

A Hamiltonian-Based Algorithm for Measurements Clustering

Daniele Casagrande and Alessandro Astolfi

Abstract—The paper describes a novel method for clustering points in the plane. The proposed algorithm is based on the notions of clustering function and level lines; the clusters are identified as the level sets corresponding to a reference value of the clustering function. The core idea is to regard the clustering function as a Hamiltonian function and to determine the level lines as the trajectories of the associated Hamiltonian system. The method is illustrated on two practical problems.

I. INTRODUCTION

The problem of clustering measurements arises in many applications such as pattern recognition, computer graphics, data analysis, image processing and market research (for a survey on the subject see, among others, the books [1], [2], [3] and [4]). The main goal of a clustering process is that of partitioning a set of points into a number of subsets in such a way that data points belonging to the same subset, or *group*, are more correlated one to the other than data points belonging to different subsets.

The results presented are motivated by the possibility of exploiting the clustering of data to derive collective properties of detected objects. In particular, we consider a scenario where a finite number of objects are moving on the two-dimensional plane; we suppose that at a given time-instant the displacements of the objects on the plane is detected and the position of each object is mapped into a point. Thus the data points, in this case, are not mere points in the space but correspond to real objects and the clustering method can be used to provide information on the collective behaviour of the objects. It is reasonable to assume, for instance, that points (objects) that are close one to the other behave in a somehow correlated way. On the other hand, isolated points are likely to correspond to objects whose behaviour is uncorrelated to the others. Moreover, tracking a group of objects is more convenient, in terms of computational load, than tracking each of them individually.

In the literature many algorithms have been proposed to perform the clustering when the number of clusters is known. The ideas underlying these algorithms are the more diverse, e.g. deterministic algorithms [5], probabilistic algorithms [6], simulated-annealing-based algorithms [7], evolutionary-based algorithms [8], tabu-search-based algorithms [9] and hierarchical algorithms [10], just to mention

This work has been partially supported by the SEAS DTC project "Dynamic Set Recognition".

D. Casagrande and A. Astolfi are with the Department of Electrical and Electronic Engineering, Imperial College, Exhibition Road, London SW7 2AZ, UK (Email: d.casagrande@ic.ac.uk, a.astolfi@ic.ac.uk). A. Astolfi is also with the Dipartimento di Informatica, Sistemi e Produzione, Università degli Studi di Roma "Tor Vergata", via del Politecnico, 1 - 00133 Roma, Italy.

a few. Other works consider the *cluster validity*, namely the problem of finding the optimal number of clusters in a given set of data points, and develop algorithms [11], [12] for automatic clustering, i.e. algorithms that solve both the cluster validity and the clustering itself.

The approach followed herein is novel as we are interested not only in a mere clustering of the single points but also in the identification of *regions of interest* surrounding the points belonging to the same cluster. The idea is somewhat similar to [13], where the clustering is achieved by defining a *clustering function* and grouping together the points belonging to the same *level set*. Nevertheless, while in [13] a level function is introduced only to implicitly define the clusters, herein the clustering function is used to define Hamiltonian dynamics. Subsequently, the trajectory of the state of the corresponding Hamiltonian system is proven to converge to the contour of the level set, namely the contour of the region of interest, no matter what the initial condition is.

The paper is organized as follows. In Section II the notation is described, while the problem is formulated in detail in Section III. Section IV describes the class of Hamiltonian functions used and illustrates the theoretical achievements. The results of some simulations are reported in Section V. Conclusions and future developments are finally drawn in Section VI.

II. NOTATION

$\overline{\mathbb{R}}$ denotes the extended set of real numbers, namely $\mathbb{R} \cup \{-\infty, +\infty\}$. For a pair of points $a \in \mathbb{R}^2$ and $b \in \mathbb{R}^2$, $d(a, b)$ denotes the distance between a and b . Given a point $a \in \mathbb{R}^2$ and a set $C \subset \mathbb{R}^2$, $d(a, C)$ denotes the distance between a and C , namely $d(a, C) = \inf\{d(a, c) : c \in C\}$. $\overline{B_{\xi_C}^R}$ denotes the closed ball of centre ξ_C and radius R , namely the set $\{\xi \in \mathbb{R}^2 : \|\xi - \xi_C\| \leq R\}$. \mathcal{C}^1 denotes the set of continuous and differentiable functions with continuous first-order derivative.

Finally, we remind that the *diameter* of a set C is defined as $\text{diam}(C) \triangleq \sup\{d(x, y) : x \in C, y \in C\}$.

