRF models, as soft constraints apt to encourage continuity, smoothness, connectedness, uniformity, etcetera, in the results of vision inference, significant analogies can be recognized with prior requirements discussed in some theories of visual perception, in particular in Gestalt theory under the heading of "tendency to Prägnanz" (Wertheimer, 2012, pp. 144-147; Rausch, 1966, pp. 904-947; Metzger, 1975). The MRF approach also has the merit of fixing a definite rule according to which the potentials in a network are to be combined into a unitary function, then used as the "objective function" in the search process by optimization. By that rule, the potentials are first combined through addition in constructing the energy function — according to Equation (1) — and then this function passes through an exponential transformation thus resulting in a probability distribution — according to Equation (2).

Typically, each individual potential in an MRF is a function acting on a small number of variables that are located in one single node or in nodes mutually adjacent in the graph. However, an entire network of potentials $F \cup R = \{f_1(W_1), \ldots, f_m(W_m), r_1(Z_1), \ldots, r_n(Z_n)\}$ usually covers the whole set (X,Y) of input and output variables that are relevant to a definite inference task. But most importantly, the scopes of the potentials in the network, although they

[4] Similar categories are found in other theories of visual perception. In particular, I mention the distinction between "constraints from optical stimuli" and "perceptual interdependencies" in constructivism (Rock, 1983, Chapter 10) and the distinction between "closeness to the optical data" and "smoothness assumption" in the regularization theory applied to vision (Poggio, Torre, & Koch, 1985; Szeliski, 2011, pp. 154-158).

are small in size, are *not* generally *disjoint* from one another, so that the scheme $\{W_1,\ldots,W_m,Z_1,\ldots,Z_n\}$ of the network is a cover but *not* a *partition* of the total variable (X,Y). Thus we may have, for example, that two output variables Y_u and Y_v are topologically distant from each other, so that they are not involved together in any potential, but there is a sequence of interaction potentials $r_{j_1}(Z_{j_1}),\ldots,r_{j_k}(Z_{j_k})$ in the network such that $Y_u \in Z_{j_1}, Y_v \in Z_{j_k}$ and $Z_{j_h} \cap Z_{j_{h+1}} \neq \varnothing$ for each h=1,…,k-1. This situation is illustrated in Figure 4, by comparing, for example, the variables Y_1 and Y_9 in the right-hand part of the figure. The same argument may be made regarding an input and an output variable involved in a model. On the whole, this signifies that, although the individual potentials express dependences that have limited spans, a network in its entirety may also imply dependences between properties that are far from one another in the topology of a model, because any chain of dependences may imply (by transitivity) a dependence between the first and last terms in it. Thus, we may have that a network is suitable to represent a dependence (of psychophysical character) between an input variable X_u and an output variable Y_v, as well as a dependence (of intra-perceptual character) between two output variables Y_u and Y_v that are far from each other in the topology implied by a given image.

This point is salient in our discussion, as it indicates how the MRF approach in computer vision resonates with the field-theoretic assumptions in perceptual psychology, in particular with the distinctive assumptions of Gestalt theory, with its emphases on the existence and analysis of actions at a distance, contextual effects, inter-local and inter-qualitative dependences, primacy of the whole, and the emergence of extensive perceptual entities by integration of stimulus data separately located (Wagemans et al., 2012; Wagemans, 2018). Note that the

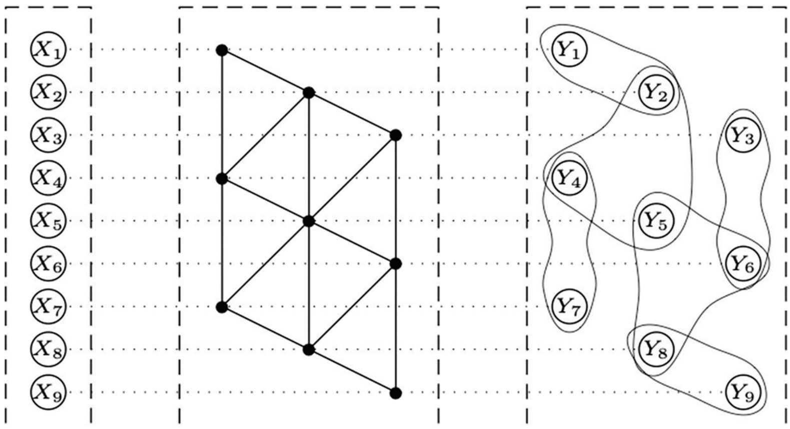

Fig. 4. Structure of an imaginary network of potentials for an MRF in computer vision, with the graph in the centre, the input variables on the left, and the output variables on the right. Contours on the right define the scopes of interaction potentials in the network.

structural aspect now highlighted implies that the MRF approach is somewhat exempt from the Gestalt-theoretic objection towards "and-summations" in explaining perceptual phenomena (Köhler, 1950, pp. 25-26; Rausch, 1966, pp. 885-888). Indeed, although the energy function, as defined by Equation (1), is constructed as the sum of potentials, these potentials are not generally independent, because some of them may overlap in their scopes. Here I presume that the Gestalt-theoretic objection is directed not so much at the use of the sum concept in itself, but at the use of that concept as applied to addends that are supposed to be totally independent of one another.

When an MRF model has been constructed for a definite inference task, by specifying a total variable (X,Y) and a network of potentials $F \cup R$, and then is applied to a particular vision episode, the MRF approach formalizes the inference task as an *optimization* problem to be solved by acting within a specific search space and under the guide of a suitable objective function. We saw that the search space could be the range Y° of the total output variable Y (or the range Y_U° of a subvariable $Y_U \subseteq Y$ of special interest), and the objective function could be the conditional distribution $p(Y|x')$ or the corresponding energy function $e(Y|x')$, where x' is the configuration taken on by the input variable X in the vision episode in question. Specifically, the inference problem is solved by finding (through computation) a point $\hat{y}$ in the space Y° that maximizes the probability $p(Y|x')$ — or, equivalently, minimizes the energy $e(Y|x')$. This general profile, that is, to cast and solve an inference problem as a problem of optimization of a quantitative criterion, is shared by various psychological theories of visual perception. A review of these theories, from this particular perspective, is that performed a while back by Hatfield and Epstein (1985). That review can consistently be updated by examining the developments of those and other theories during the forty years elapsed since then. Of course, the theories in question generally differ in the definition of the search space, the objective function subject to optimization, and the algorithms available for carrying out the optimization. However, the strategy adopted, that is, searching for "stationary" points of real-valued functions of several variables is abstractly the same.

