frontiers
in Artificial Intelligence

HYPOTHESIS AND THEORY
published: 01 December 2021
doi: 10.3389/frai.2021.768516

Visual Features and Their Own Optical Flow

Alessandro Betti[1], *Giuseppe Boccignone*[2], *Lapo Faggi*[1,3]*, *Marco Gori*[1,4] and *Stefano Melacci*[1]

[1]*Department of Information Engineering and Mathematics, Università degli Studi di Siena, Siena, Italy,* [2]*PHuSe Lab, Department of Computer Science, Università degli Studi di Milano, Milan, Italy,* [3]*Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy,* [4]*Universitè Côte D'Azur, Inria, CNRS, I3S, Maasai, Sophia-Antipolis, France*

Symmetries, invariances and conservation equations have always been an invaluable guide in Science to model natural phenomena through simple yet effective relations. For instance, in computer vision, translation equivariance is typically a built-in property of neural architectures that are used to solve visual tasks; networks with computational layers implementing such a property are known as Convolutional Neural Networks (CNNs). This kind of mathematical symmetry, as well as many others that have been recently studied, are typically generated by some underlying group of transformations (translations in the case of CNNs, rotations, etc.) and are particularly suitable to process highly structured data such as molecules or chemical compounds which are known to possess those specific symmetries. When dealing with video streams, common built-in equivariances are able to handle only a small fraction of the broad spectrum of transformations encoded in the visual stimulus and, therefore, the corresponding neural architectures have to resort to a huge amount of supervision in order to achieve good generalization capabilities. In the paper we formulate a theory on the development of visual features that is based on the idea that movement itself provides trajectories on which to impose consistency. We introduce the principle of Material Point Invariance which states that each visual feature is invariant with respect to the associated optical flow, so that features and corresponding velocities are an indissoluble pair. Then, we discuss the interaction of features and velocities and show that certain motion invariance traits could be regarded as a generalization of the classical concept of affordance. These analyses of feature-velocity interactions and their invariance properties leads to a *visual field theory* which expresses the dynamical constraints of motion coherence and might lead to discover the joint evolution of the visual features along with the associated optical flows.

Keywords: affordance, convolutional neural networks, feature flow, motion invariance, optical flow, transport equation

OPEN ACCESS

Edited by:
Andrea Tacchetti,
DeepMind Technologies Limited,
United Kingdom

Reviewed by:
Raffaello Camoriano,
Italian Institute of Technology (IIT), Italy
Chiyuan Zhang,
Google, United States

***Correspondence:**
Lapo Faggi
lapo.faggi@unifi.it

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *31 August 2021*
Accepted: *25 October 2021*
Published: *01 December 2021*

Citation:
Betti A, Boccignone G, Faggi L, Gori M and Melacci S (2021) Visual Features and Their Own Optical Flow. Front. Artif. Intell. 4:768516. doi: 10.3389/frai.2021.768516

1 INTRODUCTION

Deep learning has revolutionized computer vision and visual perception. Amongst others, the great representational power of convolutional neural networks and the elegance and efficiency of Backpropagation have played a crucial role (Krizhevsky et al., 2012). By and large, there is a strong scientific recognition of their capabilities, which is very well deserved. However, an important but often overlooked aspect is that natural images are swamped by nuisance factors such as lighting, viewpoint, part deformation and background. This makes the overall recognition problem much more difficult (Lee and Soatto, 2011; Anselmi et al., 2016). Typical CNNs architectures, not structurally modelling these possible variations, require a large amount of data with high variability to gain satisfying generalization skills. Some recent works have addressed this aspect focusing on the construction of invariant (Gens and Domingos, 2014; Anselmi et al., 2016) or equivariant (Cohen and Welling, 2016) features with respect to a priori specified symmetry groups of transformations. We argue that, when relying on massively supervised learning, we have been working on a problem that is—from a computational point of view—remarkably different and likely more difficult with respect to the one offered by Nature, where motion is in fact in charge for generating visual information. Motion is what offers us an object in all its poses. Classic translation, scale, and rotation invariances can clearly be obtained by appropriate movements of a given object (Betti et al., 2020). However, the experimentation of visual interaction due to motion goes well beyond the need for these invariances and it includes the object deformation, as well as its obstruction. Could not be the case that motion is in fact nearly all we need for learning to see? Current deep learning approaches based on supervised images mostly neglect the crucial role of temporal coherence, ending up into problems where the extraction of visual concepts can only be based on spatial regularities. Temporal coherence plays a fundamental role in extracting meaningful visual features (Mobahi et al., 2009; Zou et al., 2011; Wang and Gupta, 2015; Pan et al., 2016; Redondo-Cabrera and Lopez-Sastre, 2019) and, more specifically, when dealing with video-based tasks, such as video compression (Bhaskaran and Konstantinides, 1997). Some of these video-oriented works are specifically focused on disentangling content features (constant within the selected video clip) from pose and motion features (that codify information varying over time) (Denton and Birodkar, 2017; Villegas et al., 2017; Hsieh et al., 2018; Tulyakov et al., 2018; Wang et al., 2020). The problem of learning high-level features consistent with the way objects move was also faced in Pathak et al. (2017) in the context of unsupervised object foreground versus background segmentation.

In this vein, we claim that feature learning arises mostly from motion invariance principles that turn out to be fundamental for detecting the object identity as well as characterizing interactions between features themselves. To understand that, let us start considering a moving object in a given visual scene. The object can be thought of as made up of different material points, each one with its own identity that does not change during the object motion. Consequently, the identity of the corresponding pixels has also to remain constant along their apparent motion on the retina. We will express this idea in terms of feature fields (i.e. functions of the given pixel and the specific time instant) that are invariant along the trajectories defined by their *conjugate* velocity fields, extending, in turn, the classical brightness invariance principle for the optical flow estimation (Horn and Schunck, 1981). Visual features and the corresponding optical flow fields make up an indissoluble pair linked by the motion invariance condition that drives the entire learning process. Each change in the visual features affects the associated velocity fields and vice versa. From a biological standpoint, recent studies have suggested that the ventral and dorsal pathways may not be as independent as originally thought (Milner, 2017). Following this insight, we endorse the joint discovery of visual features and the related optical flows, pairing their learning through a motion invariance constraint. Motion information does not only confer object identity, but also its affordance. As defined by Gibson in his seminal work (Gibson, 1966, 1979), affordances essentially characterize the relation between an agent and its environment and, given a certain object, correspond to the possible actions that can be executed upon it. A chair, for example, offers the affordance of seating a human being, but it can have other potential uses. In other words, the way an agent interacts with a particular object is what defines its affordance, and this is strictly related to their relative motion. Extending and generalizing this classic notion of affordance to visual features, we will define the notion of affordance field, describing the interaction between pairs of visual features. Essentially, these interactions are defined by the relative motion of the features themselves so that the corresponding affordance fields will be required to be invariant with respect to such relative motion. Hence, in the rest of the paper, we will use the term affordance in this broader sense.

This paper is organized as follows. **Section 2** is focused on classical methods for optical flow estimation. In this case, the brightness is given by the input video and the goal is to determine the corresponding optical flow through the *brightness invariance condition*. Typical regularization issues, necessary to specify a unique velocity field, are also addressed. **Section 3** is devoted to extend the previous approach to visual features. This time, features are not given in advance but are jointly learnt together with the corresponding velocity fields. Features and velocities are tied by the motion invariance principle. After that, the classical notion of affordance by Gibson (1966), Gibson (1979) is introduced and extended to the case of visual features. Even in this case motion invariance (with respect to relative velocities) plays a pivotal role in defining the corresponding affordance fields. At the end of **Section 3** regularization issues are also considered and a formulation of learning of the visual fields is sketched out, together with the description of a possible practical implementation of the proposed ideas through deep neural networks. Finally, **Section 4** draws some conclusions.