III. DESCRIPTION OF THE PROBLEM

Suppose that at a given time-instant the position of N objects in \mathbb{R}^2 is detected and that the position of the i -th object, for $i = 1, \dots, N$, is mapped into a point ξ_i represented by a pair of coordinates: $\xi_i = (x_i, y_i)$. The set of points to be clustered is, then, $\mathcal{D} = \{\xi_1, \dots, \xi_N\} \subset \mathbb{R}^2$. The method presented in the following solves, at the same time, both the problem of finding the optimal number K of clusters and the problem of partitioning \mathcal{D} into sets $\mathcal{D}_1, \dots, \mathcal{D}_K$.

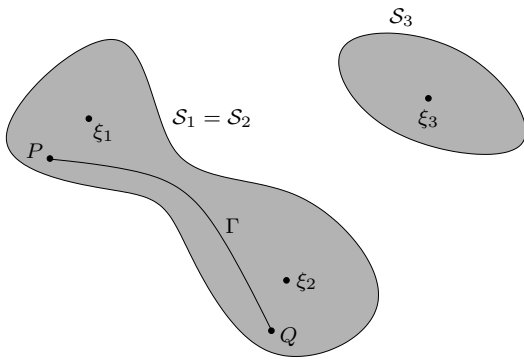


Fig. 1. A possible configuration for $\mathcal{D} = \{\xi_1, \xi_2, \xi_3\}$. According to Definition 1, the points ξ_1 and ξ_2 are linked and so are P and Q as well as P and ξ_1 . Therefore ξ_1 and ξ_2 belong to the same group \mathcal{D}_1 . The region surrounding \mathcal{D}_1 is $\mathcal{R}_1 = \mathcal{R}_2$ and its contour is $\mathcal{S}_1 = \mathcal{S}_2$. On the other hand $\xi_3 \in \mathcal{D}_2$ is isolated. The shaded regions \mathcal{D}_1 and \mathcal{D}_2 are made by points in which the level function is greater than the reference value.

This is achieved in two steps. We first define a level function H ; subsequently, we identify the clusters as the sets of points *internal* (with the meaning that is specified in the following) to the closed curves corresponding to the boundary of a particular level set. The main idea is to consider the closed curves as trajectories of Hamiltonian systems with Hamiltonian function H . We begin by defining the notion of *link* between two points.

Definition 1: Given a function $H : \mathbb{R}^2 \rightarrow \overline{\mathbb{R}}$ and a reference value H_r , two points $P \in \mathbb{R}^2$ and $Q \in \mathbb{R}^2$ are *linked with respect to H_r* if

- (i) $H(P) \geq H_r$,
- (ii) $H(Q) \geq H_r$,
- (iii) there exists a continuous curve $\Gamma \subset \mathbb{R}^2$ such that $P \in \Gamma$, $Q \in \Gamma$ and $H(R) \geq H_r$ for all points $R \in \Gamma$.

Moreover, we say that two points $\xi_A \in \mathcal{D}$ and $\xi_B \in \mathcal{D}$ *belong to the same group with respect to H_r* if they are linked (w.r.t. H_r). Finally we say that a point $\xi_C \in \mathcal{D}$ is *isolated with respect to H_r* if there does not exist any $\xi \in \mathcal{D}$ such that ξ_C and ξ are linked (w.r.t. H_r). \diamond

To provide the reader with more insight on the meaning of Definition 1 we remark that it is defined not only for the points in \mathcal{D} but for all points in \mathbb{R}^2 ; this means that once the function H and the reference level H_r are defined, grouping together two points ξ_A and ξ_B implies identifying a connected region \mathcal{R} , the contour of which is denoted by \mathcal{S} , surrounding both ξ_A and ξ_B , the points of which are all linked each other (see Figure 1)¹. Clearly, the shape and the dimension of such a connected region depend on how H and H_r are selected.

It is easy to check that the *link* relation expressed by Definition 1 has both the symmetry property (if P is linked to Q , then Q is linked to P) and the transitivity property (if P is linked to Q and Q is linked to R , then P is

¹ \mathcal{D}_j , for $j = 1, \dots, K$, denotes a cluster, namely a subset of \mathcal{D} ; \mathcal{R}_j denotes the set of points linked to any point in \mathcal{D}_j ; \mathcal{S}_j denotes the boundary of \mathcal{R}_j .

linked to R). To consider the \mathcal{D}_i 's as classes of equivalence, which is a necessary requirement for correlated objects, the reflexivity property (P is linked to itself) should also hold for the ξ 's. For, it is sufficient to suppose that H_r is such that $H(\xi_i) \geq H_r$, for all $i = 1, \dots, N$.