In particular, it is well known that this general idea of perceptual performance as the optimization of a quantitative criterion was a salient feature of the Gestalt theory since its first settling as a paradigm in perceptual psychology. With statements such as "any stationary state constitutes a salient case, compared with an infinite multiplicity of other possible states" ("Jeder zeitunabhängige … Zustand … stellt gegenüber einer unendlichen Mannigfaltigkeit von anderen Zuständen einen ausgezeichneten Fall dar"; Köhler, 1920, p. 248) or "in principle, for any figure there are many possible three-dimensional (and two-dimensional) interpretations" ("prinzipiell sind bei jeder Figur auch viele räumliche (und ebene) Auffassungen möglich"; Kopfermann, 1930, p. 338), Gestalt theorists appear to argue as if, for a given stimulus configuration, there may exist several perceptual configurations abstractly compatible with it. Hence, the formation of the prevailing and stable perceptual result may be conceptualized as the search for an optimal solution, that is, a solution that best satisfies a definite network of constraints. In comparing the MRF approach with Gestalt theory we also encounter a lexical consonance, as in both contexts it is said that the search process aims at "minimizing the energy" of a system. Of course, the exact meaning of this expression depends on how the concept of "energy" is characterized, which is quite different from the one versus the other of the two theoretical contexts. However, that lexical consonance is a meaningful cue of the "family resemblance" that exists between the two theories for the reasons I specified in the Introduction.

5. Flexibility and open questions of the MRF approach to vision

An MRF model by itself — that is, as seen from the general perspective we took in Section 2 — is a mathematical construct. It is not bounded by requirements of consistency with definite real systems, as the visual system in human observers might be. Therefore, it is a construct with intrinsic freedom and may give rise to arguments on possible variations, additions, or specializations, while keeping unchanged its peculiar logical profile. The same can be said when the MRF construct is discussed as a kind of models in computer vision, which corresponds to the perspective we took in Section 3. In this case, the boundaries to be considered are those concerning the algorithms and computational means available for the application of such models, and these are mobile boundaries in expansion, as they depend on the progress of the science and technology of information. These considerations suggest that it is formally legitimate and possibly useful to reason about the intrinsic flexibilities of the MRF paradigm. Such reasoning may be instructive not only for the prospects of the MRF paradigm in computer vision, but also for its possible impact on perceptual psychology, on considering the changes that should be made in passing from the one to the other of these branches of vision science. In this section I present some thoughts in this regard.

In Section 3, I remarked that in the standard MRF models of computer vision it happens that *one single set* of basic units determined in the image (e.g., pixels or superpixels) is the set of sites in which *both* the input and the output variables of a model are presumed to be located. In symbolic terms, this assumption corresponds to the fact that one set of nodes V is the index set of both the total input variable $X=(X_v: v\in V)$ and the total output variable $Y=(Y_v: v\in V)$ of a model. Of course, this methodological premise is not in line with what one may expect from a model of perceptual psychology. Indeed, a cornerstone of this discipline is the principle that, in the perceptual processing of a stimulus array, new units may emerge that are larger than those considered in the description of the array. Furthermore, such emergent units may be partially overlapping with one another, as it happens in the various effects of "perceptual double representation", like figure-ground segregation, static or dynamic occlusion, transparency, scission of surface colours and illumination, anomalous surfaces enclosed by subjective contours, etcetera (Koffka, 1935, pp. 178-183). A way for overcoming this limitation is implicit in an example mentioned in Section 3, relative to figure-ground segregation. I suggested that, while the elemental input and output variables may be presumed to be located in the same basic units (e.g., pixels), this assumption does not preclude that other more extensive entities be recognized in modelling, for example, the entity formed of all pixels that, in an application of the model, are classified as particles of the figure, as distinct from the ground. Thus, while preserving the characteristic aspect of an MRF model, we could enrich it by allowing for the emergence and recognition of new units, which may be described as aggregates of basic minimal units postulated in the model. We should also allow that such aggregates may fail to be disjoint (in set-theoretic sense) from one another, as required by units occurring in double representation phenomena. Regarding the formal setup of the model, this would imply that its initial graph (V,E) may undergo changes during an application because new nodes could be added to the set V for symbolizing the emergent units, and new links to the relation E for describing topological relations involving those units.

We know that the graphical structure of a standard MRF is a simple graph (V,E) in which the component E is a symmetric binary relation on the node set V. As a rule, the component E describes a topological relation of proximity or adjacency between the units (in the image) that are symbolized by the nodes in V. Thus, in the definition of the graphical structure (V,E), the method that leads to an MRF is strictly selective in the spatial information to be considered: In this structure, only adjacencies between units are recorded, not numerical distances between them, nor other geometrical quantities available in the image. This methodological choice is consistent with the theoretical views in perceptual psychology that emphasize the possible advantages of describing the optical stimuli in terms of their topological rather than geometric properties (Kadar & Shaw, 2000; Chen, 2005). There are good reasons that support those views; the topological properties are generally simpler in descriptive form and ampler in invariance than the geometric ones, so that the former are conducive both to the simplicity and to the generality of a model — granted, of course, that the model be empirically valid. However, there may be inference tasks such that the topological relation of adjacency is a piece of spatial information insufficient to guide the inference successfully. For example, a model concerning the perception of transparency, surface colour, or amodal completion, besides the adjacencies between regions in a map-like pictorial stimulus, cannot ignore the shape of the junctions formed by the margins of those regions (Todorović, 1997; Kasrai & Kingdom, 2002; Khang & Zaidi, 2002; Rubin, 2001). Similarly, it is well known that the topological relation of "enclosure" between parts of a pictorial stimulus plays a key role in the phenomenon of "depth capture" (Ramachandran & Cavanagh, 1985; Kham & Blake, 2000), the topological relation of "perceptual belongingness" is critical for the apparent lightness of surfaces (Koffka, 1935, p. 137; Agostini & Galmonte, 1999, 2002), and the same may be said concerning other relations and corresponding perceptual effects. Thus, for the validity of a model, it could be required that additional pieces of the topological information in the image be considered and suitably recorded in the graphical structure. We could have, for example, that besides the symmetric relation E of adjacency, also an asymmetric relation D is constructed for representing a relation of enclosure between parts in the image, so that the graph-theoretic support of the model would be a mixed graph (V,E,D). In the next section I will specify that the computer vision literature also includes probabilistic graphical models based on such hybrid graphs.