2 OPTICAL FLOW

The fundamental problem of optical flow estimation has been receiving a lot of attention in computer vision. In spite of the growing evidence of performance improvement (Fischer et al., 2015; Ilg et al., 2017; Zhai et al., 2021), an in-depth analysis on the precise definition of the velocity to be attributed to each pixel is still questionable (Verri, 1987; Verri and Poggio, 1989; Aubert and Kornprobst, 2006). While a simple visual inspection of some recent top level optical flow estimation systems clearly indicates remarkable performance, the definition of “optical flow” is difficult and quite slippery. Basically, we need to associate each

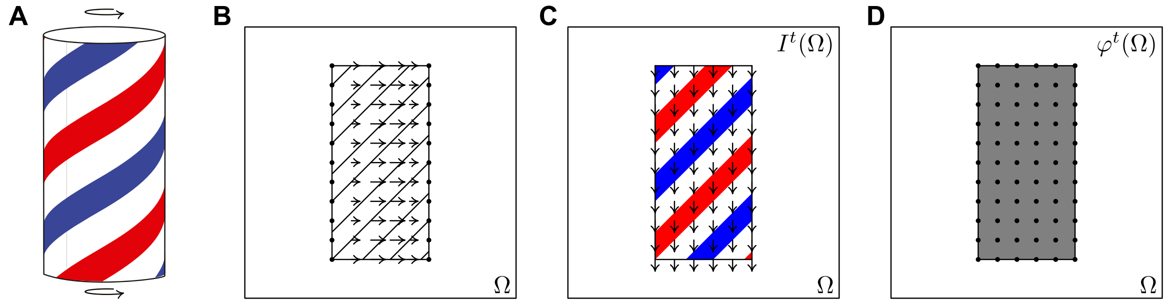

FIGURE 1 | Barber's pole example. **(A)** The 3-D object spinning counterclockwise. **(B)** The 2-D projection of the pole and the projected velocity on the retina Ω. **(C)** The brightness of the image and its optical flow pointing downwards. **(D)** A feature map that responds to the object and its conjugate (zero) optical flow.

pixel with its velocity. When considering the temporal quantization, any sound definition of such a velocity does require to know where any pixel moves on the next frame. How can we trace any single pixel? Clearly, any such pixel corresponds to a "point" of an "object" in the visual environment and the fundamental requirement to fulfill is that of tracking the point of the object.

An enlightening answer to this question was given by Horn and Shunck in a seminal paper published at the beginning of the eighties (Horn and Schunck, 1981). The basic assumption is the local-in-time constancy of *the brightness* intensity function I: $\Omega \times [0, T] \to [0, 1]$ where Ω is a subset of $\mathbb{R}^2$. In other words, $\forall t_0 > 0$ there exists a $\tau > 0$ such that for every $x_0 \in \Omega$ we can define the trajectory $\gamma_{x_0}: [t_0, t_0 + \tau] \to \Omega$ that maps $t \mapsto \gamma_{x_0}(t) \in \Omega$ for which

$$\begin{aligned} I\big(\gamma_{x_0}(t), t\big) &= I(x_0, t_0), \qquad \forall t \in [t_0, t_0 + \tau]; \\ \big(\gamma_{x_0}(t_0), t_0\big) &= (x_0, t_0). \end{aligned} \tag{1}$$

Assuming smoothness we can approximate this condition to the first order taking into account only infinitesimal temporal distances and obtain at $t = t_0$:

$$\partial_t I(x_0, t_0) + u(x_0, t_0) \cdot \nabla I(x_0, t_0) = 0, \tag{2}$$

where $u(x_0, t_0) := (d\gamma_{x_0}/dt)(t_0)$ is the optical flow and $\cdot$ is the standard scalar product in $\mathbb{R}^2$.

Assumption **Eq. 1** is reasonable when there are no occlusions and changes of the light source are assumed to be "small." Of course, in real world applications of computer vision these scenarios are not always met. On the other hand, it is clear that the optical flow u could be derived from an invariance condition of the type **Eq. 1** applied to different and possibly more "stable" visual features rather than to the brightness itself. As shown in **Figure 1**, this would give a different optical flow with respect to the one defined through the brightness invariance condition (**Figure 1C**). For example, a feature responding to the entire barber's pole, that is standing still, would have an associated optical flow that is null everywhere (**Figure 1D**). Still, we have to keep in mind that in both cases the resulting optical flow is different from the 2-D motion field (defined as the projection on the image plane of the 3-D velocity of the visual scene, see e.g. Aubert and Kornprobst (2006)) shown in **Figure 1B**.

This indeed is the main motivation to couple the problem of feature extraction together with motion invariance constraints and the derivation of robust and meaningful optical flows associated to those visual features.

2.1 Regularization of the Optical Flow

Before going on to lay out the theory for the extraction of motion invariant visual features, we need to recall some facts about the optical flow condition **Eq. 2**. Given a video stream described by its brightness intensity I as it is defined as in **Section 1**, the problem of finding for each pixel of the frame spatial support at each time instant the velocity field $u(x, t)$ satisfying

$$\partial_t I(x, t) + u(x, t) \cdot \nabla I(x, t) = 0 \quad \forall (x, t) \in \Omega \times [0, T] \tag{3}$$

is clearly ill posed since a scalar equation is not sufficient to properly constrain the two components of u. Locally, we can unequivocally determine only the component of u along ∇I.

Although many methods have been proposed to overcome this issue (see for example the work of Aubert and Kornprobst (2006)), that is usually referred to as the *aperture problem*, here we are interested in the class of approaches that aims at regularizing the optical flow:

$$\inf_v \left(A_I(v) + S(v) + H_I(v)\right) \tag{4}$$

where A_I is a functional that enforces constraint **Eq. 3**, hence a standard choice is

$$A_I(v) := \int_\Omega \left(\partial_t I(x, t) + v \cdot \nabla I(x, t)\right)^2 dx, \tag{5}$$

S imposes smoothness and H may be used to condition the extraction of the flow over spatially homogeneous regions. Depending on the regularity assumptions on I and the properties that we want to impose on the solution of this regularized problem (namely if we want to admit solutions that preserve discontinuities or not) the exact form of S and the form and presence of H_I may vary. For example, in the original approach proposed in Horn and Schunck (1981) we find:

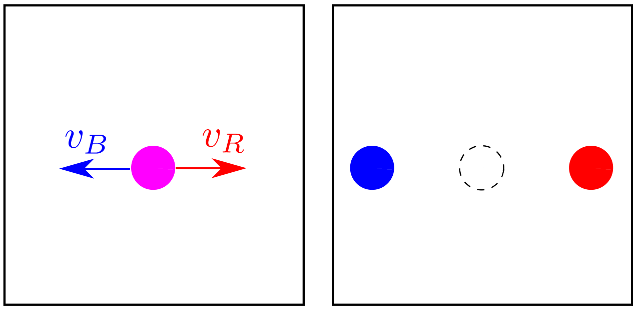

FIGURE 2 | Tracking of different color components in a synthetic example. In this case, each color component is associated to a specific velocity field.

Horn–Schunck regularization:

$$S(v) = \int_\Omega |\nabla v^1|^2 + |\nabla v^2|^2 dx, \quad H_I \equiv 0. \tag{6}$$

Note that since we are interested in extracting the optical flow for any frame of a video, namely the field $u(x, t)$, it is useful for any function $f\colon \Omega \times [0,T] \to \mathbb{R}^p$ to define $f^t\colon \Omega \to \mathbb{R}^p$ as $f^t(x) := f(x, t)$. With this notation, when the infimum in **Eq. 4** is attained, we can pose.

$$u(x,t) = u^t(x) = \arg\min_v (A_I(v) + S(v) + H_I(v)). \tag{7}$$

Notice that the brightness might not necessarily be the ideal signal to track. Since the brightness can be expressed as a weighed average of the red R, green G, and blue B components, one could think of tracking each single color component of the video signal by using the same invariance principle stated by **Eq. 3**. It could in fact be the case that one or more of the components R, G, B are more invariant in the sense of **Eq. 3** during the motion of the corresponding material point, see **Figure 2**. In that case, in general, each color can be associated with corresponding optical flows v_R, v_G, v_B that might differ. In doing so, instead of tracking the brightness, one can track the single colors. Instead, under the assumption that each color component has the same optical flow $v = v_R = v_G = v_B$ we have.

$$\frac{\partial}{\partial t}\begin{pmatrix} R \\ G \\ B \end{pmatrix} + v \cdot \nabla \begin{pmatrix} R \\ G \\ B \end{pmatrix} = 0, \tag{8}$$

where $v \cdot \nabla(R,G,B)' := (v \cdot \nabla R, v \cdot \nabla G, v \cdot \nabla B)'$. It is worth mentioning that the simultaneous tracking of different channels might contribute to a better positioning of the problem since, in general, $\mathrm{rank}\nabla(R, G, B)' = 2$ and the system **Eq. 8** admits an unique solution.[1] One can think of the color components as features that, unlike classical convolutional spatial features, are temporal features.