A. Selection of the level function

In accordance with the aim of rendering the *link* relation an equivalence relation, we first require that the function H and the value of H_r are such that each ξ_i , $i = 1, \dots, N$ is linked to itself; moreover, we prefer \widehat{H} to be positive, smooth and such that $\int_{\mathbb{R}^2} H(x, y) dx dy = \widehat{H} < +\infty$. In particular, we associate to each point $\xi = (x, y)$ a Gaussian function

$$H_i(\xi) = e^{-\rho_i(\xi)^2}, \quad (1)$$

with ³ $\rho_i(\xi) = \rho_i(x, y) \triangleq \|\xi - \xi_i\| = \sqrt{(x - x_i)^2 + (y - y_i)^2}$, and we pick

$$H(\xi) = \sum_{i=1}^N H_i(\xi). \quad (2)$$

With this choice the reflexivity property is guaranteed as long as $H_r < 1$. In fact, since $H_i(\xi_i) = 1$ we always have $H(\xi_i) > 1 > H_r$. Then each ξ_i is linked to itself.

B. On the length of the contour

If we suppose that the region within which the objects are detected is a compact set $\mathcal{C} \subset \mathbb{R}^2$, then an upper bound for the length of the contour of \mathcal{R}_i can be found. To achieve this goal we need some preliminary results.

Lemma 1: If H is as in Equation (2) and $\xi_i \in \mathcal{C} \subset \mathbb{R}^2$, with \mathcal{C} compact, then $\mathcal{S}_i \subset \mathcal{C}_B$, for all $i = 1, \dots, N$, where $\mathcal{C}_B \triangleq \left\{ \xi \in \mathbb{R}^2 : d(\xi, \mathcal{C}) \leq \sqrt{\log_e(N/H_r)} \right\}$. \square

Proof: Consider a point $\zeta \notin \mathcal{C}_B$. For all $i = 1, \dots, N$, $\rho_i(\zeta) > \sqrt{\log_e(N/H_r)}$, hence $H_i(\zeta) < e^{-(\log_e(N/H_r))} = H_r/N$. Therefore $H(\zeta) = \sum_{i=1}^N H_i(\zeta) < N \frac{H_r}{N} = H_r$. Hence $\zeta \notin \mathcal{S}_i$. \blacksquare

Now, in any practical case the objects, and hence the associated points, lie in a box⁴: $\xi_i \in [-A, A] \times [-B, B]$, for all $i = 1, \dots, N$. Therefore, as a consequence of Lemma 1, $\mathcal{C}_B \subset \overline{\mathcal{B}_O^D}$, where

$$D = \sqrt{A^2 + B^2} + \log_e(N/H_r). \quad (3)$$

Thus we have found a compact set which contains all the contours of the level function H . We now exploit this information to prove that, under some hypotheses, the length of each contour is bounded. To prove this fact, we need two preliminary results.

²This property, even though is not exploited in the paper, provides H/\widehat{H} with a behaviour similar to a probability density. This may be used to find the probability that in a future time-instant the i -th object is mapped into a point belongs to \mathcal{R}_i .

³Here we use the Euclidean norm; the results hold when using any weighted norm.

⁴Without loss of generality we consider the box symmetric with respect to the origin.

Lemma 2: Let $\gamma : [0, T] \rightarrow \mathbb{R}^2$ be a \mathcal{C}^1 curve, let $L(\gamma)$ be its length and let $(x(t), y(t))$ denotes the parametric description of γ , namely the image (in \mathbb{R}) of $t \in [0, T]$. Moreover, suppose that γ is contained in a compact set \mathcal{A} whose diameter is $\text{diam}(\mathcal{A}) = M$. If the equation $\dot{x}(t) = 0$ has z_x solutions in $[0, T)$ and the equation $\dot{y}(t) = 0$ has z_y solutions in $[0, T)$, then $L(\gamma) \leq M(z_x + z_y)$. \square

Proof: The length of the curve is given by

$$L(\gamma) = \int_0^T \sqrt{\dot{x}^2 + \dot{y}^2} dt \leq \int_0^T (|\dot{x}(t)| + |\dot{y}(t)|) dt.$$

If $\gamma \subset \mathcal{A}$, then the length spanned in the x direction in a time interval in which \dot{x} does not change sign is bounded by M . Thus, if \dot{x} changes sign (at most) z_x times, we have

$$\int_0^T |\dot{x}(t)| dt \leq M z_x.$$

An analogous result holds for y and the thesis follows immediately. \blacksquare

We apply Lemma 2 to the function H defined by (2) thus proving that the length of the contour of a level set admits an upper bound. Now, as far as the first hypothesis is concerned, namely that the curve is contained in a compact set, we have proven in Lemma 1 that it holds. On the other hand, as far as the second hypothesis is concerned, we make the following assumption⁵.

