# Multiple Video Object Extraction Using Multi-Category $\psi$-Learning

Yi Liu and Yuan F. Zheng
Dept. of Electrical and Computer Engineering
The Ohio State University
Columbus, Ohio 43210
Email: {liuyi,zheng}@ece.osu.edu

Xiaotong Shen
School of Statistics
University of Minnesota
Minneapolis, MN 55455
Email: {xshen}@stat.umn.edu

***Abstract*— As a requisite of content-based multimedia technologies, video object (VO) extraction is of great importance. In recent years, approaches have been proposed to handle VO extraction directly as a classification problem. This type of methods calls for state-of-the-art classifiers because the extraction performance is directly related to the accuracy of classification. Promising results have been reported for single object extraction using Support Vector Machines (SVM) and its extensions such as $\psi$-learning. Multiple object extraction, on the other hand, still imposes great difficulty as multi-category classification is an ongoing research topic in machine learning. This paper introduces the newly developed multi-category $\psi$-learning as the multi-class classifier for multiple VO extraction, and demonstrates its effectiveness and advantages by experiments.**

## I. Introduction

Video object (VO) extraction is of great importance for content-based video processing. In recent years, new approaches have been proposed to meet the challenge by handling VO extraction directly as a classification problem [1] [2]. Each VO is considered as a class, and VO extraction is realized by classifying every pixel to one of the available classes. By doing so, the temporal association of the objects between frames is automatically maintained through classifications, and as a result it is more robust to objects with complicated motion characteristics.

Which learning algorithm to use is key to the success of the classification-based approaches. By using powerful classifiers, high classification accuracy can be achieved which directly leads to better performance for VO extraction. For example, $\psi$-learning [3], a new learning machine which has demonstrated both theoretical and experimental advantages over SVM, is employed in [2] and shows great potential. However, most of the results reported are limited to the single object scenario. In other words, only a binary classification problem between the object and the background has been tackled. At the first glance, the extension from single object to multiple object extraction is straightforward since conceptually one only needs to replace the binary classifier with a multi-class classifier. Unfortunately, the implementation of such an extension is far more difficult because multi-category classification is still an ongoing and immature research topic itself in machine learning. Nevertheless, any advances in this area offer new tools that can help researchers tackle the multi-object problem.

The purpose of this paper is twofold. First, it introduces multi-category $\psi$-learning [4], a newly developed algorithm for multi-class learning, to solve the multiple VO extraction problem. Secondly, it reports the performance of the new learning algorithm on several MPGE4 standard video sequences instead of synthetic data which many multi-class learning algorithms are tested on.

The rest of the paper is organized as follows. Section II first gives a brief review of the technologies of multi-class classification, and then introduces multi-class $\psi$-learning. A multiple VO extraction method using multi-class $\psi$-learning is explained in Section III. Section IV provides the experimental results which is followed by conclusions in Section V.

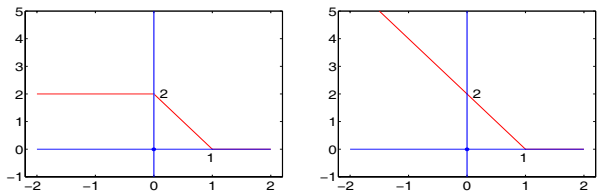
## II. Multi-Category $\psi$-Learning

### A. Review of the Multi-Class Margin-Based Classifiers

Over the last decade, margin-based classification technologies, for which the best known example is SVM [5], have drawn tremendous attention due to their theoretical merits and practical success. Instead of directly estimating the conditional probabilities, the margin-based classifiers focus on the decision boundary, which, however, makes it difficult to generalize their applications from the binary to multi-class scenario.

"Single machine" and "error correcting" are two mainstreams for multi-class margin-based classification. As its name suggests, the "single machine" approach attempts to construct a multi-class classifier by solving just a single optimization problem [6]–[10]. On the contrary, the "error correcting" approach [11] [12] works with a collection of binary classifiers, for which the primary studies are to determine what binary classifiers should be chosen to train and how to combine their classification results to make the final decision. A good overview of multi-class classification can be found in [13] and [14].