Before proceeding further, let us underline that some other optical flow methods try to directly solve the brightness invariance condition **Eq. 1** without differentiating it. This is the case, for example, of the Gunnar Farnebäck's algorithm (Farnebäck, 2003): the basic idea here is to approximate the brightness of the input images through polynomial expansions with variable coefficients, and the brightness invariance condition **Eq. 1** is then solved under this assumption. **Figure 3** shows the optical flows extracted by the Horn-Schunck and Farnebäcks methods in the barber's pole case.

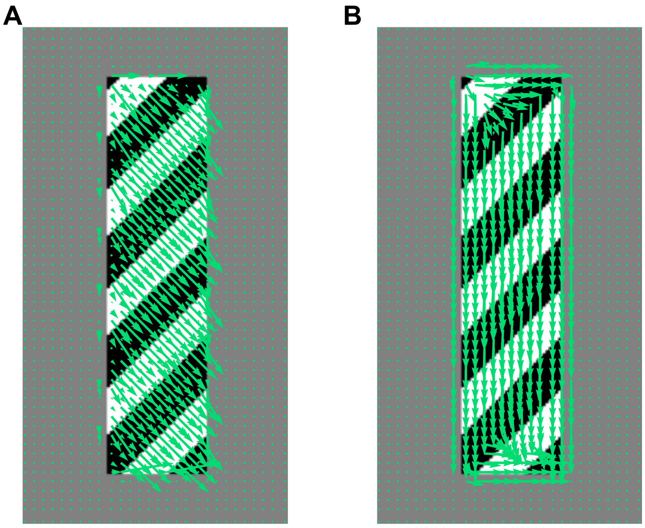

FIGURE 3 | Barber's pole optical flow (sub-sampled). **(A)** Horn–Schunck method (Horn and Schunck, 1981) with smoothing factor coefficient = 1 **(B)** Gunnar Farnebäck's algorithm (Farnebäck, 2003) in the quadratic expansion case.

In the next section, we will discuss how to use a very similar approach, based on the consistency of features along apparent motion trajectories on the frame spatial support, to derive visual features φ_i along with the corresponding *conjugate* optical flows v_{φ_i}. We anticipate that this motion consistency condition will also play a prominent role in defining affordance features, as described in **Section 3.2**.

3 FEATURE EXTRACTION AND CONJUGATE VELOCITIES

As we have already anticipated in the previous sections, the optical flow extracted by imposing an invariance condition like the one in **Eq. 3** strongly depends on the features on which we are imposing that invariance; hence it should be not surprising that different sets of features could give rise to different optical flows. This can be easily understood by considering the barber's pole example in **Figure 1**. The related classical optical flow is depicted in **Figure 1C**, see also **Figure 3**, and it is different from the projection of the 3-D velocities on the frame spatial support **Figure 1B** (the resulting optical flow is an optical illusion indeed). Let us now assume the existence of a visual feature φ_r characterizing the red stripes, that is $\varphi_r(x, t) = 1$ iff (x, t) is inside a stripe. As the barber's pole rotates, the conjugate velocity v_{φ_r} is, in this simplified case, the same that one would have obtained from the brightness invariance condition **Eq. 3**. An

[1] Under the assumption $\mathrm{rank}\nabla(R, G, B)' = \mathrm{rank}(\nabla(R, G, B)' | -\partial(R, G, B)'/\partial t)$.

additional level of abstraction can be gained when looking at the whole object. Again, we are assuming the existence of an higher level visual feature φ_{object} characterizing it. Then, considering that the barber's pole is standing still, the velocity field associated to that feature is everywhere null, as shown in **Figure 1D**.

This example clearly explains how different velocity fields can be associated to different visual features, but we still have to go one step further. Until now, mimicking the case of the classical optical flow estimation given the corresponding input brightness, we have described the construction of velocity fields starting from visual features whose existence was a priori assumed. Recent studies have suggested that the ventral and dorsal pathways may not be as independent as originally thought. Evidence for contributions from ventral stream systems to the dorsal stream indicates a crucial role in mediating complex and flexible visuomotor skills. Meanwhile, complementary evidence points to a role for posterior dorsal-stream visual analysis in certain aspects of 3-D perceptual function in the ventral stream (but see Milner, 2017 for a review). As pointed out by Milner (2017) potential cross-stream interactions might take three forms:

1) Independent processing: computations along the separate pathways proceed independently and in parallel and reintegrate at some final stage of processing within a shared target brain region; this might be achieved via common projections to the lateral prefrontal cortex or superior temporal sulcus (STS);
2) Feedback: processing along the two pathways is modulated by the existence of feedback loops which transmit information from downstream brain regions, including information processed along the complementary stream; feedback is likely to involve projections to early retinotopic cortical areas.
3) Continuous cross-talk: information is transferred at multiple stages and locations along the two pathways.

The three forms need not be mutually exclusive and a resolution of the problems of visual integration might involve a combination of such possibilities (Milner, 2017).

Yet, from a learning standpoint, the cross-talk mode is intriguing for setting some minimal conditions for an agent (either biological or artificial) in order to develop visual capabilities. Following this biological insight, we endorse the indissoluble conjunction of features and velocities and, consequently, their joint discovery based on the motion invariance condition.

$$\begin{aligned}&\partial_t \varphi_i(x,t) + v_{\varphi_i}(x,t)\cdot\nabla\varphi_i(x,t) = 0\\ &\forall (x,t)\in\Omega\times[0,T], \forall i = 1,\ldots,d,\end{aligned} \tag{9}$$

where we are considering d different visual features. Locally, this equation means that, at each pixel x of the frame spatial support and specific time instant t, features φ_i are preserved along the trajectories defined by the corresponding velocity fields v_{φ_i} and starting at (x, t). An object clearly does not change its identity while it is moving. Consequently, the identity of the corresponding pixels on the frame spatial support has to remain invariant along the apparent motion defined by the associated optical flows. Thinking of the brightness as the simplest visual feature based on single pixels, **Eq. 9** correctly reduces to the brightness invariance condition **Eq. 3**. Notice that if there is no optical flow for a given pixel $\bar{x}$, that is, if $v_\varphi(\bar{x},t) = 0$ for all $t \in [0, T]$, then $\varphi_t(\bar{x},t) = 0$. This means that the absence of the optical flow in $\bar{x}$ results into $\varphi(\bar{x},t) = c_\varphi$ for all $t \in [0, T]$, which is the obvious consistency condition that one expects in this case. Likewise, a constant field $\varphi(x, t)$ in a subregion $C \subset \Omega \times [0, T]$ makes **Eq. 9** satisfied on C independently of v_φ.

Like for the brightness, in general, the invariance condition **Eq. 9** generates an ill-posed problem. In particular, when the moving object has a uniform color, we can notice that brightness invariance holds for virtually infinite trajectories. Likewise, any of the features φ is expected to be spatially smooth and nearly constant in small portions of the frame spatial support, and this restores the ill-posedness of the classical problem of determining the optical flow that has been addressed in the previous section. Unlike brightness invariance, in the case of visual features the ill-posedness of the problem has a double face. Just like in the classic case of estimating the optical flow, v_φ is not uniquely defined (the aperture problem). On top of that, now the corresponding feature φ is not uniquely defined, too. We will address regularization issues in **Section 3.3** where, including additional information other than coherence on motion trajectories, we will make the learning process well-posed. Of course, the regularization process will also involve a term similar to the one invoked for the optical flow v, see **Eq. 6**, that will be imposed on v_φ. Given what we have discussed so far, we can also expect the presence of some regularization term concerning the features themselves and their regularity. Finally, these terms will be complemented with an additional "prediction" index necessary to avoid trivial features' solutions (we postpone its description to **Section 3.3**).

The basic notion at the core of this section is that $\{(\varphi_i, v_{\varphi_i})\}$ can be treated as indissoluble pairs bounded by the motion invariance condition that steers the entire learning process. The structure of each φ_i affects the associated velocity v_{φ_i} and vice versa—it is therefore natural to pair their learning. Leaving aside for the moment regularization issues of **Section 3.3**, learning is based on a functional generalizing to **Eq. 5**, that is

$$A\big(\{\varphi_i, v_{\varphi_i}\}\big) := \frac{1}{2}\sum_i \int_\Gamma \big(\partial_t\varphi_i(x,t) + v_{\varphi_i}(x,t)\cdot\nabla\varphi_i(x,t)\big)^2 d\mu(x,t), \tag{10}$$

where $\Gamma = \Omega \times [0, T]$ and μ is an appropriately weighted Lebesgue measure on $\mathbb{R}^2_x \times \mathbb{R}_t$; its exact form defines the dynamics of the learning process itself. The minimization of such functional (plus the additional regularitation terms) is expected to return the pairs $(\varphi_i, v_{\varphi_i})$ satisfying the motion consistency condition **Eq. 9**. Sometimes, in what follows and when the notation is clear from the context, we will drop the subscript φ of v_φ so that v_{φ_i} will be denoted as v_i.