We emphasized that any MRF model in computer vision is constructed in relation to some specific inference task, so that its input variables in X and output variables in Y are chosen in considering their relevance to that specific task. Thus, an MRF model is prepared not for simulating "the" perceptual process generally conceived — from a complete array of raw stimuli to a full perceptual scene — but for simulating a limited segment or aspect of that process. This implies a pronounced modular approach in the study and simulation of vision. This idea may be illustrated with Figure 5. In the three parts of this figure, the input and output components of imaginary models for three different inference tasks are indicated. It deserves notice, in particular, that the three models may differ not only in their output, that is, the visual properties chosen as target of inference, but also in their input, that is, the portion of optical information constituting the basis for inference, the way in which that portion is coded in the model, and its transformational distance from a hypothetical "source optical image". Actually, it is often the case that the input X of an MRF model is defined not as the set of primitive optical features of a stimulus array, but as the result of an optical transformation that is applied to an alternative (more primitive) representation of the stimuli (Zhu, Wu, & Mumford, 1998; Crouzil, Descombes, & Durou, 2003; Szeliski, 2011, Chapter 3).

In comparison with perceptual psychology, the state of affairs suggested by Figure 5 is related to the basic judgment that "stimulus definition is theory dependent" (Gordon, 1989, p. 234). Indeed, a remarkable aspect of perceptual psychology in its generality is the variety of ways in which the input information for vision episodes is operationally defined under different theoretical

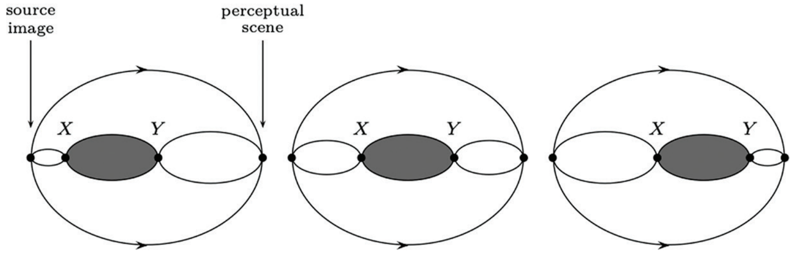

Fig. 5. Positions of input X and output Y in imaginary models for three different inference tasks.

paradigms. In any case, it is information optical in nature, but it could be specified either as a mosaic of optical particles, an image intensity function, a map-like pictorial stimulus, or else as the "ambient optic array" as defined in the ecological approach to vision (Gibson, 1979, Chapter 5). If we adopt the methodological view that is standard in computer vision, this situation gains a plausible justification. In fact, what may reasonably be expected from a psychological theory of vision is that it provides explanations of particular *segments* of the whole perceptual process. Hence, it may certainly happen that the input information is specified differently depending on the segments a theory proposes to explain, and that the input information actually hypothesized in an explanation is a transform of a more basic description of the optical data. In perceptual psychology there may be problems that are better addressed in "molecular" terms involving "lower order" stimulus variables, and other problems that are instead conveniently addressed in "molar" terms and considering "higher order" features of the stimulus (Koffka, 1935, pp. 55-58; Hochberg, 1957).

The flexibility of the MRF paradigm is especially due to the concept of potentials, which are the building blocks of a model of this kind. There are no further boundaries set in general on this concept, besides the requirements that a potential should be a non-negative valued function acting on variables that are topologically adjacent in the image, which ensures the local character of the constraint it expresses. Thus, in principle, potentials that are quite different in their algebraic form and empirical meaning may be devised as components of MRFs. In the models available in the computer vision literature, there is a prevalence of potentials suitable to express conditions of continuity, smoothness, uniformity, etcetera, in the assignment of values to the output variables of a model. As noted above, the variety of potentials of this kind resembles the variety of attributes that are headed by the Gestalt-theoretic term of "Prägnanz", which is itself a multifaceted notion. Sometimes, however, a reader may encounter potentials that are oriented otherwise. An example of this possibility is given by the "discontinuity preserving potentials" (Boykov, Veklser, & Zabih, 2001). As suggested by this expression, a potential of this kind is introduced in order to encourage the assignment of *different* values to *adjacent* output variables, to mirror a diversity that exists between the values of corresponding input variables. Formally, this is a fidelity potential $f(X_u, X_v, Y_u, Y_v)$ in which the nodes u and v represent units that are topologically adjacent in the image; X_u and X_v express optical properties possessed by those unit; whereas Y_u and Y_v are the unknown visual properties that should be inferred. Under the condition $X_u \neq X_v$, the value of the potential is the lower the larger the difference between the values assignable to Y_u and Y_v. Examples of this kind suggest that the concept of potential, because of its flexibility, could provide various sets of keys for the study of vision. In particular, it could formalize principles that are alternative to the traditional view that perceptual organization follows tendencies towards simple, regular, or uniform results (for criticism, see Kanizsa & Luccio, 1986; Kanizsa, 1994; Luccio, 2019).

Most times, an MRF in computer vision is constructed (by a modeler) in a *deductive* way, that is, by referring to general principles and plausible hypotheses and applying them to the specific inference task in question. It is remarkable, however, that (in principle) an MRF could alternatively be constructed in an *inductive* way, that is, through a learning process based on a set of data. Specifically, once a certain inference task has been specified and definite systems $X=(X_v: v \in V)$ of input variables and $Y=(Y_v: v \in V)$ of output variables relevant to it have been selected, a researcher could make an inventory $I=(I_1,\ldots,I_N)$ of images on which that task may properly be performed. Then, for each $j=1,\ldots,N$, the researcher could determine (by optical measurement) the configuration $x_j=(x_{j,v}: v \in V)$ of X on the image I_j and also determine (from judgements by human observers) the configuration $y_j=(y_{j,v}: v \in V)$ of Y on the same image. In this way, a list of paired configurations would be obtained:

$$((x_1,y_1),\ldots,(x_N,y_N))$$

which is a sample of N data on the total variable (X,Y) of a possible model. Based on this sample, and through statistical procedures, the researcher could infer a network of potentials $F \cup R$ such that the probability function implied by it — through Equations (1) and (2) — helps the researcher best reproduce the pairings between configurations of X and of Y obtained in the experiment. Such a data-driven constructive procedure would be an example of “supervised machine learning”, as methods of this kind are termed in artificial intelligence research (Bishop, 2006, p. 3). Studies on such methods are already available in the theories of MRFs and of other probabilistic graphical models (Darwiche, 2009, Chapters 17, 18; Koller & Friedman, 2009, Chapter 20; Köster, Lindgren, & Hyvärinen, 2009; Saxena, Sun, & Ng, 2009; Wang, Komodakis, & Paragios, 2013, pp. 1621-1622).

