IJICTR International Journal of Information & Communication Technology Research

ITRC

Volume 4- Number 5- December 2012

A Semi-Supervised Method for Multimodal Classification of Consumer Videos

Mahmood Karimian
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
mkarimian@ce.sharif.edu

MostafaTavassolipour
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
tavassolipour@ce.sharif.edu

Shohreh Kasaei
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
skasaei@sharif.edu

Received: January 15, 2013- Accepted: February 18, 3013

Abstract—In large databases, lack of labeled training data leads to major difficulties in classification process. Semi-supervised algorithms are employed to suppress this problem. Video databases are the epitome for such a scenario. Fortunately, graph-based methods have shown to form promising platforms for semi-supervised video classification. Based on multimodal characteristics of video data, different features (SIFT, STIP, and MFCC) have been utilized to build the graph. In this paper, we have proposed a new classification method which fuses the results of manifold regularization over different graphs. Our method acts like a co-training method that tries to find the correct label for unlabeled data during its iterations. But, unlike other co-training methods, it takes into account the unlabeled data in the classification process. After manifold regularization, data fusion is doneby a ranking method which improves the algorithm to become competitive with supervised methods. Our experimental results, run on the CCV database, show the efficiency of the proposed classification method.

Keywords*: Semi-supervised learning; co-training; video classification; multimodal features*

I. INTRODUCTION

Video is inundating different databases, most significantly the web databases. For instance, YouTube is the second largest search engine and the number of videos uploaded on it, is dramatically increasing day by day[1],[2].As such, existence of methods by which users can retrieve their desired videos fast and accurately is inevitable. On the other hand, labeling sufficient videos using human force is irrational; considering the huge volume of available video data. Therefore, semi-supervised learning methods which try to find the proper label of the query video(when there are insufficient labeled data available) must be employed to tackle this problem[3]. A special characteristic of video data is the rmultimodal property; *i.e. ,extracting* features from them could be based on their visual, audio, motional, or textual properties [2], [4]. Co-training which exploits the multimodal property of video is one of the leading semi-supervised learning methods. The underlying assumptions in co-training methods include the independency of modalities for a given class and the ability of each modality to classify data to some extent [5].Co-training method is used in[6] for video concept detection. In that method, the concepts of videos are detected using two modalities. As each modality has a limited ability to detect the concept, they are combined as a supplement to each other, so that their combination leads to improve the labeling accuracy. Till now, the engine of classification in co-training methods has been SVM, neural networks, or naive Bayes (supervised classifiers). In [7] amultimodal regularization method has been proposed to train the classifier. In general, the trend in semi-supervised multimodal algorithms is to apply an iterative classification using only labeled data and assign a label to unlabeled data in each step. The iteration is continued until all data are labeled [8], [9], [10].

Graph-based algorithms are promising semi-supervised learning methods. Two fundamental assumptions in graph-based methods are[11]:

- Samples of high similarity measures tend to have the same labels.
- Estimated label of initially labeled samples should be as close as possible to their real labels.

There is numerous graph-based semi-supervised learning methods available; such as manifold regularization [11], and local / global consistency [12].

II. Related Work

This section gives a comprehensive review on related work. In[13], a semi-supervised feature analyzing framework for multimedia data understanding is proposed and applied to three different applications; namely, image annotation, video concept detection, and 3D motion analysis. Their framework consists of two phases: first they apply a feature selection algorithm, called “$l_{2,1}$-norm” regularized feature selection. It can jointly select the most relevant features from all data points and then apply a manifold learning which analyzes the feature space by exploiting both labeled and unlabeled data. It is a widely used technique for extending many algorithms to semi-supervised scenarios; because of its capability of leveraging the manifold structure of multimedia data. Their proposed method is able to learn a classifier for different applications; by selecting the discriminating features closely related to the semantic concepts.