3.1 Feature Grouping

As already noticed, when we consider color images, what is done in the case of brightness invariance can be applied to the separated components R,G,B. Interestingly, for a material point of a certain color, given by a mixture of the three components, we can establish the same brightness invariance principle, since those components move with the same velocity. Said in other words, there could be group of different visual features φ_i, $i = 1, \ldots, m$ that share the same velocity ($v_i = v$) and are consistent with it, that is $\partial_t \varphi_i + v \cdot \nabla \varphi_i = 0 \; \forall i = 1, \ldots, m$. Thus, we can promptly see that any feature φ of span($\varphi_1, \ldots, \varphi_m$) is still conjugated with v; we can think of span($\varphi_1, \ldots, \varphi_m$) as a functional space conjugated with v.

Let us now consider the feature group $\phi = (\varphi_1, \ldots, \varphi_m)'$ and the corresponding invariance condition.

$$\partial_t \phi + v \cdot \nabla \phi = 0, \tag{11}$$

where $\nabla \phi \in \mathbb{R}^{m\times 2}$ is the matrix with elements $(\nabla \phi)_{ij} = (\nabla \phi_i)_j$ and $v \cdot \nabla \phi := (v \cdot \varphi_1, \ldots, v \cdot \varphi_m)'$. An important observation, very related to the discussion about color tracking of **Eq. 8**, is the following one. Notice that, if we consider the case in which the only scalar feature we are dealing with is the brightness, then **Eq. 11** boils down to a single equation with two unknowns (the velocity components). Differently, in the case of the feature group ϕ, we have m equations and still two unknowns. The dimension m of matrix $\nabla \phi$ can enforce the increment of its rank, which leads to a better posedness of the problem of estimating the optical flow v. Because of the two-dimensional structure of the frame spatial support, which leads to $v \in \mathbb{R}^2$, and since $\nabla \phi \in \mathbb{R}^{m\times 2}$, with $m \geq 2$, it turns out that feature grouping regularizes the velocity discovery. In order to understand the effect of feature grouping we can in fact simply notice that, under the assumption $\text{rank}\nabla\phi = \text{rank}(\nabla\phi | -\phi_t)$, a random choice of the features yields $\text{rank}\nabla\phi = 2$. As a consequence, by Rouché-Capelli theorem, linear **Eq. 11** admits a unique solution in v. However, this regularization effect of feature grouping does not prevent ill-posedness, since ϕ is far from being a random map. On the opposite, it is supposed to extract a uniform value in portions of the frame spatial support that are characterized by the same feature. Hence, $\text{rank}\nabla\phi = 1$ is still possible whenever the features of the group are somewhat dependent.

Feature groups, that are characterized by their common velocity, can give rise to more structured features belonging to the same group. This can promptly be understood when we go beyond linear spaces and consider for a set of indices $\mathcal{F}$.

$$\begin{cases} \alpha = \sum_{j\in\mathcal{F}} w_j \varphi_j \\ \eta = \sigma(\alpha). \end{cases} \tag{12}$$

Evaluating $\partial_t \eta + v \cdot \nabla \eta$ we obtain indeed

$$\partial_t \eta + v \cdot \nabla \eta = \sigma'(\alpha)(\partial_t \alpha + v \cdot \nabla \alpha) = \sigma'(\alpha) \sum_{j\in\mathcal{F}} \omega_j \big(\partial_t \varphi_j + v \cdot \nabla \varphi_j\big). \tag{13}$$

and we conclude that if $\forall j \in \mathcal{F}$ we have $\partial_t \varphi_j + v \cdot \nabla \varphi_j = 0$ then also the feature η defined by **Eq. 12** is conjugated with v, that is $\partial_t \eta + v \cdot \nabla \eta = 0$. However, the vice versa does not hold true. Basically, the inheritance of conjugation with v holds in the direction towards more abstract features. Of course, the feedforward-like recursive application of the derivation stated by **Eq. 12** yields a feature that is still conjugated with v.

3.2 Affordance-Related Features

Any learning process that relies on the motion of a given object can only aspire to discover the identity of that object, along with its characterizing visual features such as pose and shape. The motion invariance process is in fact centered around the object itself and, as such, it does reveal its own features in all possible expositions that are gained during motion. Humans, and likely most animals, also conquer a truly different understanding of visual scenes that goes beyond the conceptualization with single object identities. In the early Sixties, James J. Gibson coined the notion of affordance in (Gibson, 1966), even though a more refined analysis came later in (Gibson, 1979). In his own words: *"The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, the noun affordance is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment."* Considering this animal-centric view, we gain the understanding that affordance can be interpreted as what characterizes the "interaction" between animals and their surrounding environment. In more general terms, the way an agent interacts with a particular object is what defines its affordance, and this is strictly related to their relative motion. In the last decades, computer scientists have also being working on this general idea, trying to quantitatively implement it in the fields of computer vision and robotics (Ardón et al., 2020; Hassanin et al., 2021). As far as visual affordance is concerned, that is, extracting affordance information from still images and videos, different cognitive tasks have been considered so far, as for example affordance recognition and affordance segmentation, see (Hassanin et al., 2021) for a recent review.

In the spirit of the previous section, we will consider a more abstract notion of affordance, characterizing the interaction between different visual features along with their corresponding conjugate velocity fields. We will focus our attention on actions that are perceivable from single pictures[2] and on the related local notion of affordance, that will be defined by some function characterizing the interaction between feature φ_j and feature φ_i when considering the pixel x at the specific time instant t. As we will see, the principle of motion invariance can be extended to naturally define (explicitly or implicitly) this generalized notion of affordance. A natural choice is to consider what we will denote as the *affordance field* ψ_{ij} as a function of space and time. To implicitly codify the interaction between features i and j, $\psi_{ij}(x, t)$ has to be constrained by some relation of the form $g(\psi_{ij}, \partial_t \psi_{ij}, \nabla \psi_{ij}, (\varphi_i, v_i), (\varphi_j, v_j)) = 0$, where we are considering only first order derivatives of the affordance field and g is a scalar function. In the lower order approximation (we also need quadratic terms to build scalars from vectors):

$$\begin{aligned} & g\big(\psi_{ij}, \partial_t \psi_{ij}, \nabla \psi_{ij}, (\varphi_i, v_i), (\varphi_j, v_j)\big) \\ & \quad = a_1 + a_2 \psi_{ij} + a_3 \varphi_i + a_4 \varphi_j + a_5 \partial_t \psi_{ij} + a_6 v_i \nabla \psi_{ij} + a_7 v_j \nabla \psi_{ij}, \end{aligned} \tag{14}$$

[2] For example, we can understand that a person is sitting on or standing up from a chair just considering a still image.