Note that in this manner the individual potentials constituting the network $F \cup R$ would be adopted not because they are suggested by some theory of visual perception but because they compose a system capable of best simulating human visual performances, specifically, the performances recorded in the experiment. In this way, even the graphical structure of an MRF could be the result of a data-driven learning process, if the component E is defined as the thinnest adjacency on V relative to which all potentials in $F \cup R$ are compatible, in the sense specified in Section 2. Furthermore, I remark that the possibility now suggested, when transferred from computer vision to perceptual psychology, may constitute a significant point in the debate concerning the relationships between perception and learning. In particular, it would constitute a regular experimental way to verify whether the intrinsic tendencies or preferences attributed to the perceptual system could be the marks of statistical regularities prevailing in the physical-behavioural environment. This is a kind of hypothesis traditionally and currently discussed in the psychological theories of visual perception (Hammond & Stewart, 2001; Geisler, 2008; Murray, 2020).

The last comments I present concerning the flexibility and suggestions of the MRF approach are about the inference procedures made possible by it, as these have been outlined at the close of Section 3. There we saw that an inference procedure could be organized by first selecting a part Y_U of the total output variable Y (Y_U becomes the target of inference) and a corresponding part X_U of the total input variable X (X_U constitutes the basis of inference), and then deriving from the network of potentials in the model the probability function $p(Y_U|X_U)$, which is a family of conditional probability distributions with X_U and Y_U as the conditioning and conditioned variables, respectively. After that, when a particular image is given, the configuration x_U' taken by X_U on that image is determined, and the corresponding member $p(Y_U|x_U')$ of the family $p(Y_U|X_U)$ is highlighted. Hence, the configuration $\hat{y}_U$ inferred for the target variable Y_U will be the point in the space Y_U° that maximizes the distribution $p(Y_U|x_U')$, that is

$$\hat{y}_U = \arg\max_{y_U \in Y_U^\circ} p(y_U|x_U').$$

This procedure is a standard example of *probabilistic inference*, as by it we highlight a property possessed by the probability function $p(Y_U|x_U')$, which is known under the presumed conditions. The procedure is in line with the Bayesian way of organizing and interpreting statistical inferences and agrees with the procedures guided by Bayesian models in present time perceptual psychology (Kersten, Mamassian, & Yuille, 2004; Rescorla, 2015; Burigana & Vicovaro, 2016). Specifically, in the Bayesian lexicon, we have that the probability function $p(Y_U|x_U')$ is the "posterior distribution" which is selected (within the model) based on the "empirical evidence" $X_U=x_U'$, and that $\hat{y}_U$ is the "maximum a posteriori (MAP) configuration" of the unknown Y_U, in the light of that empirical evidence and on the basis of the assumptions that are implicit in the network $F \cup R$. It is these assumptions, that is, soft constraints expressed as potentials, that play the role of "prior information" in a probabilistic inference thus organized.

In practice, the solution $\hat{y}_U$ to an inference problem is reached by applying some suitable algorithm, whose action gives rise to walks within the search space Y_U° in trying to find a point that maximizes the probability function, or equivalently minimizes the energy function as implied by the MRF hypothesized. There are several algorithms for this use and with these characteristics (cf. Kirkpatrick, Gelatt, & Vecchi, 1983; Szeliski et al., 2008; Wang, Komodakis, & Paragios, 2013, pp. 1616-1621). The stated characteristics may suggest significant comparisons with the hypotheses within Gestalt theory and other psychological theories of vision, regarding the dynamic processes that would be implicit in the gradual formation or discovery of "stationary" perceptual results in vision episodes (Köhler, 1920, pp. 248-250; Hatfield & Epstein, 1985, pp. 178-183).

6. Concluding remarks

I began my analysis by characterizing the concept of a field as a system of entities and local constraints acting on them, which is a key concept both in the theory of MRFs and in Gestalt theory with its hypotheses concerning field dynamics in visual perception. Now I add that the term "field" is of frequent use, both in Gestalt theory and in other psychological theories of vision, and that in these contexts it takes on various meanings, some of which have little or nothing in common with the meaning brought into focus in this study. As examples, I mention the concept of "field of a property", which equals what we termed the range of a variable, that is, the set of its possible values (Koffka, 1935, p. 255); "outfield" of a unit in a system, as a region surrounding the body of that unit such that any entity falling inside it is subject to the influence of the unit in question (Kadar & Shaw, 2000, p. 147); "receptive field" of a neural component, as the part of a sensory surface that "modulates" that component, that is, influences its status (Hartline, 1940); "perceptive field", as a psychophysical analogue to the preceding physiological concept (Ehrenstein, Spillmann, & Sarris, 2003); "visual field", as a two-dimensional structure characterized as "the simultaneous structure of visual awareness" (Koenderink, van Doorn, & Wagemans, 2015); "vector field", as the assignment of vectors to distinct points in a space. Concerning this last concept, I note that since long time it is effectively used in the psychology of visual perception for the description and analysis of various phenomena, both of a static and of a dynamic type (Orbison, 1939; Koenderink, 1986; Stadler, Richter, Pfaff, & Kruse, 1991; Fantoni & Gerbino, 2003). It differs from the field concept highlighted at the start of this study because the definition of a vector field, in its basic form, does not include the specification of a network of constraints on the component entities, which are vector-valued variables located in a space.

