DISCRIMINATIVE MODEL SELECTION USING A MODIFIED BAYESIAN CRITERION: APPLICATION TO TRAJECTORY MODELING

Jacinto C. Nascimento[a] *Jorge S. Marques*[a] *Mário A. T. Figueiredo*[b]

Instituto de Sistemas e Robótica[a] Instituto de Telecomunicações[b]
Instituto Superior Técnico, 1049-001 Lisboa, Portugal

ABSTRACT

In this paper we introduce a novel method to determine the model order of a stochastic model for moving objects. The main assumption is that we make use of the knowledge that the obtained model is going to be used for some task, specifically, for trajectory classification. Particularly, the object motion is described by trajectories performed by the objects (*e.g.*, pedestrians), during their motion, by representing them by a small and meaningful mixtures of vector fields.

We present a discriminative method for model selection without resort to computationally expensive cross-validation procedures. The idea is, thus, to select the generative model achieving the best classification performance.

Although the topic of application is video surveillance, the proposed method can easily be extended to other practical situations. Experiments with both synthetic and real data concerning pedestrian activities illustrate the performance of the proposed approach.

1. INTRODUCTION

In this paper we are primarily concerned with the task of model selection, which occurs in many statistical inference/learning problems. The model selection problem that must be tackled is that of choosing a parsimonious model order, capable of describing the observed data, but also able to generalize to unobserved data. More specifically, we will focus on model selection for non-rigid trajectories performed by pedestrians in surveillance tasks. The trajectories are described by a set of vector motion fields, as proposed in [1]. Each trajectory can be split into a set of consecutive segments, each of which is generated by one vector field. Switching can occur at any point in the image domain and the switching probabilities depend on the object location. This model provides a flexible tool to represent a wide variety of motion patterns. See Fig. 1 for an illustration.

Naturally, the trajectory model described in the previous paragraph involves a model selection question: how many vector fields should be used? As is well known, the *maximum likelihood* criterion that is used to learn the vector fields and switching matrix, cannot be used to select the order of the model (*i.e.*, the number of fields), as it leads to overfitting. Naturally, an arbitrarily large number of fields allows the model to better explain the training data (trajectories), assigning larger values of the likelihood for the observed data, but yielding models with poor generalization.

Common ways for avoiding overfitting have included early stopping criterion, regularization, or cross-validation. Although, it is possible to use cross-validation for simple searches over model size, *e.g.*, if the search is restricted to a single parameter that controls the complexity of the model, the same can not be said for more general searches over many parameters in which cross-validation is computationally prohibitive.

In this paper, we propose a discriminative method for model selection. Our approach is based on the knowledge that the obtained model is going to be used for a specific task, *i.e.*, classification. The idea is, thus, to select the generative model achieving the best classification performance.

2. RELATED WORK

Model selection can be addressed using Bayesian tools, starting with some prior knowledge or assumptions about the model structure. However, fully Bayesian approaches to model selection are computationally expensive, as they require high-dimensional integrations over model parameters and latent variables; these integrations can only be performed exactly and cheaply in some special cases (conjugate priors), otherwise demanding Monte Carlos approaches. Much work has been devoted to obtaining simplified/approximate versions of Bayesian model selection, such as using the Laplace approximation; for a comprehensive review of this area, see [2]. One such approximate Bayesian method is Schwarz's *Bayesian inference criterion* (BIC) [3]. Another class of approaches to approximate Bayesian inference includes the so-called *variational Bayes* (VB) methods [4],[5].

Discriminative model selection criteria have also been proposed, *e.g.* [6], when the ultimate goal is to identify the model configuration which provides the highest classification accuracy. Our approach follows this rationale, where the underlying task is to maximize the performance of the classifier. Our discriminative model is characterized to have a grid-wise vector fields and transition matrices, as proposed in [1], *i.e*, each point in the grid has its own vector field a transition matrix, which are, in general, different within grid points (see Fig. 1). This suggests a large number of parameters to be estimated. In this context, the so-called model selection criteria such as BIC or AIC can not be directly used, in the sense that these criteria require the distribution of the data to have a known parametric form, which is not available in the presented case. Recall that the number of parameters to estimate depends on the grid nodes which is certainly large. We face this quagmire by proposing a discriminative model selection. The approach we follow is to split the data into disjoint training and test sets, and select the model parameters in the training set that minimizes the classification error on an independent validation (test) set.