Assumption 1: Let the level function be as in Equation (2) with H_i defined as in (1). The curve corresponding to the level set $H = H_r$, is such that the equation $\dot{x}(t) = 0$ ($\dot{y}(t) = 0$) has at most a finite number $\sigma_x(N)$ ($\sigma_y(N)$, respectively) of solutions, possibly depending on N .

As a consequence, the total length of the contour of any level set is upper bounded by $D(\sigma_x(N) + \sigma_y(N))$. In the following section this fact is exploited to show that the contours of the regions \mathcal{R}_i 's correspond to the restriction to a finite time-interval of a trajectory of a Hamiltonian system.

IV. MAIN RESULTS

We now show that it is possible to design a simple switching scheme, defining *modified Hamiltonian* dynamics, which in two steps, namely with only one switch, allows to determine the set \mathcal{S}_j , for some $j \in \{1, \dots, K\}$ (where K is the number of clusters), no matter what the initial condition of the Hamiltonian system is. In particular, the two dynamics corresponding to the two steps of the strategy perform the following tasks: the first dynamics are such that *in finite time* the state of the Hamiltonian system reaches \mathcal{S}_j ; the second dynamics are used so that the state covers *in finite time* a closed trajectory which corresponds to the set \mathcal{S}_j .

Suppose that a \mathcal{C}^1 function $H(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$ is given and consider the planar *Hamiltonian* system described by the equations

$$\dot{x} = \frac{\partial H}{\partial y} f(x, y), \quad \dot{y} = -\frac{\partial H}{\partial x} f(x, y), \quad (4)$$

⁵So far, Assumption 1 could not be proven; nevertheless, due to the simple structure of the functions H_i it is quite reasonable.

where $f(x, y)$ is any continuous function such that⁶ $f(x, y) > 0$ for all $(x, y) \in \mathbb{R}^2$. The main feature of system (4) is that the value of H is constant along trajectories, since

$$\dot{H} = \frac{\partial H}{\partial x} \dot{x} + \frac{\partial H}{\partial y} \dot{y} = 0.$$

Now, if the initial condition is a point $(x(0), y(0))$ on the level line corresponding to H_r , namely if $H(x(0), y(0)) = H_r$, then $H(x(t), y(t)) = H_r$ for all $t \geq 0$, thus the trajectory of the state of the system lies on \mathcal{S}_j for some $j \in \{1, \dots, K\}$. On the other hand, if the initial condition does not lie on the level line corresponding to H_r , an additional term needs to be introduced in Equation (4) so that the state reaches $\bigcup_{j=1}^K \mathcal{S}_j$.

A. Finite time convergence

To achieve the mentioned goal, namely to steer the state of the system from any initial condition to a point lying on the level line corresponding to H_r , one can choose, in particular, a term that provides finite time convergence of the trajectory to the set $\bigcup_{j=1}^K \mathcal{S}_j$. Consider the modified system

$$\begin{cases} \dot{x} = f(x, y) \left(\frac{\partial H}{\partial y} - \frac{\partial H}{\partial x} (H - H_r)^{1/3} \right), \\ \dot{y} = -f(x, y) \left(\frac{\partial H}{\partial x} + \frac{\partial H}{\partial y} (H - H_r)^{1/3} \right). \end{cases} \quad (5)$$

In this case one has

$$\begin{aligned} \dot{H} &= \frac{\partial H}{\partial x} \dot{x} + \frac{\partial H}{\partial y} \dot{y} \\ &= - \left(\left(\frac{\partial H}{\partial x} \right)^2 + \left(\frac{\partial H}{\partial y} \right)^2 \right) (H - H_r)^{1/3} f(x, y). \end{aligned}$$

By taking

$$f(x, y) = k_1 \left(\left(\frac{\partial H}{\partial x} \right)^2 + \left(\frac{\partial H}{\partial y} \right)^2 \right)^{-1} \quad (6)$$

for a given $k_1 > 0$, one obtains $\dot{H} = -k_1 (H - H_r)^{1/3}$, the solution of which is

$$H(t) = \begin{cases} H_r + \mu \sqrt{(h^{2/3} - \frac{2}{3} k_1 t)^3}, & \text{for } 0 \leq t \leq \frac{3h^{2/3}}{2k_1}, \\ H_r, & \text{for } t > \frac{3h^{2/3}}{2k_1}, \end{cases}$$

where $\mu = \text{sign}(H - H_r)$, $h = |H_0 - H_r|$ and $H_0 = H(x(0), y(0))$.