The MRFs constitute a subclass of the general class of probabilistic graphical models, as these are defined in probability theory and multivariate statistics. Within this general class, other subclasses have also proved to be reasonably and profitably applicable in computer vision as well as other areas of artificial intelligence research. In particular, I mention the Bayesian networks and the probabilistic models based on chain graphs[5]. As specifically concerns the graph-theoretic support of these models, we have that in a *Bayesian network* the support is a directed graph (V,D) in which D is an acyclic (hence also asymmetric) binary relation in the node set V, that is, a relation apt to represent links of precedence, inclusion, or implication between salient units in vision episodes (de Campos, Tong, & Ji, 2008). In turn, a *chain graph* is a structure (V,G) in which the relation G may be depicted as a net of lines some of which are oriented (asymmetric) and others are non-oriented (symmetric), so that it can be used to describe topological relations that are more complex than an adjacency or a precedence separately considered (Zhang, Zeng, & Ji, 2011). I also remark that among combinatorial constructs devised in studying vision there are *connectionist models*, which are related to neural networks and the "parallel distributed processing" in cognitive performances (Kienker, Sejnowski, Hinton, & Schumacher, 1986). Between these models and the MRFs of computer vision there are both significant similarities and differences. For example, the local dependences between variables that in an MRF are expressed as potentials in a network, in a connectionist model are defined as "excitatory or inhibitory interactions" between adjacent units in the structure. Significantly, within both paradigms reference is made to measures of "energy" and to procedures for minimizing it. First appearances of connectionist models relating to vision are contemporary to those of MRFs models in computer vision (mid-1980s).

In this study I only referred to MRF models in computer vision because this is the branch of vision science in which models of this kind so far have been designed and experimented. My ultimate goal, however, was to highlight aspects that may serve as cues for formal modelling in the psychology of visual perception. Some trials in this direction are already available in the psychological literature. For example, in Murray (2020) an MRF model is presented concerning the perception of lightness and illumination of surfaces, and the model in Michel and Jacobs (2007) is a Bayesian network relating to perceptual learning. Both these models are accompanied by experiments with human observers. As a concluding note, I remark again three problems that, in my view, are critical for a development in the stated direction. One concerns the concept of units in a model, that is, the entities symbolized by the nodes in the graph, which should be enlarged, so that new units may be installed (during processing) and the units carrying the

[5] Somewhat related are also the pioneer models discussed by Lehmann (1981), which combine geometric and probabilistic concepts for the analysis of optical-geometric illusions.

input variables may be (in some part) different from those carrying the output variables. A second problem concerns the topological relation (or relations) to be represented by links in the graph, which could be richer in information and more complex than the adjacency relation of a standard MRF, depending on the complexity of the vision inference task addressed. A third problem concerns the algebraic expression (in the form of potentials) of principles firmly established in the psychological study of vision, such as the principles of perceptual grouping, which are an exemplary part of the lasting legacy of Gestalt theory.

Authors have suggested that currently (or in the near future) there could be a decisive renaissance of the Gestalt-theoretic approach in perceptual psychology, a kind of neo-Gestalt theory, especially guided by the concept of field dynamics — in a psychophysical rather than physiological sense — which constituted a prominent and distinctive character of that approach since its first appearance as a paradigm for the explanation of perceptual performances (Epstein, 1988; Spillmann, 2012, p. 193; Wagemans, 2014, p. 17). Here I propose that research in computer vision that relates to the construction and experimentation of MRFs and other probabilistic graphical models may perhaps provide meaningful cues and comparisons in this regard.

Summary

This is a short tutorial on the concept of the Markov random field (MRF) as applied in computer vision and its relationships with salient terms in the psychology of visual perception, especially connected with Gestalt theory. The concepts discussed are those of graphical structure of an MRF model, variables involved in the model, potentials defined as soft constraints on neighbouring variables, energy and probability functions implied by a network of potentials, and inference procedures aiming at the minimization of energy or, equivalently, the maximization of probability. These concepts are first defined for MRFs in general, then characterized in relation to MRFs associated with vision tasks, then compared with analogous concepts in the psychology of visual perception, and finally evaluated in their flexibility and heuristic power for formal modelling in the psychological study of vision.
Keywords: Computer vision, Vision inference task, Markov random field, Field dynamics, Probabilistic inference.

Zusammenfassung

Feld Dynamiken in der visuellen Wahrnehmung und Markov Random Fields in der Computer Vision

Dies ist ein kurzes Tutorial zum Konzept des Markov Random Field (MRF), wie es in der Computer Vision angewendet wird, und seiner Bezüge zu zentralen Begriffen der Psychologie der Visuellen Wahrnehmung, insbesondere solchen im Zusammenhang mit der

Gestalttheorie. Diskutiert werden Konzepte der grafischen Struktur eines MRF-Modells, im Modell involvierte Variablen, Potenzialen als weiche Randbedingungen in Bezug auf angrenzende Variablen, Energie- und Wahrscheinlichkeitsfunktionen eines Netwerks von Potenzialen sowie Inferenzmethoden mit dem Ziel einer Energieminimierung bzw., damit gleichbedeutend, einer Wahrscheinlichkeitsmaximierung. Diese Konzepte werden zunächst für MRFs im Allgemeinen definiert, anschließend für MRFs im Zusammenhang mit der visuellen Wahrnehmung näher beschrieben, dann mit analogen Konzepten der Psychologie der visuellen Wahrnehmung verglichen und schließlich bezüglich ihrer Flexibilität und ihres Erklärungspotentials für die formale Modellierung in der psychologischen Forschung zur visuellen Wahrnehmung bewertet.
Schlüsselwörter: Computer Vision, visuelle Inferenzleistung, Markov Random Field, Feld Dynamiken, Wahrscheinlichkeitsaussagen.

References

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 339-351). Cambridge, MA: MIT Press.

Agostini, T., & Galmonte, A. (1999). Spatial articulation affects lightness. *Perception & Psychophysics, 61*, 1345-1355. doi: 10.3758/bf03206185

Agostini, T., & Galmonte, A. (2002). Perceptual organization overcomes the effects of local surround in determining simultaneous lightness contrast. *Psychological Science, 13*, 89-93. doi: 10.1111/1467-9280.00417

Awasthi, P., Gagrani, A., & Ravindran, B. (2007). Image modeling using tree structured conditional random fields. In R. Sangal, H. Mehta, & R. K. Bagga (Eds.), *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence* (pp. 2060-2065). San Francisco, CA: Morgan Kaufmann.

Besag, J. E. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B, 48*, 259-302. doi: 10.1111/j.2517-6161.1986.tb01412.x

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blake, A., & Kohli, P. (2011). Introduction to Markov random fields. In A. Blake, P. Kohli, & C. Rother (Eds.), *Markov random fields for vision and image processing* (pp. 1-28). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8579.003.0001

Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*, 1124-1137. doi: 10.1109/TPAMI.2004.60

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*, 1222-1239. doi: 10.1109/34.969114

Bozzi, P. (1969). Direzionalità e organizzazione interna della figura. *Memorie della Accademia Patavina di Scienze Lettere ed Arti, 81*, 135-170.