3. ACTIVITY REPRESENTATION

This section describes how trajectories are modeled using switching vector fields.

We define a trajectory as a length-n sequence of positions of the person 'mass center' in the image, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ with $\mathbf{x}_t \in R^2$ and it is generated by a set of m velocity vector fields $\mathcal{T} =$

This work was supported by FCT (ISR/IST plurianual funding) through the PIDDAC Program funds, and the ARGUS project PTDC/EEA-CRO/098550/2008.

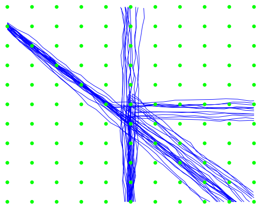

Fig. 1. Several trajectories with switching motion fields (blue lines) in a grid of 11×11 nodes (green dots).

$\{\mathbf{T}_1, \ldots, \mathbf{T}_m\}$, with $\mathbf{T}_k : R^2 \rightarrow R^2$, for $k \in \{1, \ldots, m\}$. The velocity vector at point $\mathbf{x} \in R^2$ of the k-th field is denoted as $\mathbf{T}_k(\mathbf{x})$. At each time instant, one of the velocity fields is *active*, *i.e.*, is driving the motion. The object trajectory is, thus, generated by

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{T}_{k_t}(\mathbf{x}_{t-1}) + \mathbf{w}_t, \quad t = 2, ..., n, \tag{1}$$

where $k_t \in \{1, ..., K\}$ is the label of the active field at time t, $\mathbf{w}_t \sim \mathcal{N}(0, \sigma^2_{k_t}\mathbf{I})$ is white Gaussian noise with zero mean and variance $\sigma^2_{k_t}$, and n is the length (number of points) of the trajectory. The initial position follows a known distribution $p(\mathbf{x}_1)$; $\{k_t\}$ is assumed to be a first order Markov sequence characterized by a space varying transition matrix $\mathbf{B}(\mathbf{x})$. The model parameters are defined by a triplet $\boldsymbol{\theta} = (\mathcal{T}, \boldsymbol{\sigma}, \mathbf{B})$ which includes the parameters of the motion fields $\mathcal{T}$, noise standard deviations $\boldsymbol{\sigma}$ and space-varying switching matrix $\mathbf{B}$.

Different types of activities may occur in a scene *e.g.*, people walking, entering or leaving. In many cases, the type of activity can be inferred from the trajectory. We will therefore assume that each activity is represented by a set trajectory patterns to be learned from the video. We will represent each activity j by a different dynamical model with a triplet of model parameters $\boldsymbol{\theta}^{(j)}_{m_j} = (\mathcal{T}^{(j)}_{m_j}, \mathbf{B}^{(j)}_{m_j}, \boldsymbol{\sigma}^{(j)}_{m_j})$ where m_j denotes the number of active fields used to represent the j-th activity. The estimation of the model parameters from the training trajectories is detailed in [1] assuming that the number of fields is known. We will now address the estimation of the number of fields m_j for each type of activity.

4. MODEL SELECTION ALGORITHM

The goal is to estimate the model order m_j for all activity-classes, *i.e.*, for $j \in \{1, ..., J\}$. To accomplish this, we will use a labeled training set $D = \{D^{(1)}, \ldots, D^{(J)}\}$, where $D^{(j)}$ is the set of trajectories associated to the j-th class, *i.e.*, $D^{(j)} = \{\mathbf{x}^{(j)}_1, \ldots, \mathbf{x}^{(j)}_l\}$, where $\mathbf{x}^{(j)}_l$ is a trajectory described by (1).

One possible way to determine the model order is considering all possible combinations of m_j assuming that $1 \leq m_j \leq M$, where M is the maximum number of vector fields as proposed in [7]. However, this approach leads to a M^J possible classifications. In this paper, we follow a better approach which allows to reduce the complexity (number of classifications) maintaining the performance as in [7].

