ACTIVE CLASSIFICATION FOR HUMAN ACTION RECOGNITION

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

Depeartment of Informatics, Aristotle University of Thessaloniki, Greece
{aiosif,tefas,pitas}@aiia.csd.auth.gr

ABSTRACT

In this paper, we propose a novel classification method involving two processing steps. Given a test sample, the training data residing to its neighborhood are determined. Classification is performed by a Single-hidden Layer Feedforward Neural network exploiting labeling information of the training data appearing in the test sample neighborhood and using the rest training data as unlabeled. By following this approach, the proposed classification method focuses the classification problem on the training data that are more similar to the test sample under consideration and exploits information concerning to the training set structure. Compared to both static classification exploiting all the available training data and dynamic classification involving data selection for classification, the proposed active classification method provides enhanced classification performance in two publicly available action recognition databases.

Index Terms— Active classification, dynamic classification, human action recognition, Single-hidden Layer Feedforward Neural network, Extreme Learning Machine

1. INTRODUCTION

Supervised classification methods can be categorized depending on the way they utilize the available training data, in static and dynamic ones. Static classification methods employ all the available training data, and the corresponding class labels, in order to train a (universal) classification model, that will be used in order to classify any (unknown) test sample. Dynamic classification methods involve a model adaptation process based on the test sample to be classified. It has been shown that, by exploiting the information appearing in the test sample under consideration, dynamic classification methods can provide enhanced classification performance, compared to the static ones.

A dynamic classification method exploiting sparsity constraints has been proposed in [1]. A given test sample is involved in a L1-minimization-based class-independent regression process by using an overcomplete dictionary formed by all the available training data. Multiple reconstruction samples are, subsequently, produced by exploiting the reconstruction weights corresponding to each class independently and the test sample under consideration is classified based on the minimum reconstruction error classification rule. The Dynamic Committee Machine (DCM) has been proposed in [2]. DCM employs five state-of-the-art classifiers in order to determine five classification results for a given test sample. The obtained classification results are, finally, fused by using test sample-specific combination weights. A dynamic classification scheme has been proposed in [3] for human action recognition. The classification process involved person identification and action classification based on a classifier trained by using training data of the recognized person. A dynamic classification method involving training data selection and Linear Discriminant Analysis (LDA)-based data classification is proposed in [4]. The procedure used in order to determine an appropriate training set for LDA-based data projection and classification is intuitive and effective. However, the LDA-based classification approach in this setting sets the assumption of linearly separable classes and is prone to the Small Sample Size problem relating to statistical learning models [5]. In order to overcome these drawbacks, the method has been extended so as to exploit an Artificial Neural Network-based non-linear classification scheme [6]. A dynamic classification scheme involving optimization-based feature space partitioning has been proposed in [7]. In [7], multiple linear classifiers, each performing on a feature space region, are learned, while a test sample is classified by the classifier responsible for the corresponding region.

In this paper, we propose a classification method inspired by relevant work in Active Vision [8, 9]. Motivated by the fact that the structure of the human retina is such that only a small neighborhood around the fixation point [10] is captured in high resolution by the fovea, while the rest of the scene is captured in lower resolution by the sensors in the periphery of retina, it has been shown [11, 12, 13] that by following an Active Vision-based approach several computer vision tasks, like image segmentation and motion estimation, can be better described and addressed. Based on this fact, we investigate the applicability of this approach to classification problems. To this end, we propose a dynamic, noted as active hereafter, classification method involving two processing steps. Given a test sample, which is considered to be a fixation point in a high-dimensional feature space, we determine the training data appearing in its neighborhood by exploiting its similarity

with all the available training data. We, subsequently, perform semi-supervised classification by employing all the available training data and the class labels of the training data appearing in the test sample neighborhood. By following this approach, the proposed classification method focuses the classification problem on the training data that are more similar to the test sample under consideration. In addition, it exploits information concerning to the training set structure, which is lost in other dynamic classification schemes involving labeled data selection for classification [3, 4, 6].

The proposed method employs the, recently proposed, semi-supervised Extreme Learning Machine (SELM) algorithm [14] for Single-hidden Layer Feedforward Neural (SLFN) network training and is evaluated in human action recognition exploiting the, recently proposed, Action Bank [15] action video representation. It should be noted though, that the same methodology can be applied by employing other semi-supervised classification schemes. Compared to both static classification exploiting all the available training data and dynamic classification involving data selection for classification, the proposed method provides enhanced classification performance in two publicly available databases.