Bozzi, P. (2019). A new factor of perceptual grouping: demonstration in terms of pure experimental phenomenology. In I. Bianchi, & R. Davies (Eds.), *Paolo Bozzi's experimental phenomenology* (pp. 246-266). New York: Routledge. [English translation of Bozzi, 1969.]

Burigana, L., & Vicovaro, M. (2016). Inflections of the Bayesian paradigm in perceptual psychology. *Perception, 45*, 1412-1425. doi: 10.1177/0301006616669959

Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition, 12*, 553-637. doi: 10.1080/13506280444000256

Cross, G. R., & Jain, A. K. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*, 25-39. doi: 10.1109/TPAMI.1983.4767341

Crouzil, A., Descombes, X., & Durou, J. D. (2003). A multiresolution approach for shape from shading coupling deterministic and stochastic optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*, 1416-1421. doi: 10.1109/TPAMI.2003.1240116

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge, UK: Cambridge University Press.

Dass, S. C., Jain, A. K., & Lu, X. (2002). Face detection and synthesis using Markov random field models. *Proceedings of the Sixteenth International Conference on Pattern Recognition* (pp. 201-204). New York: IEEE. doi: 10.1109/ICPR.2002.1047432

De Campos, C. P., Tong, Y., & Ji, Q. (2008). Constrained maximum likelihood learning of Bayesian networks for facial action recognition. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Proceedings of the 2008 European Conference on Computer Vision*, Lecture Notes in Computer Science, Volume 5304 (pp. 168-181). Berlin: Springer. doi: 10.1007/978-3-540-88690-7_13

Ehrenstein, W. H., Spillmann, L., & Sarris, V. (2003). Gestalt issues in modern neuroscience. *Axiomathes, 13*, 433-458. doi: 10.1023/B:AXIO.0000007203.44686.aa

Epstein, W. (1988). Has the time come to rehabilitate Gestalt theory? *Psychological Research, 50*, 2-6. doi: 10.1007/BF00309403

Fantoni, C., & Gerbino, W. (2003). Contour interpolation by vector-field combination. *Journal of Vision, 3*, 281-303. doi: 10.1167/3.4.4

Fechner, G. T. (1966). *Elements of psychophysics*. New York: Holt, Rinehart and Winston

Freeman, W. T., & Liu, C. (2011). MRFs for superresolution and texture synthesis. In A. Blake, P. Kohli, & C. Rother (Eds.), *Markov random fields for vision and image processing* (pp. 155-165). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8579.003.0012

Geiger, D., & Girosi, F. (1991). Parallel and deterministic algorithms from MRF's: Surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 401-412. doi: 10.1109/34.134040

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59*, 167-192. doi: 10.1146/annurev.psych.58.110405.085632

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741. doi: 10.1109/TPAMI.1984.4767596

Geman, S., & Graffigne, C. (1987). Markov random field image models and their applications to computer vision. In A.M. Gleason (Ed.), *Proceedings of the 1986 International Congress of Mathematicians*, Volume 1 (pp. 1496-1517). Providence, RI: American Mathematical Society.

Gibbs, J.W. (1902). *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*. New York: C. Scribner's sons.

Gibson, J. J. (1950). *The perception of the visual world*. Cambridge, MA: The Riverside Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Gogel, W. C. (1978). The adjacency principle in visual perception. *Scientific American, 238*, 126-139. doi: 10.1038/scientificamerican0578-126

Gordon, I. E. (1989). *Theories of visual perception*. New York: Wiley.

Hammond, K. R., & Stewart, T. R. (Eds.) (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Hartline, H. K. (1940). The receptive fields of optic nerve fibers. *American Journal of Physiology, 130*, 690-699. doi: 10.1152/AJPLEGACY.1940.130.4.690

Hatfield, G. C., & Epstein, W. (1985). The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin, 97*, 155-186. doi: 10.1037/0033-2909.97.2.155

Hochbaum, D. S. (2013). Multi-label Markov random fields as an efficient and effective tool for image segmentation, total variations and regularization. *Numerical Mathematics: Theory, Methods, and Applications, 6*, 169-198. doi: 10.4208/nmtma.2013.mssvm09

Hochberg, J. E. (1957). Effects of the Gestalt revolution: The Cornell symposium on perception. *Psychological Review, 64*, 73-84. doi: 10.1037/h0043738

Hua, Y., & Tian, H. (2016). Depth estimation with convolutional conditional random field network. *Neurocomputing, 214*, 546-554. doi: 10.1016/j.neucom.2016.06.029

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik, 31*, 253-258. doi: 10.1007/BF02980577

Kadar, E. E., & Shaw, R. E. (2000). Toward an ecological field theory of perceptual control of locomotion. *Ecological Psychology, 12*, 141-180. doi: 10.1207/S15326969ECO1202_02

Kanizsa, G. (1994). Gestalt theory has been misinterpreted, but has also had some real conceptual difficulties. *Philosophical Psychology, 7*, 149-162. doi: 10.1080/09515089408573117

Kanizsa, G., & Gerbino, W. (1982). Amodal completion: seeing or thinking? In J. Beck (Ed.), *Organization and representation in perception* (pp. 167-190). Hillsdale, NJ: Erlbaum.

Kanizsa, G., & Luccio, R. (1986). Die Doppeldeutigkeiten der Prägnanz. *Gestalt Theory, 8*, 99-135.