The new algorithm proposed herein is based on the BIC criterion in which we vary a γ parameter. The BIC approximation is twofold: (i) it does not depend on the prior; and (ii) it does not take into account the local geometry of the parameter space and hence is invariant to reparameterisations of the model. In this work, the BIC is simply obtained

$$\text{BIC}(\mathcal{L}, \gamma) = -2\mathcal{L} + \gamma(\mu \log \mu) \tag{2}$$

where $\mathcal{L}$ is the log-likelihood and $\mu = md + (md + m^2)N^2$; m is the number of vector fields, $d = 2$ is the dimension of the 2D motion filed, N^2 is the number of grid nodes (N=11). γ is the new parameter that is defined in an interval, $\gamma \in R_\gamma = [0, 1]$, *i.e.* the "range of γ". Thus, in the new framework, we propose a quaternion of model parameters $\Gamma = (\boldsymbol{\theta}, \gamma)$.

In the training stage, we first take the set $D^{(j)}$ and compute the parameters set $\boldsymbol{\theta}^{(j)}_{m_j}$ and the corresponding log-likelihoods $\mathcal{L}^{(j)}_{m_j}$. We repeat this procedure for all models within the j-th class activity, and then for all activities, thus obtaining the duplet $(\boldsymbol{\theta}^{(j)}_{m_j}, \mathcal{L}^{(j)}_{m_j})$ for $j \in \{1, ..., J\}$ and $m \in \{1, ..., M\}$. Secondly, for a given value of γ, we compute the BIC criterion for each activity, varying the number of models, *i.e*, $\text{BIC}(\mathcal{L}^{(j)}_{m_j}, \gamma)$, with $m_j = 1, ..., M$ and $j = 1, ..., J$. From this set, we determine for each class, the best model order, say m^*_j, with BIC. We repeat this procedure for the range of γ, R_γ.

The synopsis of the training stage for each activity (the variable j is fixed) is described next.

Algorithm 1 Modified Bayesian criterion.

- Compute the parameters and log-likelihood sets $\left\{\boldsymbol{\theta}^{(j)}_{m_j}, \mathcal{L}^{(j)}_{m_j}\right\}$, for $m_j = 1, ..., M$ and $j = 1, ..., J$ using the EM algorithm,

for $\gamma \in [\gamma_{\min}, ..., \gamma_{\text{Max}}]$ **do**
- Compute $\text{BIC}(\mathcal{L}^{(j)}_{m_j}, \gamma)$, for $m_j = 1, ..., M$ and $j = 1, ..., J$
- For each j-th class activity, compute the best parameters set $\boldsymbol{\theta}^{(j)}_{m^*_j}$, such that $m^*_j = \arg\min_{m_j} \text{BIC}(\mathcal{L}^{(j)}_{m_j}, \gamma)$

end for

For each value of $\gamma \in [\gamma_{\min}, ..., \gamma_{\text{Max}}]$ the raining procedure above described allows to obtain the following parameters set:

$$\boldsymbol{\Theta} = \left\{\boldsymbol{\theta}^{(j)}_{m^*_j}\right\}, \quad \text{with} \;\; j = 1, ..., J \tag{3}$$

where $\sharp\boldsymbol{\Theta} = J \times \sharp R_\gamma$.

Now assuming, in addition to the training set, we have a *selection set*, with the trajectories from all classes $\mathcal{D} = (\mathcal{D}^{(1)}, ..., \mathcal{D}^{(J)})$, where $\mathcal{D}^{(j)} = \{\mathbf{x}^{(j)}_1, ..., \mathbf{x}^{(j)}_{D_j}\}$ denotes a set of D_j trajectories from class j.

The classification procedure is taken on the set $\mathcal{D}$ with the estimated parameters $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta}^{(j)}_{m^*_j}\right\}$, with $j = 1, ..., J$ and for each value of $\gamma \in R_\gamma$. In this way, the proposed strategy allows to reduce the number of classifications from M^J to simply R_γ [1], *i.e.* the range of γ (see (3)).

The performance of each classification, *i.e.*, for each value of $\gamma \in [\gamma_{\min}, ..., \gamma_{\text{Max}}]$, is obtained by evaluating the overall classification over the *selection set*. The number of misclassifications $\mathcal{M}(\gamma, \mathcal{D})$ is computed as

$$\mathcal{M}(\gamma, \mathcal{D}) = \sum_{j=1}^{J} \sum_{l=1}^{D_j} e^{(j)}_l(\gamma) \tag{4}$$

$$e^{(j)}_l(\gamma) = \delta(j, \arg\max_r(p(\mathbf{x}^{(j)}_l | \boldsymbol{\theta}^{(r)}_{\mathbf{K}})))$$

where $r \in \{1, ..., J\}$; $p(\mathbf{x}^{(j)}_l | \boldsymbol{\theta}^{(r)}_{\mathbf{K}})$ is the likelihood of the trajectory $\mathbf{x}^{(j)}_l$ given the r-th model determined by BIC and for a given value γ; $\mathbf{K}$ is the model configuration $\mathbf{K} = \{m^*_1, ..., m^*_J\}$ containing the best model order for all activities and $\delta(a, b)$ is a Kronecker symbol, *i.e.* $\delta(a, b) = 1$, if $a = b$, and zero otherwise.