The paper is structured as follows. The proposed method is described in Sections 2. Section 3 presents experiments conducted in order to evaluate its performance. Finally, conclusions are drawn in Section 4.

2. PROPOSED METHOD

As it has been previously described, the proposed method involves training data selection and semi-supervised classification. We describe these two processing steps in Subsections 2.1 and 2.2, respectively. We, subsequently, describe the proposed active classification method in Subsection 2.3.

2.1. Data Selection

Let us denote by $\mathcal{U}$ a vector database containing vectors $\mathbf{v}_i \in \mathbb{R}^D$, $i = 1, \ldots, N$, each belonging to one of the C classes forming a class set $\mathcal{C}$. We denote the class label of vector $\mathbf{v}_i$ by using c_i. Let us, also, assume that a test sample is represented by the corresponding test vector $\mathbf{v}_t \in \mathbb{R}^D$. We would like to determine the l training vectors that reside to the test vector neighborhood in the high-dimensional feature space $\mathbb{R}^D$. To this end, we calculate the similarity of $\mathbf{v}_t$ with all the training vectors $\mathbf{v}_i$, i.e.,:

$$s_i = \|\mathbf{v}_i - \mathbf{v}_t\|_2^{-1}. \quad (1)$$

The obtained similarity values are sorted in a descending order and the l training vectors that reside to $\mathbf{v}_t$ neighborhood are the ones providing the l highest similarity values. In our experiments, l is automatically determined by $l = N/K$, where K is a user specified parameter value.

Alternatively, one may cluster the training vectors $\mathbf{v}_i$ in K groups, e.g., by applying K-Means algorithm, and select the training vectors belonging to the group where $\mathbf{s}_t$ belongs to, similar to [4, 6]. This approach has the advantages that the number of selected training data l is dynamically determined by the test vector under consideration and that the training set can be clustered in an offline stage, leading to faster classification. However, in the cases where the test vector $\mathbf{v}_t$ is far from the corresponding group center, this approach would not result to optimal training data selection.

2.2. SELM-based classification

Let us denote by $\mathbf{v}_i$, $i = 1, \ldots, l, l+1, \ldots, l+u$ the training vectors that will be used in order to classify $\mathbf{v}_t$. $u = N - l$ is the number of the training vectors that do not reside to $\mathbf{v}_t$ neighborhood. Here we assume that the training vectors have been ordered by using the similarity values s_i in (1). We would like to employ $\mathbf{v}_i$, $i = 1, \ldots, N$ and the class labels c_i, $i = 1, \ldots, l$ corresponding to the training vectors residing to $\mathbf{v}_t$ neighborhood in order to train a SLFN network for $\mathbf{v}_t$ classification. We employ the SELM algorithm [14] to this end.

In SELM, the network input weights $\mathbf{W}_{in} \in \mathbb{R}^{D \times H}$ and the hidden layer bias values $\mathbf{b} \in \mathbb{R}^H$ are randomly assigned, while the network output weights $\mathbf{W}_{out} \in \mathbb{R}^{H \times C}$ are analytically calculated. H refers to the number of neurons forming the network hidden layer. The network target vectors $\mathbf{t}_i \in \mathbb{R}^C$ are set to $t_{ij} = 1$ for $c_i = j$, $t_{ij} = -1$ for $c_i \neq j$ and $t_{ij} = 0$ otherwise. Let $\mathbf{w}_j$, $\mathbf{u}_k$ and u_{kj} denote the j-th column of $\mathbf{W}_{in}$, the k-th row of $\mathbf{W}_{out}$ and the j-th element of $\mathbf{u}_k$, respectively. For a given hidden layer activation function $\Phi()$, the output $\mathbf{o}_i = [o_{i1}, \ldots, o_{iC}]^T$ of the SELM network corresponding to training action vector $\mathbf{v}_i$ is calculated by:

$$o_{ik} = \sum_{j=1}^{H} u_{kj}\, \Phi(\mathbf{w}_j, b_j, \mathbf{v}_i), \quad k = 1, ..., C. \quad (2)$$