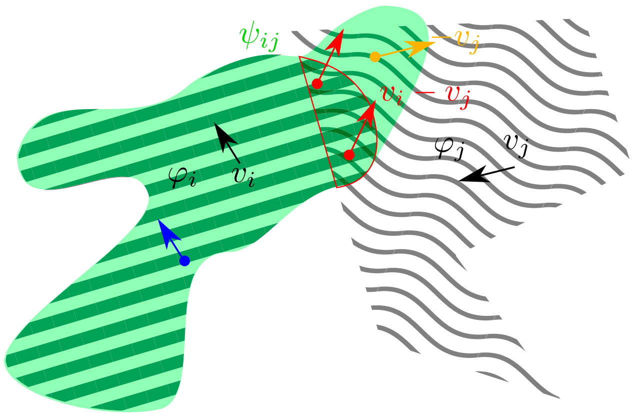

FIGURE 4 | Illustration of **Eq. 15** with $b_j = -1$. The two considered features φ_i (diagonal lines), φ_j (wavy lines) translate over the frame spatial support with uniform velocities v_i, v_j. The green area represents where ψ_{ij} is on, while the red border identifies the region where φ_i and φ_j overlap. On the overlapping region the velocity fields of the two features are both present and here the affordance field $\psi_{ij}(x, t)$ is constrained to be consistent along the direction $v_i - v_j$ (red arrows). Outside and on the left of the red border, the consistency term in **Eq. 15** essentially collapses to the feature identity constraint **Eq. 9** defined by the invariance motion property with respect to v_i (blue arrow). Finally, in those region where $v_i = 0$, motion consistency of ψ_{ij} is required along $-v_j$ (orange arrow).

where $a_1, \ldots, a_7$ are scalars. Considering the case in which the motion field associated to feature φ_j is everywhere null, the affordance field ψ_{ij} will codify a property only related to φ_i itself and strictly related to its identity, that has to be invariant with respect to v_i. Thus, from this observation, we can infer $a_1 = a_2 = a_3 = a_4 = 0$ and $a_5 = a_6$ so that the above constraint becomes:

$$\partial_t \psi_{ij} + \left(v_i + b_j v_j\right) \cdot \nabla \psi_{ij} = 0 \tag{15}$$

where $b_j = a_7/a_5$. Requiring $b_j = -1$ this constraint assumes a very reasonable physical meaning, that is the motion invariance of the affordance field ψ_{ij} in the reference of feature φ_j. Within this choice, the affordance field is conjugated with the velocity $v_i - v_j$ indeed, which is in fact the relative velocity of feature φ_i in the reference of feature φ_j. Considering points at the border of φ_i, this can lead to slightly expand ψ_{ij} outside the region defined by φ_i itself, as shown in **Figure 4**. In the case $v_i = 0$, the motion consistency is forced "backward" along the pixels' trajectories defined by $-v_j$. In the case $b_j = 1$ **Eq. 15** becomes symmetric under permutations instead so that ψ_{ij} and ψ_{ji} will be developed exploiting the same constraint. This will likely result in the same affordance feature unless some other factor (let us think for example to different initializations in neural architectures) breaks that symmetry. From the classic affordance perspective this is not a desirable property as we can easily understand considering, for example, a knife that is used to slice bread: the affordance transmitted by the knife to the bread would be strictly related to the possibility of being cut or sliced, that is clearly a property that could not be attached to the knife.

Another viable and different alternative to codify the interaction between features may be the one of directly evaluating the affordance as function of the feature fields and their respective velocities: $\tilde{\psi}_{ij}(x, t, (\varphi_i, v_i), (\varphi_j, v_j))$. Here, we are giving up the previous field theory approach, being explicitly codifying the interaction between features in the computational scheme of the $\tilde{\psi}_{ij}$ function. On the other hand, since we have already distinguished $\tilde{\psi}_{ij}$ and φ_i by the different computational structure, requiring the same motion invariance property of φ_i with respect to v_i for the affordance function $\tilde{\psi}_{ij}$ appears a very natural choice, see also **Figure 5**:

$$\partial_t \tilde{\psi}_{ij} + v_i \cdot \nabla \tilde{\psi}_{ij} = 0. \tag{16}$$

Given the possible great variability of velocity fields in a visual scene, let us underline that within this second approach some problems in the learning of the affordance function $\tilde{\psi}_{ij}(x, t, (\varphi_i, v_i), (\varphi_j, v_j))$ may emerge. Moreover, to pursue the fascinating idea to describe all the visual processes entirely through visual fields defined on the frame spatial support, in the following we will only consider the affordance field $\psi_{ij}(x, t)$ and the related motion invariance property **Eq. 15** with $b_j = -1$.

Given a certain visual environment we can easily realize that, as time goes by, object interactions begin obeying statistical regularities and the interactions of feature φ_i with the others become very well defined. Hence, the notion of ψ_{ij} can be evolved towards the *inherent affordance* ψ_i of feature φ_i, which is in fact a property associated with φ_i while living in a certain visual environment. For example, thinking in terms of the classic notion of affordance, when considering a knife the related inherent affordance property is gained by being manipulated, in a certain way, by a virtually unbounded number of different people. Based on **Eq. 15** ($b_j = -1$) we define the inherent feature affordance as the function $\psi_i(x, t)$ which satisfies.

$$\partial_t \psi_i + \left(v_i - v_j\right) \cdot \nabla \psi_i = 0, \qquad 1 \leq i, j \leq n. \tag{17}$$

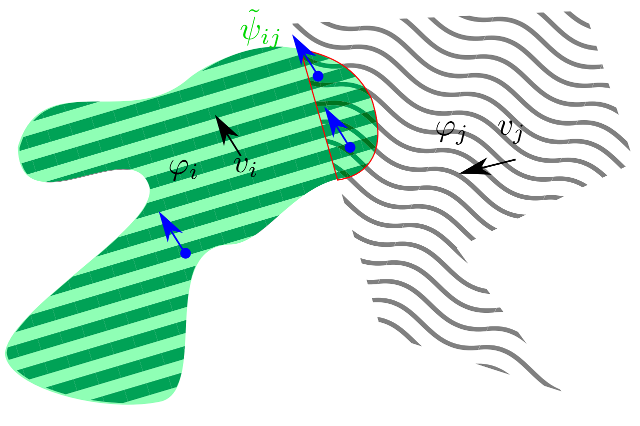

FIGURE 5 | Illustration of **Eq. 16**. The two considered features φ_i (diagonal lines), φ_j (wavy lines) translate over the frame spatial support with uniform velocities v_i, v_j. The green area represents where $\tilde{\psi}_{ij}$ is on, while the red border identifies the region where φ_i and φ_j overlap. In this case, the motion invariance property of the affordance feature $\tilde{\psi}_{ij}$ is the same of the original feature field $\varphi_i(x, t)$. Blue arrows identify the direction along which motion coherence of $\tilde{\psi}_{ij}$ is required.

Let us note that the above formula can also be interpreted as the motion invariance property of ψ_i with respect to the velocity $v_i - \sum_{j=1}^{n} v_j/n$. The identification feature φ_i pairs with the corresponding affordance feature ψ_i, and the visual scene turns out to be effectively described by the collection of visual fields $\mathcal{V} = \{(\varphi_i, \psi_i, v_i)\}$. In a sense, ψ_i can be thought of as the abstraction of φ_i, as it arises from its environmental interactions. A few comments are in order concerning these visual fields.

- The pairing of φ_i and ψ_i relies on the same optical flow which comes from φ_i. This makes sense, since the inherent affordance is a feature that is expected to gain abstraction coming from the interactions with other features, whereas the actual optical flow can only come from identifiable entities that are naturally defined by φ_i.
- The inherent affordance features still bring with them a significant amount of redundant information. This can be understood when considering especially high level features that closely resemble objects. For example, while we may have many different chairs in a certain environment, one would expect to have only a single concept of chair. On the opposite, ψ assigns many different affordance variables that are somewhat stimulated by a specific identifiable feature. This corresponds to thinking of these affordance features as entities that are generated by a corresponding identity feature.
- The collection of visual fields $\mathcal{V}$ is the support for high-level decisions. Of course, the recognition of specific objects does only involve the field φ_i, whereas the abstract affordance semantic labeling is supported by features ψ_i.

In order to abstract the notion of affordance even further we can, for instance, proceed as follows: for each $\kappa = 1,\ldots, n$ we can consider another set of fields $\chi_\kappa\colon \Gamma \to \mathbb{R}$ each of which satisfies the following condition

$$\partial_t \chi_\kappa + v_j \cdot \nabla \chi_\kappa = 0, \quad j = 1, \ldots, n. \tag{18}$$

In this way the variables χ_κ do not depend, unlike for ψ, on a particular v_i, which contributes to lose the link with its firing feature. Moreover, they need to take into account, during their development, multiple motion fields which results in a motion invariant property with respect to the average velocity $\sum_{j=1}^{n} v_j/n$ and in a greater level of abstraction.

Once the set of the χ_κ is given, a method to select the most relevant affordances could simply be achieved through a linear combination. In other words, a subselection of $\chi_1, \ldots, \chi_n$ can be performed by considering for each $l = 1, \ldots, n_\chi < n$ the linear combinations

$$X_l := \sum_{\kappa=1}^{n} a_{l\kappa} \chi_\kappa, \tag{19}$$

where $(a_{l\kappa}) \in \mathbb{R}^{n_\chi \times n}$ is a matrix of learnable parameters. Notice that since $X_l \in \operatorname{span}(\chi_1, \ldots, \chi_n)$, as we remarked in **Section 3.4**, then $\partial_t X_l + v_j \cdot \nabla X_l = 0$ for all $j = 1, \ldots, n$. It is worth mentioning that the learning of coefficients $a_{l\kappa}$ does not involve motion invariance principles. Interestingly, they can be used for additional developmental steps like that of object recognition. For example, they can be learned under the classic supervised framework along with the correspondent regularization.