Recall that, the error criterion (4) may be minimized for several values of γ. Thus, in addition to (4) we also take into account the model complexity

$$\widetilde{\mathcal{M}}(\gamma, \mathcal{D}) = \mathcal{M}(\gamma, \mathcal{D}) + \mathcal{C}(\gamma) \tag{5}$$

[1] Recall that cardinality of the R_γ corresponds to the number of classifications. In this paper we set this to a small number of ten

where $\mathcal{C}(\gamma)$ is the cost associated with a particular model γ defined as the sum of m_j. The minimization of (5) is the *discriminative model selection* criterion proposed herein.

5. EXPERIMENTS

In this section we provide results in both synthetic and real scenarios. In both scenarios, a comparison between the proposed method and the one in [7] is presented. In the synthetic example the equivalence of the two methods is illustrated. In the real case, the presented example allows to focus how the method avoids the exponential growth of classifications, maintaining at the same time equivalent results.

5.1. Synthetic data

We consider a synthetic example which contains two type of trajectories which resemble trajectories performed by pedestrians (see Fig. 2 top left). The trajectories are generated according to the HMM vector field (1). The trajectories presented in this example are classified into two activities: "*straight*" (green lines) which contains a single vector field, and "*spread*" (red lines) which contains three different vector fields.

Four experiments are performed (see Fig. 2) to illustrate te robustness of the model with respect to model mismatch. For each experiment, we generated 100 training and testing trajectories. For the training trajectories we set $\sigma^2_{\text{trn}} = 1e-3$, for the testing set we used the following values for the dynamic noise: (i) $\sigma^2_{\text{tst}} = \sigma^2_{\text{trn}}$ (Fig. 2 top left); (ii) $\sigma^2_{\text{tst}} = 5\sigma^2_{\text{trn}}$ (Fig. 2 top right); (iii) $\sigma^2_{\text{tst}} = 10\sigma^2_{\text{trn}}$ (Fig. 2 bottom left); and (iv) $\sigma^2_{\text{tst}} = 20\sigma^2_{\text{trn}}$ (Fig. 2 bottom right).

The first step is to determine the classes model order for each value of γ. To accomplish this, we varied the number of models from one to three ($M = 3$) for each γ value. This is done independently for each class and for $\sigma^2_{\text{trn}} = 1e-3$ (*i.e.* the smallest value), since we want to figure out the intrinsic model order for each class. Fig. 3 (top left) shows the models number determined using the BIC criterion, for the different values of $\gamma \in [0.1, 0.9]$.

Fig. 3 (top right) shows the classification accuracy for all the experiments and for $\gamma \in [0.1, 0.9]$. It can be concluded that the best accuracy is obtained in the interval $\gamma \in [0.1, 0.4]$. The most parsimonious choice in this range corresponds to have:"*spread*" class $\rightarrow$ 2 vector fields and "*straight*" class $\rightarrow$ 1 vector fields (see top left of the Fig. 2).

We also present the results of discriminative approach (see [7]). In this case, 9 different classifications (as in the proposed method) are possible. The labels in the X-axis of the Fig. 3 (bottom), "N_1N_2", means the number of models for the class "1" ("spread" trajectories) and for class "2" ("straight" trajectories), respectively. From this figure it is shown that the accuracy improves from the configurations "1X" to "2X" (X - stands for any number of models). This means that the first activity is suitably described by using at least two models. The most un-expensive configuration is that of "21". This is the same result provided by the proposed method.