Many activation functions $\Phi()$ can be employed for the calculation of the hidden layer output, such as sigmoid, sine, Gaussian, hard-limiting and Radial Basis function (RBF). The most popular choice is the sigmoid function, i.e.:

$$\Phi_{sigmoid}(\mathbf{w}_j, b_j, \mathbf{v}_i) = \frac{1}{1 + e^{-(\mathbf{w}_j^T\, \mathbf{v}_i + b_j)}}, \quad (3)$$

By storing the hidden layer neuron outputs in a matrix $\mathbf{\Phi}$:

$$\mathbf{\Phi} = \begin{bmatrix} \Phi(\mathbf{w}_1, b_1, \mathbf{v}_1) & \cdots & \Phi(\mathbf{w}_1, b_1, \mathbf{v}_N) \\ \ldots & \ddots & \ldots \\ \Phi(\mathbf{w}_H, b_H, \mathbf{v}_1) & \cdots & \Phi(\mathbf{w}_H, b_H, \mathbf{v}_N) \end{bmatrix}, \quad (4)$$

SELM solves the following optimization problem:

$$\mathbf{W}_{out} = \underset{\mathbf{W}_{out}}{argmin} \|\mathbf{W}_{out}^T \mathbf{\Phi} - \mathbf{T}\|_F,$$
$$s.t.: \quad \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \left(\mathbf{W}_{out}^T \boldsymbol{\phi}_i - \mathbf{W}_{out}^T \boldsymbol{\phi}_j\right)^2 = 0, \quad (5)$$

where $\boldsymbol{\phi}_i$ is the i-th column of $\mathbf{\Phi}$, i.e., the network hidden layer output for $\mathbf{v}_i$, $\mathbf{T} \in \mathbb{R}^{C \times N}$ is a matrix containing the network target vectors $\mathbf{t}_i$ and w_{ij} is a value denoting the similarity between $\boldsymbol{\phi}_i$ and $\boldsymbol{\phi}_j$.

By solving (5), $\mathbf{W}_{out}$ is given by:

$$\mathbf{W}_{out} = \left(\left(\mathbf{J} + \lambda \mathbf{L}^T\right) \mathbf{\Phi}\right)^{\dagger} \mathbf{J}\mathbf{T}^T, \quad (6)$$

where $\mathbf{J} = diag(1, 1, \ldots, 0, 0)$ with the first l diagonal entries as 1 and the rest 0, $\mathbf{L}$ is the Graph Laplacian matrix [16] encoding the similarity between the training vectors and $\mathbf{A}^{\dagger} = \left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{A}$ is the Moore-Penrose pseudoinverse of $\mathbf{A}^T$.

2.3. Dynamic Classification

Let us assume that a test sample is represented by the corresponding test vector $\mathbf{v}_t$. We employ $\mathbf{v}_t$ in order to determine the l training vectors that reside to its neighborhood by following the procedure described in subsection 2.1. We, subsequently, train a SLFN network by exploiting all the N training vectors and the class labels corresponding to the training vectors residing to $\mathbf{v}_t$ neighborhood by following the procedure described in subsection 2.2. Finally, $\mathbf{v}_t$ is introduced to the trained SLFN network and is classified to the class corresponding to the maximal network output, i.e.,:

$$c_t = \underset{j}{argmax}\, o_{tj}, \;\; j = 1, \ldots, C. \quad (7)$$

3. EXPERIMENTS

In this Section we present experiments conducted in order to evaluate the proposed active classification method. We conducted experiments on the KTH action database [17] containing daily actions and the UCF sports action database [18] containing actions appearing in sports. We provide a comprehensive description of these databases in the following. In all the presented experiments we employ the, recently proposed, Action Bank [15] action video representation. We compare the performance of the proposed dynamic classification method with that of static classification using all the available training vectors and with dynamic classification based on training data selection, similar to [4, 6]. We, finally, compare the performance of the proposed action classification scheme with that of other methods proposed in the literature evaluating their performance on the KTH and the UCF sports databases. Regarding the parameter values used in the presented experiments, the values $K = 10$, $H = 1000$ have been used, while the optimal parameter λ value has been determined by following a grid search strategy and using values $\lambda = 10^r$, $r = -3, \ldots, 3$.