3.3 Regularization Issues

We have already discussed the ill-posed definition of features conjugated with their corresponding optical flow. Interestingly, we have also shown that a feature group $\phi = (\varphi_i, \ldots, \varphi_m)'$ with conjugate velocity v exhibits an inherent regularization that, however, does not prevent ill-positioning, especially when one is interested in developing abstract features that are likely constant over large regions of the frame spatial support.

Let us assume that we are given n feature groups ϕ_i, $i = 1, \ldots n$, each one composed of m_i single features (m_i-dimensional feature vector) $\phi_i \in \mathbb{R}^{m_i}$. Furthermore, let v_i be the velocity field shared by each component of feature group ϕ_i and let us also denote with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ and with $\boldsymbol{v} = (v_1, \ldots, v_n)$. We can then impose the following generalization of the smoothness term (6), used for the classical optical flow estimation, to the velocities and the corresponding visual features:

$$E = \frac{1}{2} \sum_{i=1}^{n} \int_\Gamma \left(\|\nabla v_i(x,t)\|^2 + \lambda_\varphi \left|\phi_i(x,t)\right|^2 + \lambda_\nabla \left\|\nabla \phi_i(x,t)\right\|^2 \right) d\mu(x,t). \tag{20}$$

Here, the notation $\|Z\|^2$ (generic argument matrix Z) means $\sum_{i,j} Z_{ij}^2$, while λ_φ, λ_∇ are positive constants that express the relative weight of the regularization terms. First of all, notice that E is a functional of the pairs $\{(\phi_i, v_i)\}$, that is, once they are given, we can compute $E(\boldsymbol{\phi}, \boldsymbol{v})$. On the contrary, the index used to regularize the classical Horn–Schunck optical flow **Eq. 6** only depends on v. The dependence on visual features and their temporal dynamic in **Eq. 20** is explained considering that, while the brightness is given, the features are learned as time goes by, which is just another facet of the feature-velocity conjugation. Moreover, it is worth mentioning that E only involves spatial smoothness whereas it doesn't contain any time regularization term. There is also another difference with respect to the classic optical flow regularization **Eq. 6**, that is the penalizing term $(1/2)|\phi_i|^2$ which favors the development of $\phi_i = 0$. Of course, there is no such requirement in classic optical flow since, as already stated, the brightness is given. On the opposite, the discovery of visual features is expected to be driven by motion information, but their "default value" is expected to be null. We can promptly see that the introduction of the regularization term **Eq. 20** does not suffice to achieve a well-posed learning problem. The motion invariance condition **Eq. 9** is still satisfied by the trivial constant solution $\varphi = c_\varphi$ indeed.

Important additional information comes from the need of exhibiting the human visual skill of reconstructing pictures from our symbolic representation. At a certain level of abstraction, the features that are gained by motion invariance possess a certain degree of semantics that is needed to interpret the scene. However, visual agents are also expected to deal with actions and react accordingly. As such, a uniform cognitive task that visual agents are expected to carry out is that of predicting what will happen next, which is translated into the capability of guessing the next incoming few frames in the scene. We can think of a predictive computational scheme based on the φ_i codes $y(x,t) = \alpha_y(\boldsymbol{\phi}(\cdot,t),t)(x)$, where α_y: $(\mathbb{R}^{\Omega})^n \times [0,T] \to \mathbb{R}^{\Omega}$ and the prediction y needs to satisfy the condition established by the index

$$R = \frac{1}{2}\int_{\Gamma} \left(y(x,t) - I_t(x,t)\right)^2 d\mu(x,t). \tag{21}$$

Of course, as the visual agent gains the capability of predicting what will come next, it means that the developed internal representation based on features $\boldsymbol{\phi}$ cannot correspond with the mentioned trivial solution. Interestingly, it looks like visual perception does not come alone: the typical paired skill of visual prediction that animals exhibit helps regularizing the problem of developing invariant features. Clearly, the $\varphi = c_\varphi$ which satisfies motion invariance is no longer acceptable since it does not reconstruct the input. This motivates the involvement of prediction skills typical of action that, again, seems to be interwound with perception.

Having described all the regularization terms necessary to the well-posedness of the learning problem, we can introduce the following functional.

$$S(\boldsymbol{\phi},\boldsymbol{v}) = A(\boldsymbol{\phi},\boldsymbol{v}) + \lambda_E E(\boldsymbol{\phi},\boldsymbol{v}) + \lambda_R R(\boldsymbol{\phi},\boldsymbol{v}) \tag{22}$$

where $A(\boldsymbol{\phi}, \boldsymbol{v})$ is the direct generalization to feature groups of **Eq. 10**, that is $A(\boldsymbol{\phi},\boldsymbol{v}) = \frac{1}{2}\sum_{i=1}^{n}\int_{\Omega\times[0,T]}|\partial_t\phi^i + v\cdot\nabla\phi^i|^2 d\mu(x,t)$. Here, λ_R and $\lambda_E > 0$ are the regularization parameters. Learning to see means to discover the indissoluble pair $(\boldsymbol{\phi}^\star, \boldsymbol{v}^\star)$ such that

$$\left(\boldsymbol{\phi}^\star,\boldsymbol{v}^\star\right) = \underset{(\boldsymbol{\phi},\boldsymbol{v})}{\arg\min}\, S(\boldsymbol{\phi},\boldsymbol{v}). \tag{23}$$

Basically, the minimization is expected to return the pair $(\boldsymbol{\phi}, \boldsymbol{v})$, whose terms should nominally be conjugated. The case in which we reduce to consider only the brightness, that is when the only $\boldsymbol{\phi}$ is I, corresponds with the classic problem of optical flow estimation. Of course, in this case the term R is absent and the problem has a classic solution. Another special case is when there is no motion, so as the integrand in the definition **Eq. 10** of A is simply null $\forall i = 1, \ldots, n$. In this case, the learning problem reduces to the unsupervised extraction of features $\boldsymbol{\phi}$.

The learning of ψ_i can be based on a formulation that closely resembles what has been done for φ_i, for which we have already considered the regularization issues. In the case of ψ_i we can get rid of the trivial constant solution by minimizing.

$$I_\psi = \sum_{i=1}^{n}\int_{\Gamma}(1-\psi_i(x,t))\varphi_i(x,t)\,d\mu(x,t), \tag{24}$$

which comes from the p-norm translation (Gori and Melacci, 2013; Gnecco et al., 2015) of the logic implication $\varphi_i \Rightarrow \psi_i$. Here we are assuming that φ_i, ψ_i range in [0, 1], so as whenever φ_i gets close to 1, it forces the same for ψ_i. This yields a well-posed formulation thus avoiding the trivial solution.

As far as the χ are concerned, like for ψ, we ask for the minimization of

$$I_\chi = \sum_{k=1}^{n}\int_{\Gamma}(1-\chi_\kappa(x,t))\psi_\kappa(x,t)\,d\mu(x,t), \tag{25}$$

that comes from the p-norm translation of the logic implication $\psi_\kappa \Rightarrow \chi_\kappa$. While this regularization term settles the value of χ_κ on the corresponding ψ_κ, notice that the motion invariance condition (18) does not assume any privilege with respect to the *firing feature* ψ_κ.

3.4 Deep Networks-Based Realization of Vision Fields

In the previous sections we discussed invariance properties of visual features that lead to model the processes of computational vision as transport equations on the visual fields, see **Eqs 9**, **15**, **17**, **18**. Some of those properties are based on the concept of consistency under motion, others lead to a generalization of the concept of affordance. In this section we will discuss how the features φ, ψ and χ, along with the velocity fields v_φ, can be represented in terms of neural networks that operate on a visual stream and how the above theory can be interpreted in a classical framework of machine learning.