Kasrai, R., & Kingdom, F. A. A. (2002). Achromatic transparency and the role of local contours. *Perception, 31*, 775-790. doi: 10.1068/p3357

Kersten, D. (1991). Transparency and the cooperative computation of scene attributes. In M. S. Landy, & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 209-228). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/2002.003.0022

Kersten, D., Mamassian, P., & Yuille, A. L. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304. doi: 10.1146/annurev.psych.55.090902.142005

Kham, K., & Blake, R. (2000). Depth capture by kinetic depth and by stereopsis. *Perception, 29*, 211-220. doi: 10.1068/p3011

Khang, B. G., & Zaidi, Q. (2002). Cues and strategies for color constancy: perceptual scission, image junctions and transformational color matching. *Vision Research, 42*, 211-226. doi: 10.1016/S0042-6989(01)00252-8

Kienker, P. K., Sejnowski, T. J., Hinton, G. E., & Schumacher, L. E. (1986). Separating figure from ground with a parallel network. *Perception, 15*, 197-216. doi: 10.1068/p150197

Kim, I. Y., & Yang, H. S. (1994). A systematic way for region-based image segmentation based on Markov random field model. *Pattern Recognition Letters, 15*, 969-976. doi: 10.1016/0167-8655(94)90028-0

Kindermann, R., & Snell, J. L. (1980). *Markov random fields and their applications*. Providence, RI: American Mathematical Society.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*, 671-680. doi: 10.1126/science.220.4598.671

Koenderink, J. J. (1986). Optic flow. *Vision Research, 26*, 161-179. doi: 10.1016/0042-6989(86)90078-7

Koenderink, J. J., van Doorn, A., & Wagemans, J. (2015). The nature of the visual field, a phenomenological analysis. *Pattern Recognition Letters, 64*, 71-79. doi: 10.1016/j.patrec.2015.02.003

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt, Brace and Company.

Köhler, W. (1920). *Die physischen Gestalten in Ruhe und im stationären Zustand. Eine naturphilosophische Untersuchung*. Braunschweig: Friedrich Vieweg und Sohn. doi: 10.1007/978-3-663-02204-6

Köhler, W. (1940). *Dynamics in psychology*. New York: Liveright.

Köhler, W. (1950). Physical Gestalten. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 17-54). New York: The Humanities Press. [Translated excerpts from Köhler, 1920.]

Kohli, P., Ladický, L., & Torr, P. H. S. (2011). Enforcing label consistency using higher-order potential. In A. Blake, P. Kohli, & C. Rother (Eds.), *Markov random fields for vision and image processing* (pp. 311-328). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8579.003.0024

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT Press.

Kolmogorov, V., & Zabih, R. (2004). What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*, 147-159. doi: 10.1109/TPAMI.2004.1262177

Kopfermann, H. (1930). Psychologische Untersuchungen über die Wirkung zweidimensionaler Darstellungen körperlicher Gebilde. *Psychologische Forschung, 13*, 293-364. doi: 10.1007/BF00406771

Köster, U., Lindgren, J. T., & Hyvärinen, A. (2009). Estimating Markov random field potentials for natural images. In T. Adali, C. Jutten, J. M. T. Romano, & A. K. Barros (Eds.), *Proceedings of the Eighth International Conference on Independent Component Analysis and Signal Separation*, Lecture Notes in Computer Science, Volume 5441 (pp. 515-522). Berlin: Springer. doi: 10.1007/978-3-642-00599-2_65

Kubovy, M., & van den Berg, M. (2008). The whole is equal to the sum of its parts: A probabilistic model of grouping by proximity and similarity in regular patterns. *Psychological Review, 115*, 131-154. doi: 10.1037/0033-295X.115.1.131

Kumar, S., & Hebert, M. (2006). Discriminative random fields. *International Journal of Computer Vision, 68*, 179-201. doi: 10.1007/s11263-006-7007-9

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley, & A. P. Danyluk (Eds.), *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 282-289). San Francisco, CA: Morgan Kaufmann.

Lehmann, G. (1981). *Figurale Wechselwirkungen im Gesichtsfeld. Experimentelle Analysen stochastischer Feldtheorien der subjektiven Figurbildung*. Göttingen: Hogrefe.

Luccio, R. (2019). Perceptual simplicity: The true role of Prägnanz and Occam. *Gestalt Theory, 41*, 263-276. doi: 10.2478/gth-2019-0024

Malfait, M., & Roose, D. (1997). Wavelet-based image denoising using a Markov random field a priori model. *IEEE Transactions on Image Processing, 6*, 549-565. doi: 10.1109/83.563320

Maxwell, J. C. (1873). *A treatise on electricity and magnetism*. Oxford, UK: Clarendon Press.

Metelli, F. (1974). The perception of transparency. *Scientific American, 230*, 90-98. doi: 10.1038/scientificamerican0474-90

Metzger, W. (1975). Die Entdeckung der Prägnanztendenz. Die Anfänge einer nicht-atomistischen Wahrnehmungslehre. In G. B. Flores D'Arcais (Ed.), *Studies in perception. Festschrift for Fabio Metelli* (pp. 3-47). Milano: Martello-Giunti.

Michel, M. M., & Jacobs, R. A. (2007). Parameter learning but not structure learning: A Bayesian network model of constraints on early perceptual learning. *Journal of Vision, 7*(1):4, 1-18. doi: 10.1167/7.1.4

Ming, Y., & Hu, Z. (2010). Modeling stereopsis via Markov random field. *Neural Computation, 22*, 2161-2191. doi: 10.1162/NECO_a_00005-Ming

Modestino, J. W., & Zhang, J. (1992). A Markov random field model-based approach to image interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14*, 606-615. doi: 10.1109/34.141552

Murray, R. F. (2020). A model of lightness perception guided by probabilistic assumptions about lighting and reflectance. *Journal of Vision, 20*(7):28, 1-22. doi: 10.1167/jov.20.7.28

Newell, G. F., & Montroll, E. W. (1953). On the theory of the Ising model of ferromagnetism. *Reviews of Modern Physics, 25*, 353-389. doi: 10.1103/RevModPhys.25.353

Nowozin, S., & Lampert, C. H. (2009). Global connectivity potentials for random field models. *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 818-825). New York: IEEE. doi: 10.1109/CVPR.2009.5206567

Orbison, W. D. (1939). Shape as a function of the vector-field. *American Journal of Psychology, 52*, 31-45. doi: 10.2307/1416658

Ouali, S., Courbot, J. B., Pierron, R., & Haeberlé, O. (2024). Bayesian image segmentation under varying blur with triplet Markov random field. *Inverse Problems, 40*, Article Number 095010. doi: 10.1088/1361-6420/ad6a34

Panagopoulos, A., Wang, C., Samaras, D., & Paragios, N. (2013). Simultaneous cast shadows, illumination and geometry inference using hypergraphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 437-449. doi: 10.1109/TPAMI.2012.110

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*, 314-319. doi: 10.1038/317314a0

Ramachandran, V. S., & Cavanagh, P. (1985). Subjective contours capture stereopsis. *Nature, 317*, 527-530. doi: 10.1038/317527a0

Rausch, E. (1966). Das Eigenschaftsproblem in der Gestalttheorie der Wahrnehmung. In W. Metzger (Ed.), *Handbuch der Psychologie*. Band 1.1. *Allgemeine Psychologie: Der Aufbau des Erkennens (Wahrnehmung und Bewusstsein)* (pp. 866-953). Göttingen: Hogrefe.