Hence $H(t) = H_r$ for $t \geq t^* = 3h^{2/3}/2k_1$.

Remark 1: The function (6) is well-defined for all points in \mathbb{R}^2 except the points ξ^* such that

$$\frac{\partial H}{\partial x} \Big|_{\xi^*} = \frac{\partial H}{\partial y} \Big|_{\xi^*} = 0. \quad (7)$$

The closed level lines containing the points for which (7) holds are *homoclinic* level lines. In the following we suppose that the level line corresponding to the reference level is not homoclinic, i.e. for none of its points (7) holds. Moreover,

⁶This assumption is needed to maintain the orientation of the trajectory.

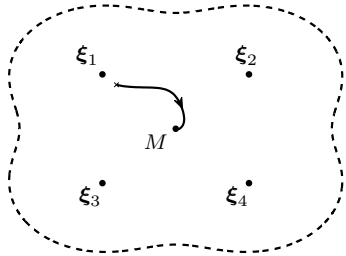


Fig. 2. A possible situation in which the trajectory of system (5) tends to a local minimum (the point M) which does not belong to the desired level line (dashed line).

we suppose that the trajectories of system (5) do not intersect (and do not tend to) stationary points of the function H . \diamond

Remark 2: Condition (7) is verified, in particular, for the local minima of the level function, e.g. the point M in Figure 2. In this case, if the value of the function in the stationary point is greater than H_r and $H(x(0), y(0)) > H_r$, the trajectory of system (5) might get stuck in the local minima (see again Figure 2), without reaching the level line. In the examples presented in the paper due to the particular displacement of the points this does not occur; nonetheless, a generic procedure to avoid this occurrence should be investigated. \diamond

Note that in (2) all the functions H_i 's are bounded, namely $H_i(x, y) \in [0, 1]$, for all $(x, y) \in \mathbb{R}^2$. Therefore $H(x, y) < N$ and $t^* \leq 3N^{2/3}/2k_1$, which means that the time of convergence of the trajectory to the desired level line is bounded no matter what the initial condition $(x(0), y(0))$ is. Moreover, by properly selecting the value of k_1 the time of convergence can be made arbitrarily small.

B. Constant linear velocity along the level lines

Consider again system (5). If the initial condition $\xi_0 = (x_0, y_0)$ is such that $H(\xi_0) = H_r$ and the level line does not contain stationary points of H , then the system is identical to system (4) and $H(t) = H_r$ for all $t \geq 0$. Moreover, in Section III we have proven that the locus of points ξ such that $H(\xi) = H_r$ has a finite length. These two facts imply that a system initialized at ξ_0 and evolving according to (4) (or to (5)) either runs along a homoclinic orbit, i.e. admits an equilibrium state ξ_* such that (7) holds, or its trajectory is periodic. In the latter case we define the *period* of the trajectory as the minimum positive constant T such that $x(t+T) = x(t)$ and $y(t+T) = y(t)$, for all $t \in \mathbb{R}$. If the condition (7) is verified, we set $t = \infty$.

In general, the period of the trajectory depends on the value H_r and on the function f . Nevertheless, by properly selecting the expression of f it is possible to find an upper bound for T . In fact, let s denote the curvilinear coordinate of a point moving along a periodic trajectory (see Figure 3). It is easy to see that the dynamics of s are described by $\dot{s} = \sqrt{\dot{x}^2 + \dot{y}^2}$ (apply Pythagoras's theorem to the triangle

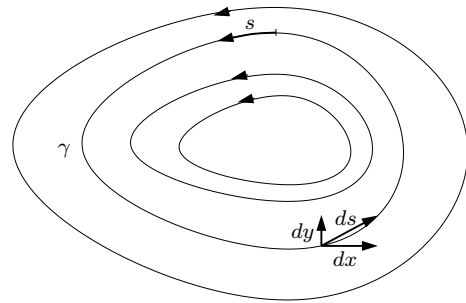


Fig. 3. Some possible trajectories of the Hamiltonian system (4) corresponding to different values of H .