5.2. Real data

In this section we present results in a real settings concerning different and typical activities/trajectories that unfold in a shopping scenario. Typical trajectories in this scenario are "entering", "leaving" or "passing" in the front of a shopping mall. Fig. 4 illustrates the scenario after a homography transformation in which the trajectories are superimposed with yellow dots. Three different classes are considered: (i) *converge*, (ii)*spread* and (iii) *passing*. The trajectories of each class concern the following:

- *converge*: It contains all the trajectories in which the person enters into the mall. The pedestrian may enter from the left or the right side of the scenario. We call it converge, since all trajectories converge at the shop entrance (Fig. 4 left).
- *spread*: This class contains the trajectories in which the person leaves the mall and it has the opposite direction of the previous class. People can leave to the right or left side of the scenario, say that, the trajectories spread or diverge at the entrance of the mall (Fig. 4 center).
- *passing*: Here, all the trajectories concern the situations in which the pedestrian passes in the front of the mall in two opposite directions (Fig. 4 right).

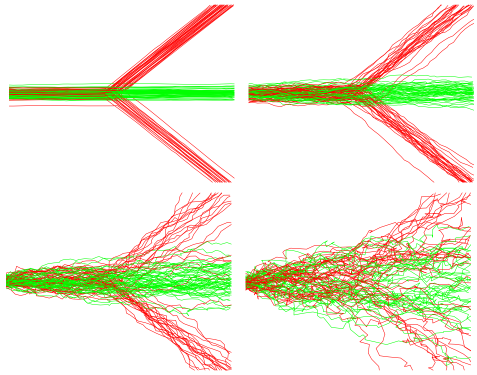

Fig. 2. Synthetic examples of trajectories concerning two activities marked with different colors: "spread" (in red), "straight" (in green) with $\sigma^2_{\text{tst}} = \sigma^2_{\text{trn}}$ (top left), $\sigma^2_{\text{tst}} = 5\sigma^2_{\text{trn}}$ (top right), $\sigma^2_{\text{tst}} = 10\sigma^2_{\text{trn}}$ (bottom left) $\sigma^2_{\text{tst}} = 20\sigma^2_{\text{trn}}$ (bottom right).

All the trajectories exhibit a non-smooth direction, and some of them contain gaps due to the failures in tracking which is based on foreground region detection as in [8].

Fig. 5 (top row) shows the results of the proposed method. The results on the left were obtained using a training set to determine the number of models per class for each value of γ. On the right it is shown the classification accuracy on a disjoint test set. It can be easily concluded that the classification is highest for the three first values of γ. From these, the most un-expensive configuration is that of (see Fig. 5 top left): *converge* $\rightarrow$ 1 model, *diverge* $\rightarrow$ 2 models, and passing $\rightarrow$ 2 models.

Fig. 5 (bottom) shows the results of discriminative approach without restrictions proposed in [7]. We see that $M^J = 27$ classifications are needed, whereas in the proposed method just $\sharp R_\gamma = 9$ classifications are enough. The X-axis of the Fig. 5 (bottom), contains the configurations of the number of models (vector fields) for the classes in the following order: "*converge*", "*spread*", "*passing*". Thus, "1" means the configuration (111), and the "27" corresponds to the configuration (333). From the graphic, several configurations are possible which alow to obtain the maximum classification score (12 configurations). From these, we select the one that uses the minimum number of models which is the configuration 5 or (122). This is precisely the result already obtained with the proposed algorithm.

6. CONCLUSIONS

We have presented a discriminative method for model selection. The approach presented herein avoids to consider all possible combinations of the activities vs. number of fields reducing the complexity from $O(m^2)$ to simply $O(m)$ maintaining equivalent results.

We applied the proposed framework to surveillance scenarios, where observed data are trajectories. Further work will address this methodology for the recognition of human activities and vehicle traffic using the presented framework. These examples often contain a