As a natural extension of binary large margin classification, the "single machine" approach is intuitively appealing. Recently, it has drawn even more attention when certain formulations are reported to yield classifiers with consistency approaching the optimal Bayes error rate in the large sample limit. Multi-class $\psi$-learning is such a learning algorithm. Unlike the other margin-based classifiers, $\psi$-learning aims directly at minimizing the generalization error (GE), and as



(a) $\psi_b$ function (b) $F_{\text{SVM}}$ function

Fig. 1. $\psi_b$ function for binary $\psi$-learning and $F_{\text{SVM}}$ function for SVM.

a result its binary version has shown significant advantage over SVM in terms of generalization both theoretically and experimentally [3]. The extended multi-class $\psi$-learning retains the desirable properties of its binary counterpart. In addition, a computational tool based on the recent advance in global optimization has been developed to reduce the time of the training of the "single machine" [15].

### B. Notations of Multi-Category $\psi$-Learning

In the frame work of multi-category $\psi$-learning, the class label is coded as $y \in \{1, 2, \ldots, M\}$, and the decision rule is

$$y = \arg\max_{i=1,\ldots,M} f_i(x), \tag{1}$$

where $M$ is the number of classes and $f_i$ is the decision function of class $i$ for $i = 1, \ldots, M$. For the linear classifier $f_i(x) = w_i^T x_i + b_i$.

As a characteristic of multi-class problems, multiple comparisons between classes need to be performed. To simplify the notations, an $(M-1)$-dimensional function vector $g(x, y)$ and a multivariate sign function sign$(u)$ where $u = (u_1, \ldots, u_{M-1})$ are defined as follows

$$g(x, y) = (f_y - f_1, \ldots, f_y - f_{y-1}, f_y - f_{y+1}, \ldots, f_y - f_M),$$

$$\text{sign}(u) = \begin{cases} 1, & \text{if } u_{\min} = \min(u_1, u_2, \ldots, u_{M-1}) \geq 0; \\ 0, & \text{if } u_{\min} < 0. \end{cases} \tag{2}$$

As mentioned before, the most prominent feature of $\psi$-learning is the direct consideration of GE. Defined as the probability of misclassification, GE yielded by an $M$-class classifier is GE $= E[Y \neq \arg\max_{i=1,\ldots,M} f_i(x)]$. It can be shown that with the notations of $g(x, y)$ and sign$(u)$ GE can be rewritten as GE $= \frac{1}{2}E[1 - \text{sign}(g(x, y))]$.

### C. Multi-Category $\psi$-Learning

Seeking the function $f$ to minimize GE is the ultimate goal for any learning algorithm. For example, in the coding system described above[1], the cost function of linear SVM can be rewritten as [4]

$$\begin{aligned} &\text{minimize} : \tfrac{1}{2}\textstyle\sum_{j=1}^{2}\|w_j\|^2 + C\sum_{i=1}^{N} F_{\text{SVM}}\big(f_{y_i}(x_i) - f_{3-y_i}(x_i)\big), \\ &\text{subject to} : \textstyle\sum_{j=1}^{2} f_j(x) = 0 \text{ for } \forall x, \end{aligned} \tag{3}$$

where $N$ is the number of training samples and the sum-to-zero constraint is invoked to eliminate the redundancy in $(f_1, f_2)$. Here the so-called hinge loss $F_{\text{SVM}}(u) = 0$ if $u \geq 1$, and $2(1-u)$ if $u \leq 1$ is a convex upper envelope of $(1 - \text{sign}(u))$. However, there is a difference between this convex envelope and $(1 - \text{sign}(u))$ itself especially for the nonseparable case for which the training error is inevitable. Motivated by this consideration, Shen et. al. proposes to replace $F_{\text{SVM}}$ with a non-convex $\psi$ function [3] [4] as

$$\begin{aligned} &\text{minimize} : \tfrac{1}{2}\textstyle\sum_{j=1}^{2}\|w_j\|^2 + C\sum_{i=1}^{N} \psi_b\Big(f_{y_i}(x_i) - f_{3-y_i}(x_i)\Big), \\ &\text{subject to} : \textstyle\sum_{j=1}^{2} f_j(x) = 0 \text{ for } \forall x. \end{aligned} \tag{4}$$

Here $\psi_b$ can be any function satisfying $R \geq \psi_b(u) \geq 0$ if $u \in [0\ \tau]$ and $\psi_b(u) = 1 - \text{sign}(u)$ otherwise, where $\psi_b(u)$ is non-increasing in $u$ and $\tau \in (0\ 1]$. An example of such a function is shown in Fig. 1(a). Evidentally because of the constant penalty for misclassification $\psi_b$ is much closer to $(1 - \text{sign}(u))$ than $F_{\text{SVM}}$ (Fig. 1(b)), which explains why $\psi$-learning is expected to deliver higher accuracy performance for the nonseparable case.