Ren, X., Fowlkes, C. C., & Malik, J. (2008). Learning probabilistic models for contour completion in natural images. *International Journal of Computer Vision, 77*, 47-63. doi: 10.1007/s11263-007-0092-6

Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (Ed.), *The Oxford handbook of philosophy of perception* (pp. 694-716). New York. Oxford University Press. doi: 10.1093/oxfordhb/9780199600472.013.010

Richey, M. (2010). The evolution of Markov Chain Monte Carlo methods. *American Mathematical Monthly, 117*, 383-413. doi: 10.4169/000298910X485923

Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.

Rubin, N. (2001). The role of junctions in surface completion and contour matching. *Perception, 30*, 339-366. doi: 10.1068/p3173

Saxena, A., Chung, S. H., & Ng, A. Y. (2008). 3-D depth reconstruction from a single still image. *International Journal of Computer Vision, 76*, 53-69. doi: 10.1007/s11263-007-0071-y

Saxena, A., Sun, M., & Ng, A. Y. (2009). Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*, 824-840. doi: 10.1109/TPAMI.2008.132

Sedgwick, H. A. (2001). Visual space perception. In E. B. Goldstein (Ed.), *Blackwell handbook of perception* (pp. 128-167). Malden, MA: Blackwell. doi: 10.1002/9780470753477.ch5

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). TextonBoost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision, 81*, 2-23. doi: 10.1007/s11263-007-0109-1

Spillmann, L. (2012). The current status of Gestalt rules in perceptual research: Psychophysics and neurophysiology. In L. Spillmann (Ed.), *On perceived motion and figural organization* (pp. 191-251). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9222.003.0008

Spitzer, F. (1971). Markov random fields and Gibbs ensembles. *American Mathematical Monthly*, 78, 142-154. doi: 10.2307/2317621

Stadler, M., Richter, P. H., Pfaff, S., & Kruse, P. (1991). Attractors and perceptual field dynamics of homogeneous stimulus areas. *Psychological Research, 53*, 102-112. doi: 10.1007/BF01371818

Sun, J., & Tappen, M. F. (2013). Separable Markov random field model and its applications in low level vision. *IEEE Transactions on Image Processing, 22*, 402-408. doi: 10.1109/TIP.2012.2208981

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning, 4*, 267-373. doi: 10.1561/2200000013

Szeliski, R. (2011). *Computer vision: algorithms and applications*. New York: Springer.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M. F., & Rother C. (2008). A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*, 1068-1080. doi: 10.1109/TPAMI.2007.70844

Todorović, D. (1997). Lightness and junctions. *Perception, 26*, 379-394. doi: 10.1068/p260379

Wagemans, J. (2014). How much of Gestalt theory has survived a century of neuroscience? In A. Geremek, M. W. Greenlee, & S. Magnussen (Eds.), *Perception beyond Gestalt: Progress in vision research* (pp. 9-21). New York: Psychology Press.

Wagemans, J. (2018). Perceptual organization. In J. T. Wixted, & J. Serences (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience*. Volume 2. *Sensation, perception, and attention* (pp. 803-872). Hoboken, NJ: Wiley. doi: 10.1002/9781119170174.epcn218

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., & van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin, 138*, 1218-1252. doi: 10.1037/a0029334

Wang, C., Komodakis, N., & Paragios, N. (2013). Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding, 117*, 1610-1627. doi: 10.1016/j.cviu.2013.07.004

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung, 4*, 301-350. doi: 10.1007/BF00410640

Wertheimer, M. (2012). Investigations on Gestalt principles. In L. Spillmann (Ed.), *On perceived motion and figural organization* (pp. 127-182). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9222.003.0006 [English translation of Wertheimer, 1923.]

Wilson, A. D., & Bobick, A. F. (1999). Parametric hidden Markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*, 884-900. doi: 10.1109/34.790429

Wilson, R., & Li, C. T. (2002). A class of discrete multiresolution random fields and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*, 42-56. doi: 10.1109/TPAMI.2003.1159945

Winn, J., & Shotton, J. (2011). Markov random fields for object detection. In A. Blake, P. Kohli, & C. Rother (Eds.), *Markov random fields for vision and image processing* (pp. 389-404). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8579.003.0030

Zhang, L., & Ji, Q. (2010). Image segmentation with a unified graphical model. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 1406-1425. doi: 10.1109/TPAMI.2009.145

Zhang, L., Zeng, Z., & Ji, Q. (2011). Probabilistic image modeling with an extended chain graph for human activity recognition and image segmentation. *IEEE Transactions on Image Processing, 20*, 2401-2413. doi: 10.1109/TIP.2011.2128332

Zhou, Z., Zhong, L., & Wang, L. (2014). Locally incremental visual cluster analysis using Markov random field. *Neurocomputing, 136*, 49-55. doi: 10.1016/j.neucom.2014.01.032

Zhu, S. C. (1999). Embedding Gestalt laws in Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*, 1170-1187. doi: 10.1109/34.809110

Zhu, S. C., & Wu, Y. N. (1999). From local features to global perception – A perspective of Gestalt psychology from Markov random field theory. *Neurocomputing, 26-27*, 939-945. doi: 10.1016/S0925-2312(99)00089-2

Zhu, S. C., Wu, Y. N., & Mumford, D. (1998). Filters, random fields and maximum entropy (FRAME). Towards a unified theory for texture modeling. *International Journal of Computer Vision, 27*, 107-126. doi: 10.1023/A:1007925832420

Luigi Burigana, born 1944, is a retired professor of Psychometrics, University of Padua, Italy. From 1972 to 2015, he gave courses in Psychometrics, Methods of Data Analysis, and Mathematical Psychology for students of psychology at the University of Padua. His current scientific interests are in epistemological issues in vision research, probabilistic theories of visual perception, and discrete mathematical models in psychology.

Address: Department of General Psychology, University of Padua, via Venezia 8, I-35131 Padova, Italy

Email: luigi.burigana@unipd.it

ORCID: 0000-0001-9898-8317