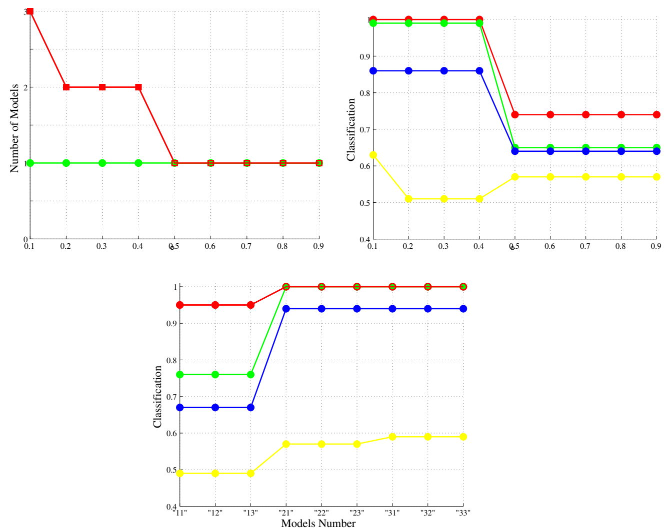

Fig. 3. Results with the proposed method (top row) and the method in [7] (bottom row). Top left: model order for spread class (red) and straight class (green). Top right: classification accuracy of the proposed method. Bottom: classification accuracy of the method in [7]. For the classifications, each color denotes a different experiment: $\sigma^2_{\text{tst}} = \sigma^2_{\text{trn}}$ (red), $\sigma^2_{\text{tst}} = 5\sigma^2_{\text{trn}}$ (green), $\sigma^2_{\text{tst}} = 10\sigma^2_{\text{trn}}$ (blue) $\sigma^2_{\text{tst}} = 20\sigma^2_{\text{trn}}$ (yellow).

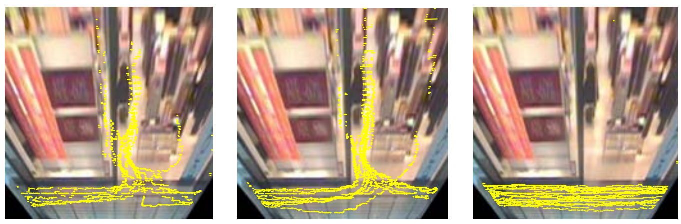

Fig. 4. Tracked trajectories in the shopping mall. From left to right: converge, diverge and passing classes.

much large number of classes. The presented framework will be suitable, since it reduces the number of training/test experiments, and it reduces the computational burden. Also, extensions of the proposed method including approaches such as variational based or sampling methods will be addressed in the future.

7. REFERENCES

[1] J. C. Nascimento, M. A. T. Figueiredo, and J. S. Marques, "Vector fields estimation for motion in natural images," in *Proc. of the IEEE Int. Conf. on Imag. Proc.*, 2009.

[2] A. D. Lanterman, "Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection," *International Statistical Review*, vol. 69, no. 2, pp. 185–212, 2000.

[3] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[4] M. J. Beal, "Variational Algorithms for Approximate Bayesian Inference," Ph.D. dissertation, University College London, 2003.

[5] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data, with application to scoring graphical model structures," in *Bayesian Statistics 7*. Oxford University Press, 2003.

[6] B. Thiesson and C. Meek, "Discriminative model selection for density models," 2003.

[7] J. C. Nascimento, M. A. T. Figueiredo, and J. S. Marques, "Discriminative model selection for object motion recognition," in *Proc. of the IEEE Int. Conf. on Imag. Proc.*, 2010.

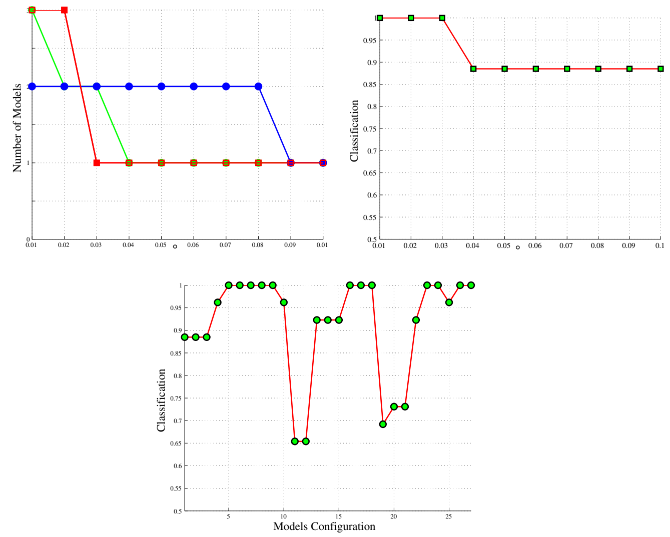

Fig. 5. Results with the proposed method (top row) and the method in [7] (bottom row). Top left: model order for converge class (red), diverge class (green) and passing class (blue). Top right: classification accuracy of the proposed method. Bottom: classification accuracy of the method in [7].

[8] T. Boult, R. Micheals, X. Gao, and M. Eckmann, "Into the woods: Visual surveillance of non-cooperative camouflaged targets in complex outdoor settings," in *Proceedings of the IEEE*, October 2001, pp. 1382–1402.