In analogy to Eq. (4) which is for binary classification, the multi-category $\psi$-learning is formulated as

$$\begin{aligned} \text{minimize} : \quad & \frac{1}{2}\sum_{j=1}^{M}\|w_j\|^2 + C\sum_{i=1}^{N}\psi\Big(g(x_i, y_i)\Big), \\ \text{subject to} : \quad & \sum_{j=1}^{M} f_j(x_i) = \sum_{j=1}^{M}(w_i^T x_i + b_i) = 0. \end{aligned} \tag{5}$$

The $\psi$ function here is a multivariate version of $\psi_b$ with $(M-1)$ arguments which is defined as

$$\begin{cases} R \geq \psi(u) > 0, & \text{if } u \in (0\ \tau_1] \times \ldots \times (0\ \tau_{M-1}]; \\ \psi(u) = 1 - \text{sign}(g(x, y)), & \text{otherwise}, \end{cases} \tag{6}$$

where $0 < \tau_1, \ldots, \tau_{M-1} \leq 1$ and $\psi(u)$ is non-increasing in each $u_j$. The multi-category $\psi$-learning preserves the desired properties of its binary counterpart. More specifically, for any $x$ satisfying $\text{sign}(g(x, y)) = -1$, $\psi$ assigns a constant penalty which is in the same spirit as GE. As a result, it is less sensitive to outliers and offers better learning ability.

## III. Multi-Object Extraction Using Multi-Category $\psi$-Learning

Because of the variety of video contents, the background and the objects are usually nonseparable. For this reason, a VO extraction method employing binary $\psi$-learning as the classifier is proposed in [2], and the results of single VO extraction demonstrate its superior advantage. The basic idea of [2] is to decompose each frame into small blocks, and use $\psi$-learning to classify them as object or background class. The object of interest is then formed by all the foreground blocks. For the remainder of this paper, we will integrate multi-category $\psi$-learning into this method to tackle the task of multi-object extraction.

There are two phases: the training phase and the tracking phase. Suppose we have $M$ VOs of interest. The training phase begins with dividing the first frame, chosen as the training frame, into $(M+1)$ types of blocks (the number

[1]Conventionally, the formulation of SVM is expressed in the coding system where the class label $y \in \{-1, 1\}$.

(a) (b) (c) (d) (e) (f)

Fig. 2. The extraction performance of *Students*. (a) Frame 15. (b) The first object extracted from frame 15. (c) The second object extracted from 15. (d) Frame 90. (e) The first object extracted from 90. (f) The second object extracted from frame 90.

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3. The extraction performance of *Trevor*. (a) Frame 6. (b) The first object extracted from frame 6. (c) The second object extracted from frame 6. (d) The third object extracted from frame 6. (e) Frame 41. (f) The first object extracted from frame 41. (g) The second object extracted from frame 41. (h) The third object extracted from frame 41.

(a) (b) (c) (d) (e)

Fig. 4. The extraction performance of *Sun Flower Garden*. (a) Frame 2. (b) The first object extracted from frame 2. (c) The second object extracted from 2. (d) Frame 114. (e) The first object extracted from 114.

of different VOs plus background) depending on which object or background the pixel at the center of the block belongs to. Discrete Cosine Transform (DCT) is applied to each block and based on the DCT coefficients the local and neighboring features are constructed to represent every block as well as the centering pixel. Then by solving the optimization problem Eq.(5), $(M+1)$ decision functions that separate the $M$ objects as well as the background are obtained.