whose sides are dx , dy and ds). Then by picking

$$f(x, y) = k_2 \left(\left(\frac{\partial H}{\partial x} \right)^2 + \left(\frac{\partial H}{\partial y} \right)^2 \right)^{-1/2}, \quad (8)$$

for a given $k_2 > 0$, which is well-defined for all points which do not belong to a homoclinic orbit, one has

$$\dot{s} = \sqrt{\left(\frac{\partial H}{\partial x} \right)^2 f(x, y)^2 + \left(\frac{\partial H}{\partial y} \right)^2 f(x, y)^2} = k_2, \quad (9)$$

which means that the linear velocity of a point moving along the trajectory is constant and may be rendered arbitrarily large. This, together with the fact that the length of the curve is $L(\gamma) \leq D(\sigma_x(N) + \sigma_y(N))$, with D defined by (3), implies that $T \leq D(\sigma_x(N) + \sigma_y(N))/k_2$.

Then we can determine in finite time the contour of \mathcal{R}_j , for some $j \in \{1, \dots, K\}$, by applying the following two-steps strategy.

S1. Select an initial condition $\xi(0)$ and integrate the dynamics defined by (5) with f as in (6) until the state of the system reaches a point $\bar{\xi} \in \bigcup_{j=1}^K \mathcal{S}_j$.

S2. Starting from $\bar{\xi}$, integrate the dynamics defined by (5), with f as in (8), until a whole period of the trajectory has elapsed.

By applying this strategy only once, one finds \mathcal{S}_j for only one $j \in \{1, \dots, K\}$, while the final goal is to find $\bigcup_{j=1}^K \mathcal{S}_j$. A naive methodology to achieve this result is that of applying the strategy **S1-S2** N times, taking ξ_i as initial condition for the i -th iteration. Each ξ_i , in fact, belongs to some region \mathcal{R}_j and is therefore internal to the contour \mathcal{S}_j which is a closed curve contained in a compact set. This exhaustive approach certainly allows to find all the contours; nevertheless it can be improved by exploiting a well-known fact in differential geometry, as explained in the following.

C. The cluster validity problem

Let $\gamma(t) : [0, 1] \rightarrow \mathbb{R}^2$ be a differentiable closed curve (i.e. $\gamma(1) = \gamma(0)$) such that for all $t_1 \in (0, 1)$ and $t_2 \in (0, 1)$, with $t_1 \neq t_2$, $\gamma(t_1) \neq \gamma(t_2)$. Let $x = x(t)$ and $y = y(t)$ be the parametric description of γ and define, for a point $\xi^* = (x^*, y^*) \in \mathbb{R}^2$, the differential one-form

$$d\theta_{\xi^*} = \frac{-(y - y^*)dx + (x - x^*)dy}{(x - x^*)^2 + (y - y^*)^2}.$$

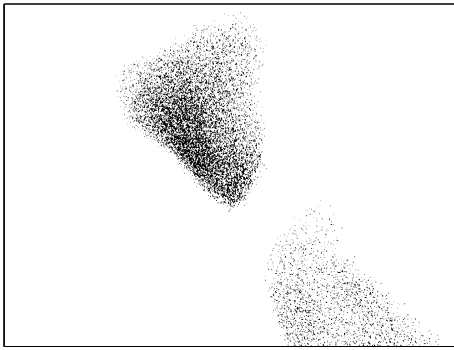


Fig. 4. A black-and-white picture of two flocks of birds.

The *winding number* [14] (or *index*) of γ with respect to ξ^* is defined as

$$\beta(\xi^*, \gamma) \triangleq \frac{1}{2\pi} \oint_{\gamma} d\theta_{\xi^*}.$$

and by Cauchy's theorem

$$\beta(\xi^*, \gamma) = \begin{cases} 0 & \text{if } \xi^* \text{ lies outside the curve,} \\ 1 & \text{if } \xi^* \text{ lies inside the curve.} \end{cases}$$

As a result, in order to determine all the contours corresponding to the set of detected points $\mathcal{D} = \{\xi_1, \xi_2, \dots, \xi_N\}$, to the level function H and to the level set H_r , one can run the following algorithm.

A0. Let $i = 1$ and $j = 1$.