The first step we need to perform consists in moving to a discretized frame spatial support $\Omega^\diamond = \{(i,j) \in \mathbb{N}\text{: } 0 < i \le w, \quad 0 < i \le h\}$. As a consequence, the fields φ_i and the velocities can be viewed as vector-valued functions of time $t \mapsto \varphi_i(t) \in \mathbb{R}^{\Omega^\diamond}$ and $t \mapsto v_{\varphi_i}(t) \in (\mathbb{R}^{\Omega^\diamond})^2$; similarly the discretized brightness can be seen as a map $t \mapsto I(t) \in \mathbb{R}^{\Omega^\diamond}$.[3] Then, features φ_i (and similarly the fields ψ and χ) can be modelled as neural networks Φ_i: $\mathbb{R}^N \times \mathbb{R}^{\Omega^\diamond} \to \mathbb{R}^{\Omega^\diamond}$ that given the brightness I at a certain instant and a set of N weights $w_{\Phi_i} \in \mathbb{R}^N$ yield the value $\Phi_i(I, w_{\Phi_i})$ of φ_i. Similarly, the velocities v_φ can be estimated by a neural network V_i: $\mathbb{R}^M \times \mathbb{R}^{\Omega^\diamond} \times (\mathbb{R}^{\Omega^\diamond})^2 \to (\mathbb{R}^{\Omega^\diamond})^2$ that takes as inputs the temporal partial derivative $\dot{I}$ (that is the discrete version of the term $\partial_t I$), the discrete spatial gradient ∇I of I and a given a set of M weights w_{v_i} in order to predict $V_i(w_{v_i}, \dot{I}, \nabla I)$ as the velocity field v_{φ_i}.

It should be noted that, within this framework, the learning problem for the fields φ, χ, η and v_φ, that is based on the principles described in this paper and that is defined by the optimization problem of the form described in **Eq. 23**, becomes a finite-dimensional learning problem on the weights of the neural models.

Thus, learning will be affected by the structure (i.e. the architecture) of the network that we choose. Recent successes of deep learning within the realm of computer vision suggest that natural choices for Φ would be Deep Convolutional Neural Networks (DCNN). More precisely, the features extracted at level ℓ of a DCNN can be identified with a group of Φ_i; in this way we are establishing a hierarchy between features that, in turn, suggests a natural way in which we could perform the

[3] Here we are overloading the symbols φ, v_φ and I in order to avoid a cumbersome notation. In the previous sections φ, v_φ and I are functions defined over the spatio-temporal cylinder $\Omega \times [0, T]$, here they are instead regarded as vector-valued functions of time only.

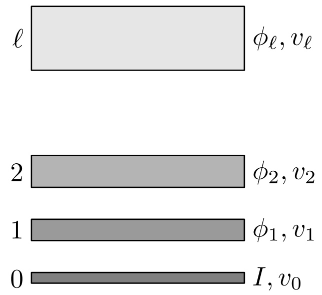

FIGURE 6 | Different visual features along with the corresponding velocity fields.

grouping operation that we discussed in **Section 3.1**. In this way, features that are at the same level of a CNN share the same velocity (see **Figure 6**). In the case of velocities, CNN-based architectures like the one employed by FlowNet (see Fischer et al. (2015)) have been already proven to be suitable to model velocity fields.

It is also important to bear in mind that the choice of a specific neural architecture has strong repercussions on the way the invariance conditions are satisfied. For instance, let us consider the case of Convolutional Networks together with the fundamental condition expressed by **Eq. 9**. In this case, since CNN are equivariant under translations, any feature that tracks a uniformly translating motion of the brightness will automatically satisfy **Eq. 9** with the same velocity of the translation of the input.

4 DISCUSSION

In this paper we have proposed motion invariance principles that lead to discover identification features and more abstract features that are somewhat inspired to the notion of affordance. Those principles are expressed by motion invariance equations that characterize the interesting visual field interactions. The conjunction with features φ leads to believe that those features and their own velocities represent an indissoluble pair. Basically, the presence of a visual feature in the frame spatial support corresponds with its own optical flow, so as they must be jointly detectable. Apparently, in case of a visual agent without relative movement with respect to an object, this sounds odd. What if the object is fixed? Interestingly, foveate animals always experiment movement in the frame of reference of their eyes, and something similar can be experimented in computers by the simulation of focus of attention Zanca et al. (2019), Faggi et al. (2020). Hence, apart from the case of saccadic movements, foveate animals, like haplorhine primates, are always in front of motion, and conjugation of features with the corresponding optical flow does not result in trivial conditions.

The overall field interaction of features and velocities leads to compose a more abstract picture, since in the extreme case of features that represent objects, as already pointed out, we see the emergence of the classic notion of affordance. Interestingly, the described mechanisms of field interaction go well beyond the connection with such a high-level cognitive notion. We can promptly realize that it is impossible to understand whether the discussed field interactions come from different objects or if they are in fact generated within the same object. Overall, the discussed field interactions represent a natural mechanism for transmitting information from the video by local mechanisms.

It has been shown that in order to get a well-posedness of the motion invariance problems of **Eqs 9**, **15**, **17**, **18** we need to involve appropriate regularization. In particular, the development of visual features φ requires the correspondent minimization of **Eq. 21**, that somewhat indicates the need of involving action together with perception. Indeed, visual perception coupled with gaze shifts should be considered the *Drosophila* of perception-action loops. Among the variety of active behaviors the organism can fluently engage to purposively act upon and perceive the world (e.g, moving the body, turning the head, manipulating objects), oculomotor behavior is the minimal, least energy, unit. The history of these ideas has been recently reviewed in Bajcsy et al. (2018). At that time, such computational approaches were pervaded by the early work of Gibson (1950) who proposed that perception is due to the combination of the environment in which an agent exists and how that agent interacts with it. He was primarily interested in optic flow that is generated on the frame spatial support when moving through the environment (as when flying) realizing that it was the path of motion itself that enabled the perception of specific elements, while disenabling others. That path of motion was under the control of the agent and thus the agent chooses how it perceives its world and what is perceived within it Bajcsy et al. (2018). The basic idea of Gibson's view was that of the exploratory behaviour of the agent. It is worth noting that despite of the pioneering work of Aloimonos et al. (1988), Ballard (1991), and Bajcsy and Campos (1992), gaze dynamics has been by and large overlooked in computer vision. The current state of affairs is that most effort is spent on salience modelling Borji (2021), Borji and Itti (2013) as a tool for predicting where/what to look at (the tacit though questionable assumption is that, once suitably computed, salience would be predictive of gaze). Interestingly enough, and rooted in the animate vision approach, Ballard set out the idea of predictive coding Rao and Ballard (1999):

We describe a model of visual processing in which feedback connections from a higher-to a lower-order visual cortical area carry predictions of lower-level neural activities, whereas the feedforward connections carry the residual errors between the predictions and the actual lower level activities. When exposed to natural images, a hierarchical network of model neurons implementing such a model developed simple cell-like receptive fields. A subset of neurons responsible for carrying the residual errors showed endstopping and other extra-classical receptive field effects. These results suggest that rather than being exclusively feedforward phenomena, nonclassical surround effects in the visual cortex may also result from corticocortical feedback as a consequence of the visual system using an efficient hierarchical strategy for encoding natural images.

This idea has gained currency in recent research covering many fields from theoretical cognitive neuroscience (e.g., Knill and Pouget, 2004; Ma et al., 2006) to philosophy Clark (2013). Currently, the most influential approach in this perspective has been proposed by Friston (e.g., Feldman and Friston, 2010; Friston, 2010) who considered a variational approximation to Bayesian inference and prediction (free energy minimization, minimization of action functionals, etc).

The principles on visual feature flow introduced in this paper might also have an impact in computer vision, since one can reasonably believe that the proposed invariances might overcome one of the major current limitation of supervised learning paradigms, namely the need of a huge amount of labeled examples. This being said, deep neural networks, along with their powerful approximation capabilities, could provide us with the ideal computational structure to complete the theoretical framework here proposed.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING AND ACKNOWLEDGMENTS

This work was partially supported by the PRIN 2017 project RexLearn (Reliable and Explainable Adversarial Machine Learning), funded by the Italian Ministry of Education, University and Research (grant no. 2017TWNMH2).

REFERENCES

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active Vision. *Int. J. Comput. Vis.* 1, 333–356. doi:10.1007/bf00133571

Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A., and Poggio, T. (2016). Unsupervised Learning of Invariant Representations. *Theor. Comput. Sci.* 633, 112–121. doi:10.1016/j.tcs.2015.06.048

Ardón, P., Pairet, È., Lohan, K. S., Ramamoorthy, S., and Petrick, R. (2020). *Affordances in Robotic Tasks–A Survey*. arXiv preprint arXiv:2004.07400.

Aubert, G., and Kornprobst, P. (2006). *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*, Vol. 147. New York City, NY: Springer Science & Business Media.