In the tracking phase, each subsequent frame is also divided into blocks, and for each block the $M+1$ decision functions are evaluated to decide which object the centering pixel belongs to, which consequently determines the class label of the block. Then the tracking mask of every object is formed by the blocks that have been classified to be in the corresponding class. At this point the resolution of object's boundary is as large as the size of the block. Then by applying a *pyramid boundary refining algorithm* [2], the object boundary can be refined and the pixel-wise accuracy can be achieved. Interested readers are referred to [2] for more details of the latter algorithm.

## IV. Experimental Results

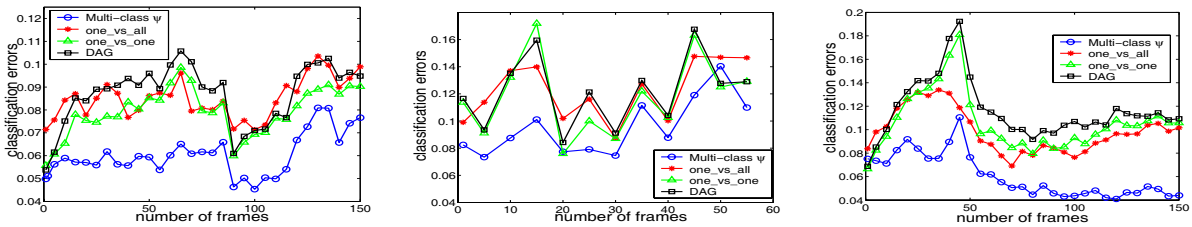
Experiments are conducted on some standard MPEG-4 test video sequences, and the performance comparisons are made between multi-category $\psi$-learning and three popular multi-class algorithms, namely one-vs-all, one-vs-one and directed acyclic graph (DAG) [16].

The first one to test is *Students*. As the major content of this sequence, the two students are chosen as two objects of interest and along with the background this is a three-class classification problem. As one can see from frame 15 and 90 as shown in Fig. 2(a) and 2(d) respectively, *Students* is a typical head-and-shoulder type of sequences. The extracted students are shown in Fig. 2(b), 2(c), 2(e), and 2(f).

Another sequence containing three people is also tested, and the three people are considered as three objects which makes it a four-class classification problem. The original frames and the extracted objects are shown in Fig. 3.

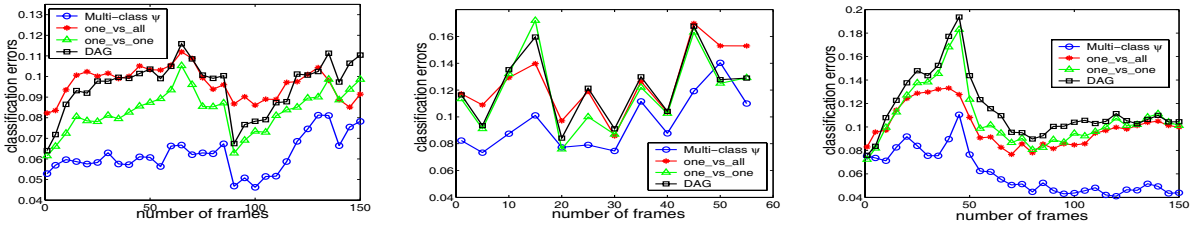
Among the sequences tested in the experiments, *Sun Flower Garden* is the most challenging one. Unlike the previous video-conference kind of sequences, it displays a natural scene that is rich of colors and textures with a non-stationary camera. There are two objects of interest: the house and the tree. For the first few frames, the house is occluded by the tree. One of such frames is shown in Fig. 4(a), and the two extracted objects (house and tree) by using $\psi$-learning are shown in Fig. 4(b) and 4(c). With the camera moving, the tree shifts toward the left hand side of the frame and finally disappears as in Fig. 4(d). From that point on, only the house can be extracted by the proposed method as shown in Fig. 4(e).

The computational complexity of the new approach deserves a discussion. Assume there are $N$ classes and each pixel is represented by a $R$-dimensional feature vector $x$. In the tracking phase, we need to evaluate $N$ functions $f_i = w_i^T x + b_i$ each of which performs $R$ multiplications to decide the class label




(a) Students (b) Trevor (c) Sun Flower Garden

Fig. 5. The comparison of classification errors between multi-category $\psi$-learning, one-vs-all, one-vs-one and DAG. SVM is the underlying binary classifier employed by one-vs-all, one-vs-one and DAG.