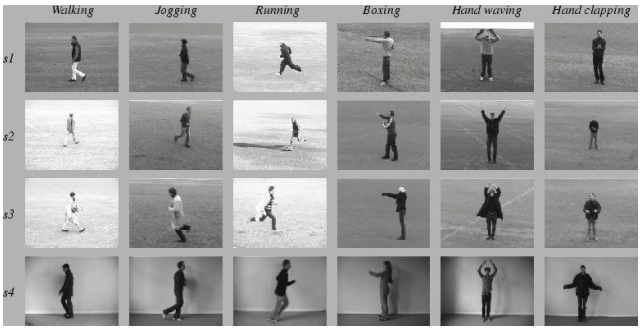

Fig. 1. *Video frames of the KTH action database for the four different scenarios.*

3.0.1. The KTH action database

The KTH action database [17] consists of 600 low-resolution (120×160 pixel) videos depicting 25 persons, performing six actions each. The actions appearing in the database are: 'walking', 'jogging', 'running', 'boxing', 'hand waving' and 'hand clapping'. Four different scenarios have been recorded: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4), as illustrated Figure 1. The most widely adopted experimental setting on this data set is based on a split (16 training and 9 test persons) that has been used in [17].

3.0.2. The UCF sports action database

The UCF sports action database [18] consists of 150 low-resolution (720×480 pixel) videos depicting actions collected from ten sports, which are typically featured on broadcast television channels, such as the BBC and ESPN. The actions appearing in the database are: 'diving', 'golf swinging', 'kicking', 'lifting', 'horse riding', 'running', 'skating', 'bench swinging', 'swinging' and 'walking'. The videos were obtained from a wide range of stock footage websites including BBC Motion gallery and GettyImages. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. The Leave-One-Video-Out cross-validation procedure is used by most action recognition methods evaluating their performance on this data set. Example video frames are illustrated in Figure 2.

3.0.3. Experimental Results

Table 1 illustrates the classification rates obtained by applying the three competing algorithms on the KTH and the UCF

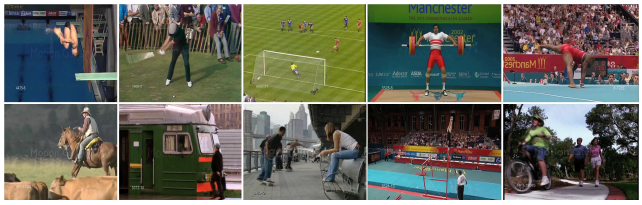

Fig. 2. *Video frames of the UCF sports action database.*

Table 1. Comparison results with static and dynamic classification schemes.

	KTH	UCF sports
Static	91.7%	90.24%
Dynamic	95.83%	93.42%
Active	**97.7%**	**95%**

sports action databases. As can be seen, dynamic classification results to enhanced classification performance, compared to the static classification case. Furthermore, it can be seen that the proposed active classification method, by incorporating information concerning to the action classes structure in the learning process, further increases the classification performance for both databases providing 97.7% and 95% in the KTH and the UCF sports databases, respectively. The corresponding confusion matrices are illustrated in Figures 3 and 4.

Finally, we compare the performance of the proposed action classification scheme with that of other methods evaluating their performance on the KTH and the UCF sports databases in Table 2. As can be seen, the proposed action classification scheme compares favorably with other state of the art methods, recently proposed in the literature.

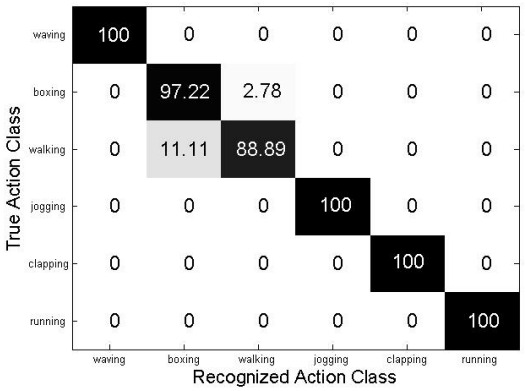

Fig. 3. *Confusion matrices on the KTH database.*

Table 2. Comparison results on the KTH and UCF sports action databases with other methods.