A1. Run **S1-S2** with $\xi(0) = \xi_i$, thus determining the contour \mathcal{S}_j of the region surrounding ξ_i , and for all $k = i + 1, \dots, N$ compute $\beta(\xi_k, \mathcal{S}_j)$.

A2. For all $k = i + 1, \dots, N$, if $\beta(\xi_k, \mathcal{S}_j) = 1$ then $\mathcal{D} \leftarrow \mathcal{D} \setminus \{\xi_k\}$. If $\mathcal{D} \neq \emptyset$ then go to step **A1** with $i \leftarrow i + 1$ and $j \leftarrow j + 1$, otherwise stop.

The algorithm does not need any preliminary information on the final number K of clusters that one expects. On the contrary, it computes K iteratively in the following way: as long as there exists a point ξ which do not belong to any cluster already determined, the strategy **S1-S2** is applied with $\xi(0) = \xi$ and the number K is incremented.

V. SIMULATIONS

In this section we present the result of the application of the method to two practical cases. In the first one, we consider two flocks of birds, we map each bird into a (set of) pixels of a black-and-white image (see Figure 4 which has been obtained from an original coloured image) and we apply the algorithm to cluster all the measurements (the black pixels) thus identifying the two flocks. The image has 700×961 pixels, 27016 of which are black (each of which potentially corresponds to a bird).

To simplify the computation, the image has been sampled with a decimation rate of 10. The new image has 70×97 pixels, 272 of which are black. Figure 5 shows the set of sampled points with the contours determined according to the algorithm in Section IV.

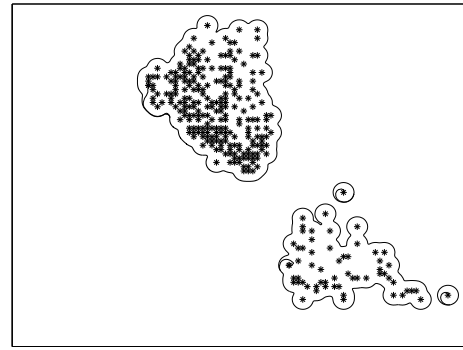


Fig. 5. The result of the algorithm overlapped with the set of sampled measurements.

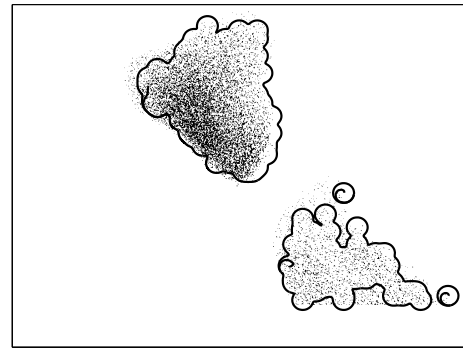


Fig. 6. The result of the algorithm overlapped with the original image of Figure 4.

The algorithm has been performed with a reference level $H_r = 5 \times 10^{-6}$ and has clustered the 272 measured data into 4 groups. Two of them correspond to the original flocks while two contain only one point (one bird). These may correspond to birds which are moving out from (or joining) the flock or may be due to the fact that the sampling cut out some pixels between them and the nearest flock. It can be noted that for each cluster only one iteration of the strategy **S1-S2** has been performed. In particular, in each cluster one can see the piece of trajectory (a small "curl") from the initial point ξ_i to the level line corresponding to H_r .

Finally, in Figure 6 the level line provided by the algorithm has been compared with the original picture. As one can see, the sampling does not affect significantly the clustering.

To show how the selection of the level set can affect the result of the clustering process, we have applied the algorithm to another practical case where the measurements are the black pixels of a black-and-white image of a flock of geese flying "in formation". The algorithm has been applied with two different values of H_r . In the first case (see Figure 7) the algorithm detects the single birds, while in the second case (see Figure 8) the birds are grouped into three clusters.

To have a more specific idea about the dependency of the number of clusters on the reference value H_r , all the values from 0.005 to 0.995, with step 0.005 have been tested. The result is reported in Figure 9.

Note that the values of k_1 and k_2 do not affect the

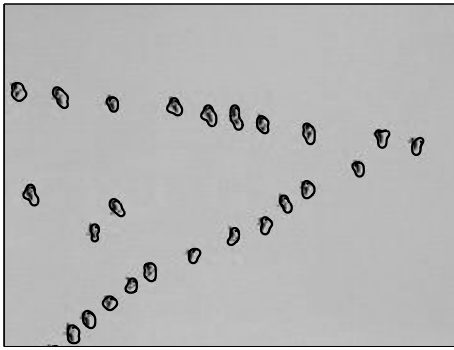


Fig. 7. The result of the algorithm overlapped with the original image of a flock of geese. The pixels corresponding to each single goose are clustered together.