Bajcsy, R., and Campos, M. (1992). Active and Exploratory Perception. *CVGIP: Image Understanding* 56, 31–40. doi:10.1016/1049-9660(92)90083-f

Bajcsy, R., Aloimonos, Y., and Tsotsos, J. K. (2018). Revisiting Active Perception. *Auton. Robot* 42, 177–196. doi:10.1007/s10514-017-9615-3

Ballard, D. H. (1991). Animate Vision. *Artif. Intell.* 48, 57–86. doi:10.1016/0004-3702(91)90080-4

Betti, A., Gori, M., and Melacci, S. (2020). Learning Visual Features under Motion Invariance. *Neural Networks* 126, 275–299. doi:10.1016/j.neunet.2020.03.013

Bhaskaran, V., and Konstantinides, K. (1997). *Image and Video Compression Standards: Algorithms and Architectures*. Boston, MA: Springer.

Borji, A., and Itti, L. (2013). State-of-the-art in Visual Attention Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 185–207. doi:10.1109/tpami.2012.89

Borji, A. (2021). Saliency Prediction in the Deep Learning Era: Successes and Limitations. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 679–700. doi:10.1109/TPAMI.2019.2935715

Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behav. Brain Sci.* 36, 181–204. doi:10.1017/s0140525x12000477

Cohen, T., and Welling, M. (2016). "Group Equivariant Convolutional Networks," in International conference on machine learning (PMLR), New York City, NY, June 20–22, 2016, 2990–2999.

Denton, E., and Birodkar, V. (2017). Unsupervised Learning of Disentangled Representations from Video. *Advances in Neural Information Processing Systems*. Editors I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Red Hook, NY: Curran Associates, Inc.) 30. Available at https://proceedings.neurips.cc/paper/2017/file/2d2ca7eedf739ef4c3800713ec482e1a-Paper.pdf

Faggi, L., Betti, A., Zanca, D., Melacci, S., and Gori, M. (2020). Wave Propagation of Visual Stimuli in Focus of Attention. *CoRR*. arXiv preprint arXiv:2006.11035.

Farnebäck, G. (2003). "Two-frame Motion Estimation Based on Polynomial Expansion," in Scandinavian Conference on Image Analysis, Halmstad, Sweden, June 29–July 2, 2003 (Springer), 363–370. doi:10.1007/3-540-45103-x_50

Feldman, H., and Friston, K. J. (2010). Attention, Uncertainty, and Free-Energy. *Front. Hum. Neurosci.* 4, 215. doi:10.3389/fnhum.2010.00215

Fischer, P., Dosovitskiy, A., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., et al. (2015). Flownet: Learning Optical Flow with Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2758–2766.

Friston, K. (2010). The Free-Energy Principle: a Unified Brain Theory? *Nat. Rev. Neurosci.* 11, 127–138. doi:10.1038/nrn2787

Gens, R., and Domingos, P. M. (2014). Deep Symmetry Networks. *Adv. Neural Inf. Process. Syst.* 27, 2537–2545. doi:10.5555/2969033.2969110

Gibson, J. J. (1950). Boston, MA: *The Perception of the Visual World*. Houghton Mifflin.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*, Vol. 2. Boston, MA: Houghton Mifflin Boston.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin Comp.

Gnecco, G., Gori, M., Melacci, S., and Sanguineti, M. (2015). Foundations of Support Constraint Machines. *Neural Comput.* 27, 388–480. doi:10.1162/neco_a_00686

Gori, M., and Melacci, S. (2013). Constraint Verification with Kernel Machines. *IEEE Trans. Neural Netw. Learn. Syst.* 24, 825–831. doi:10.1109/tnnls.2013.2241787

Hassanin, M., Khan, S., and Tahtali, M. (2021). Visual Affordance and Function Understanding. *ACM Comput. Surv.* 54, 1–35. doi:10.1145/3446370

Horn, B. K. P., and Schunck, B. G. (1981). Determining Optical Flow. *Artif. Intell.* 17, 185–203. doi:10.1016/0004-3702(81)90024-2

Hsieh, J.-T., Liu, B., Huang, D.-A., Li, F.-F., and Niebles, J. C. (2018). Learning to Decompose and Disentangle Representations for Video Prediction. *NeurIPS*, 514–524. Available at http://papers.nips.cc/paper/7333-learning-to-decompose-and-disentangle-representations-for-video-prediction.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). "Flownet 2.0: Evolution of Optical Flow Estimation with Deep Networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, honolulu, HI, July 21–26, 2017, 2462–2470. doi:10.1109/cvpr.2017.179

Knill, D. C., and Pouget, A. (2004). The Bayesian Brain: the Role of Uncertainty in Neural Coding and Computation. *Trends Neurosci.* 27, 712–719. doi:10.1016/j.tins.2004.10.007

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105.

Lee, T., and Soatto, S. (2011). Video-based Descriptors for Object Recognition. *Image Vis. Comput.* 29, 639–652. doi:10.1016/j.imavis.2011.08.003

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian Inference with Probabilistic Population Codes. *Nat. Neurosci.* 9, 1432–1438. doi:10.1038/nn1790

Milner, A. D. (2017). How Do the Two Visual Streams Interact with Each Other? *Exp. Brain Res.* 235, 1297–1308. doi:10.1007/s00221-017-4917-4

Mobahi, H., Collobert, R., and Weston, J. (2009). "Deep Learning from Temporal Coherence in Video," in Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, June 14–18, 2009, 737–744. doi:10.1145/1553374.1553469

Pan, Y., Li, Y., Yao, T., Mei, T., Li, H., and Rui, Y. (2016). "Learning Deep Intrinsic Video Representation by Exploring Temporal Coherence and Graph Structure," in IJCAI, New York City, NY, July 9–15, 2016, (Citeseer), 3832–3838.

Pathak, D., Girshick, R., Dollár, P., Darrell, T., and Hariharan, B. (2017). "Learning Features by Watching Objects Move," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, July 21–26, 2017, 2701–2710. doi:10.1109/cvpr.2017.638

Rao, R. P. N., and Ballard, D. H. (1999). Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-Field Effects. *Nat. Neurosci.* 2, 79–87. doi:10.1038/4580

Redondo-Cabrera, C., and Lopez-Sastre, R. (2019). Unsupervised Learning from Videos Using Temporal Coherency Deep Networks. *Comput. Vis. Image Understanding* 179, 79–89. doi:10.1016/j.cviu.2018.08.003

Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J. (2018). "Mocogan: Decomposing Motion and Content for Video Generation," in Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, June 18–23, 2018, 1526–1535. doi:10.1109/cvpr.2018.00165

Verri, A., and Poggio, T. (1989). Motion Field and Optical Flow: Qualitative Properties. *IEEE Trans. Pattern Anal. Machine Intell.* 11, 490–498. doi:10.1109/34.24781

Verri, A. (1987). "Against Quantititive Optical Flow," in Proc. First Int'l Conf. Computer Vision, London, UK, June 8–11, 1987 (London, 171–180).

Villegas, R., Yang, J., Hong, S., Lin, X., and Lee, H. (2017). Decomposing Motion and Content for Natural Video Sequence Prediction. *ICLR*. arXiv preprint arXiv: 1706.08033.

Wang, X., and Gupta, A. (2015). "Unsupervised Learning of Visual Representations Using Videos," in Proceedings of the IEEE international conference on computer vision, Santiago, Chile, December 7–13, 2015, 2794–2802. doi:10.1109/iccv.2015.320

Wang, Y., Bilinski, P., Bremond, F., and Dantcheva, A. (2020). "G3an: Disentangling Appearance and Motion for Video Generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, June 13–19, 2020, 5264–5273. doi:10.1109/cvpr42600.2020.00531

Zanca, D., Melacci, S., and Gori, M. (2019). Gravitational Laws of Focus of Attention. *IEEE Trans. Pattern Anal. Mach Intell.* 42, 2983–2995. doi:10.1109/TPAMI.2019.2920636

Zhai, M., Xiang, X., Lv, N., and Kong, X. (2021). Optical Flow and Scene Flow Estimation: A Survey. *Pattern Recognit.* 114, 107861. doi:10.1016/j.patcog.2021.107861

Zou, W. Y., Ng, A. Y., and Yu, K. (2011). "Unsupervised Learning of Visual Invariance with Temporal Coherence," in NIPS 2011 workshop on deep learning and unsupervised feature learning, Granada, Spain, December 12–17, 2011.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.