In[14], a novel approach is proposed for interesting event annotation in videos that is based on semi-supervised learning from available information on the Internet. Concretely, a *fast graph-based semi-supervised multiple instance learning* (FGSSMIL) algorithm, which aims to simultaneously tackle these difficulties for various video domains (*e.g.*, sports, news, and movies) in a generic framework, is proposed. It jointly explores small-scale expert labeled videos and large-scale unlabeled videos to train the models. Two critical issues of FGSSMIL are as follows. How to calculate the weight assignment for a graph construction, where the weight of an edge specifies the similarity between two data points, and how to solve the algorithm efficiently for large-scale dataset through an optimization approach. They use a multiple instance algorithm to solve the first problem and propose a fast iterative convex optimization to tackle the second problem. They perform the extensive experiments in three popular video domains; namely, movies, sports, and news.

It has been proved that graph-based semi-supervised learning approaches can effectively and efficiently solve the problem of labeled data limitation in many real-world applications; such as video concept detection. A multi-graph based *semi-supervised learning* (SSL) method is proposed in [15] that utilizes the different modalities of videos. They show in their proposed method, named *optimized multi-graph-based semi-supervised learning* (OMG-SSL),that various crucial factors in video annotation, including multiple modalities, multiple distance functions, and temporal consistency, all correspond to different relationships among video units, and hence they can be represented by different graphs. These factors have been simultaneously dealt with by learning multiple graphs. They have tested their approach on the TREC video retrieval evaluation (TRECVID) benchmark. The similarity measure in graph-based SSL is very important and can be considered as the key for these methods. In a video concept detection task, the spatial features are not sufficient for classification.

In [16], a novel framework based on spatio-temporal correlation is proposed for video concept detection. The framework is characterized by simultaneously taking into account both the spatial and temporal properties of video data to improve the similarity measurement. In[17], a novel method, called *joint spatio-temporal correlation learning* (JSTCL), is proposed to improve the accuracy of video annotation. The method is characterized by simultaneously considering both the spatial and temporal properties of video data to well represent the pair wise similarity. As multimodal approaches in video classification are becoming more popular, finding a method to fuse the results of learning in different modalities becomes important. The main idea behind multimodal learning is that separately learning from these representations can lead to better gains over merging them into a single dataset. In the same way as ensembles combine results from different classifiers, the outputs given by classifiers in different modalities have to be combined in order to provide a final class label for the query.

In real-world, video databases and specially consumer videos (home videos), can be found in very large sizes. Indexing the concepts of these videos, are more difficult than the TV or another videos that are captured by professional cameras. This is because of the lack of preprocessing stage in their capturing process which makes the classification and indexing of these types of videos to be more challenging. Due to the lack of text descriptions as well as the difficulties in analyzing the content of consumer videos, little work has been conducted to provide video search engines in the consumer domain. In [19], a content-based consumer video search system based on multimodal concept classifications developed. The system supports the query-by-example access mechanism, by exploiting the query-by-concept search paradigm underneath, where online concept classification is conducted over the query video by integrating both visual and audio information. The system adopts an audio-visual group let representation that captures salient audio-visual signatures to describe the video content for efficient concept classification.

In [20], a new method for automatically classifying consumer video clips based on their soundtracks is presented. The main audio feature that is used is the *mel-frequency cepstral coefficient* (MFCC) .We have also used this audio feature to build a graph on audible features of videos.

In this paper, we introduce a method in which the multimodal property of video data is combined with graph-based semi-supervised algorithms. The underlying assumption of our algorithm is that data are scattered over a manifold for different modalities. This is a rationale assumption as discussed in [3]. Furthermore, since in our algorithm the multimodal property is utilized, even if data in some of the modalities are not structured as manifold, our algorithm is robust. The merits of each modality are employed in our method using a ranking fusion method. First, a graph is formed by using each of the existing modalities. Then, a ranking-based decision criterion over manifold regularization output of each graph is utilized to label the most efficient samples in our ranking metric. The algorithm iterates until all samples are labeled. The proposed *semi co-training with graph fusion* (SCGF) method has several benefits including:

1. In each step, not only the labeled data but also the unlabeled ones are utilized in the classifier training phase. As such, the unlabeled data are also involved in choosing the best samples to be labeled; which is why the method is called semi co-training.
2. Different modalities of video data are exploited by their related merit. In this paper, we have used SIFT,STIP, and MFCC features which are visual, motional, and audible features, respectively. All of these features are normalized to be comparable using the bag-of-xmethod [2].
3. Fusion criterion over different graphs guarantees superior results over the single graph scenario; which means that the efficiency of each graph is combined with that of the others.
4. The more proper the graphs are formed, the better the classification results would be. SCGF is generic in its graph; which means that the classification results can be improved if the graphs are improved.