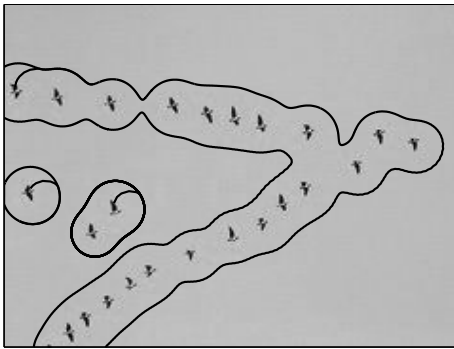


Fig. 8. The result of the algorithm overlapped with the same image of Figure 7 but with a different value of the reference level. In this case the geese are clustered in three groups.

final result of the clustering process but only the speed of execution of the algorithm.

VI. CONCLUSIONS

In this paper we have presented a novel approach to the problem of clustering a set of measurements. The algorithm makes use of the notion of level function and it clusters together measurements internal to a level line. The goal is achieved by considering the level function as a Hamiltonian function and integrating the corresponding Hamiltonian system. Future developments will be devoted to improve the algorithm in order to cope with situations like the one in Figure 10, where the region identified by the level lines is not simply connected. The present version of the algorithm **A1-A3**, in fact, cannot detect both contours.

ACKNOWLEDGEMENTS

The authors would like to thank Neil Cade from SELEX Sensors and Airborne Systems Limited for many useful discussions that contributed to this work.

REFERENCES

- [1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, ser. Morgan Kaufmann. Academic Press, 2001.
- [2] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*. John Wiley and sons, 2001.
- [3] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*. John Wiley and sons, 1973.

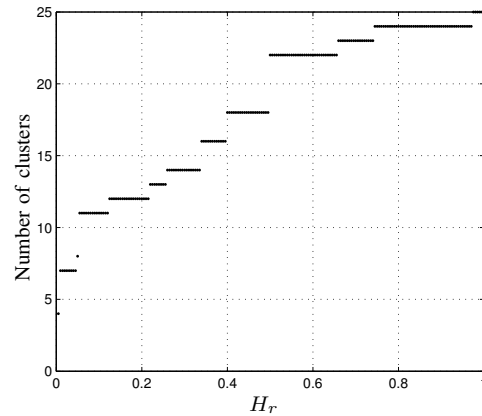


Fig. 9. Dependency of the number of clusters on the value of H_r .

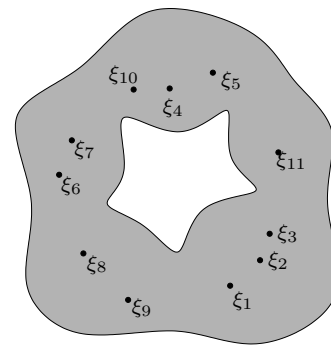


Fig. 10. A non-simply connected region. The algorithm **A1-A2** cannot detect both the contours.

- [4] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic press, 1999.
- [5] Z. Huang, "Extension to the k -means algorithm for clustering data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, pp. 283–304, 1998.
- [6] S.L. Lauritzen, "The em algorithm for graphical association models with missing data," *Computational Statistics and Data Analysis*, vol. 19, pp. 191–201, 1995.
- [7] S.Z. Selim and K. Alsultan, "A simulated annealing algorithm for the clustering problem," *Pattern Recognition*, vol. 24, no. 10, pp. 1003–1008, 1991.
- [8] K. Krishna and M. Narasimha Murty, "Genetic k -means algorithm," *IEEE Transactions on systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 29, no. 3, pp. 433–439, 1999.
- [9] S.M. Pan and K.S. Cheng, "Evolution-based tabu search approach to automatic clustering," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 37, no. 5, pp. 827–838, 2007.
- [10] G. Karypis, E. H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999.
- [11] N.R. Pal and J.C. Bezdek, "On cluster validity for the fuzzy c -means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, 1995.
- [12] X.L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.
- [13] G. V. Nosovskiy, D. Liu, and O. Sourina, "Automatic clustering and boundary detection algorithm based on adaptive influence function," *Pattern Recognition*, 2008, to appear.
- [14] I. Stewart and D. Tall, *Complex Analysis*. Cambridge University Press, 1983.