	KTH	UCF sports
Method [19]	94.3%	-
Method [20]	94.5%	-
Method [21]	-	85.2%
Method [22]	-	85.6%
Method [23]	93.9%	86.5%
Method [24]	94.5%	87.3%
Method [25]	94.5%	91.3%
Method [15]	**98.2%**	**95%**
Proposed Scheme	**97.7%**	**95%**

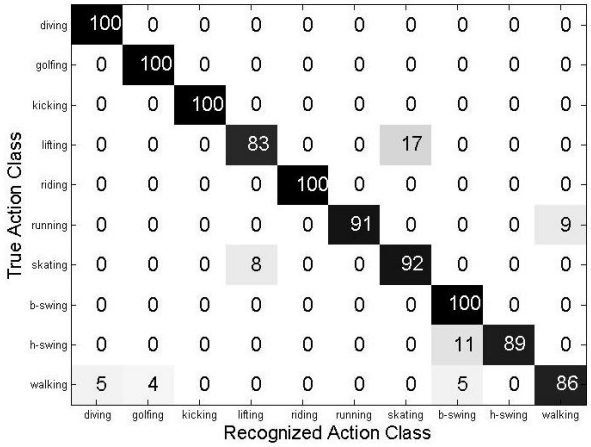

Fig. 4. *Confusion matrices on the UCF sports database.*

4. CONCLUSION

In this paper we proposed a novel classification method inspired by relevant work in Active Vision. The proposed method focusses the classification problem on the training data that are more similar to the test sample under consideration and exploits information concerning to the training set structure. Compared to both static classification exploiting all the available training data and dynamic classification involving data selection for classification, the proposed active classification method provides enhanced classification performance in two publicly available action recognition databases.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 316564 (IMPART). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

5. REFERENCES

[1] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[2] H.M. Tang, M.R. Lyu, and I. King, "Face recognition committee machines: dynamic vs. static structures," in *International Conference on Image Analysis and Processing*, 2003, pp. 121–126.

[3] A. Iosifidis, A. Tefas, and I. Pitas, "Person specific activity recognition using fuzzy learning and discriminant analysis," *European Signal Processing Conference*, pp. 1974–1978, 2011.

[4] M. Kyperountas, A. Tefas, and I. Pitas, "Dynamic training using multistage clustering for face recognition," *Pattern Recognition*, vol. 41, no. 3, pp. 894–905, 2008.

[5] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face recognition using lda-based algorithms," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195–200, 2003.

[6] A. Iosifidis, A. Tefas, and I. Pitas, "Dynamic action recognition based on dynemes and extreme learning machine," *Pattern Recognition Letters*, p. in press, 2013.

[7] "Local supervised learning through space partitioning, author =," .

[8] Y. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *International Journal of Computer Vision*, vol. 48, no. 1, pp. 333–356, 1988.

[9] R. Bajcsy, "Active perception," *International Journal of Computer Vision*, vol. 48, no. 8, pp. 966–1005, 1988.

[10] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 4, pp. 1395–1407, 2006.

[11] A. Mishra, Y. Aloimonos, and C.L. Fah, "Active segmentation with fixation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 468–475, 2009.

[12] K. Daniilidis, "Fixation simplifies 3d motion estimation," *Computer Vision and Image Understanding*, vol. 62, no. 2, pp. 158–169, 1997.

[13] K. Pahlavan, T. Uhlin, and J.O. Eklundh, "Dynamic fixation and active perception," *International Journal of Computer Vision*, vol. 17, no. 2, pp. 113–135, 1996.

[14] J. Liu, Y. Cheng, M. Liu, and Z. Zhao, "Semi-supervised elm with application in sparse calibrated location estimation," *Neurocomputing*, vol. 74, pp. 2566–2572, 2011.

[15] S. Sadanand and J.J. Corso, "Action bank: A high-level representation of activity in video," *IEEE Conference on Computer Vision and Pattern Recgnition*, 2012.

[16] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7.

[17] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," *International Conference on Pattern Recognition*, vol. 3, pp. 32–36, 2004.

[18] M.D. Rodriguez and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[19] J. Liu and M. Shah, "Learning human actions via information maximization," *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[20] A. Gilbert, J. ILLingworth, and R. Bowden, "Fast realistic multi-action recognition using minde dense spatio-temporal features," *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[21] M. Varma and B.R. Babu, "More generality in efficient multiple kernel learning," *International Conference on Machine Learning*, 2011.

[22] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *British Machine Vision Conference*, 2009.

[23] Q. Le, W. Zou, S. Yeung, and A. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[24] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[25] X. Wu, D. Xu, L. Duan, and J. Juo, "Action recognition using context and appearance distribution features," *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.