(a) Students (b) Trevor (c) Sun Flower Garden

Fig. 6. The comparison of classification errors between multi-category $\psi$-learning, one-vs-all, one-vs-one and DAG. Binary $\psi$-learning is the underlying binary classifier employed by one-vs-all, one-vs-one and DAG.

of a given pixel. As a result, the computational complexity is $O(NR)$, a linear function of the number of objects $N$, which gives the approach low complexity and good scalability.

For their simplicity and effectiveness, one-vs-all, one-vs-one and DAG are three widely-used multi-category algorithms. To see how $\psi$-learning performs against these three methods, the classification errors yielded by all the four methods are displayed every 5 frames in Fig. 5 and Fig. 6, where SVM and binary $\psi$-learning are the underlying binary classifiers respectively. As one can see, for all the three sequences, $\psi$-learning achieves the lowest classification errors almost for every test frame. Although the training is conducted only once by using the first frame, the superior generalization ability of $\psi$-learning enables it to survive nearly the whole sequence.

## V. Conclusions

VO extraction is of great importance for content-based video analysis, and the multi-object scenario imposes great challenges. Following the idea that handles VO extraction directly as a classification problem, this paper attempts to tackle multiple object extraction by solving a multi-class classification problem and using multi-category $\psi$-learning, a newly developed learning algorithm, as the classifier. The experimental comparison between $\psi$-learning and other three popular multi-category classifiers has shown the potential of this new learning machine in the application of VO extraction. Low complexity, which is another advantage of the proposed method, scales well when the number of objects increases.

## References


[1] A. Doulamis, N. Doulamis, K. Ntalianis, and S. Kollias, "An Efficient Fully Unsupervised Video Object Segmentation Scheme Using an Adaptive Neural-Network Classifier Architecture", *IEEE Tran. on Neural Networks*, vol. 14, no. 3, pp. 616-630, May 2003.

[2] Y. Liu and Y.F. Zheng, "Video Object Segmentation and Tracking Using $\psi$-Learning Classification", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 7, pp. 885-899, July 2005.

[3] X. Shen, G. Tseng, X. Zhang, and W.H. Wong, "On $\psi$-learning", *Journal of the American Statistical Association*, vol. 98, no. 463, pp. 724-734, 2003.

[4] Y. Liu, X. Shen, and H. Doss, "Multicategory $\psi$-Learning", *Journal of the American Statistical Association*, to appear.

[5] C. Cortes and V.N. Vapnik, "Support Vector Networks", *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[6] V.N. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.

[7] J. Weston and C. Watkins, "Multi-Class Support Vector Machines", Technical Report CSD-TR-98-04, Royal Holloway, Department of Computer Science, University of London, 1998.

[8] Y. Lin, "Support Vectors Machines and the Bayes Rule in Classification", *Data Mining and Knowledge Discovery*, vol. 6, pp. 259-275, July 2002.

[9] Y. Lee, Y. Lin, and G. Wahba, G., "Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data", *Journal of American Statistical Association*, vol. 99, no. 465, pp. 67-81, Mar. 2004.

[10] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines", *Journal of Machine Learning Research*, vol. 2, pp. 265-292, Dec. 2001.

[11] T.G. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes", *Journal of Artificial Intelligence Research*, vol. 2, pp. 263-286, 1995.

[12] E.L. Allwein, R.E. Schapire, and Y. Singer, "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers", *Journal Machine Learning Research*, vol. 1, pp. 113-141, 2000.

[13] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification" *Journal of Machine Learning Research*, vol. 5, pp. 101-141, 2004.

[14] C. Hsu and C. Lin, "A Comparison of Methods for Multiclass Support Vector Machines", *IEEE Tran. on Neural Networks*, vol. 13, no. 2, 2002.

[15] Y. Liu, X. Shen, and H. Doss, " Multicategory $\psi$-Learning and Support Vector Machine: Computational Tools", *Journal of Computational and Graphical Statistics*, vol. 14, no. 1, pp. 219-236, 2005.

[16] J.C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAG's for Multiclass Classification", *Advances in Neural Information Processing Systems*, vol. 12, pp. 547-553, Cambridge, MA: MIT Press, 2000.