The rest of the paper is organized as follows. In Section 2 the graph-based semi-supervised learning methods are reviewed. Section 3 is devoted to our proposed algorithm. In Section 4the experimental results are discussed. Finally, Section 5 draws the conclusion.

III. PROBLEM FORMULATION

We have a multitude of videos as inputs to the algorithm; each of them is attributed with three feature vectors corresponding to three modalities. If the *i*-th video is denoted byv_i,then

$$v_i = \{x_{i1}, x_{i2}, x_{i3}\} \quad (1)$$

in which $\boldsymbol{x_{ij}}$ is the *j*-thmodality of videov_i.

All videos are separated into two groups of labeled,$V_{labeled}$,and unlabeled,$V_{unlabeled}$. Let$D$denote the set of all videos, then

$$D = \{V_{labeled}\ , V_{unlabeled}\ \} = \{(\boldsymbol{v_i}, \boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{l}} + \{\boldsymbol{v_i}\}_{\boldsymbol{i=l+1}}^{\boldsymbol{l+u}} \quad (2)$$

in which $y_i$is the label of the *i*-thlabeled video. The problem now becomes to find the label of unlabeled videos$\{y_{\boldsymbol{i}}\}_{\boldsymbol{i=l+1}}^{\boldsymbol{l+u}}$. To put it in another way, a function $f$should be obtained which detects the membership of each video to the positive or negative group.

In the graph-based semi-supervised learning methods, each sample is considered as a node of the graph and each edge is weighted by the similarity of its ending nodes. Each graph is denoted by matrix$\boldsymbol{W}$ defined by:

$$W_{ij} = similarity(v_i, v_j). \quad (3)$$

A function f, containing the domain of all nodes, must be found such that fulfills two criteria of:

1. Correctness: the estimated label of $f$for labeled samples should be as close as possible to their real labels.
2. Smoothness: estimated labels of $f$must project the similarity of nodes. It means that nodes with higher similarity are more probable to have the same label.

Manifold regularization meets these two criteria by solving the optimization problem

$$argmin_f\{\frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A\|f\|_K^2 + \gamma_I\|f\|_I^2\} \quad (4)$$

where$V$ is an arbitrary loss function, $K$is a ‘base kernel’, e.g. a linear or RBFkernel. I is a regularization term induced by thelabeled and unlabeled data and γ_A, γ_I are the regularization parameters [3].

In local and global consistency method, the aforementioned two criteria are met but normalization is also applied. The optimization problem of local and global consistency method is

$$argmin_f\{\sum_{i,j=1}^{n}\left\|\frac{f_i}{\sqrt{D_i}} - \frac{fj}{\sqrt{D_j}}\right\|^2 W_{ij} + \mu\sum_{i=1}^{n}\|f_i - y_i\|^2\} \quad (5))$$

in which $D_i = \sum_{j=1}^{n}\boldsymbol{W}_{ij}$, μ is the regularization parameter, and n in is the total number of samples consisting of both labeled (l) and unlabeled (u) ones.

IV. PROPOSED SEMI CO-TRAINING METHOD

The block diagram shown in **Error! Reference source not found.**, shows the processing steps in SCGF method. We will develop SCGF in two modes; namely, deterministic mode (SCGF-D) and probabilistic mode (SCGF-P).

A. Deterministic Mode

As mentioned above, we have a multitude of videos, each of them presented by three feature vectors corresponding to three modalities. First, a graph is formed for each of the modalities. Each video sample forms a node, and each edge represents the magnitude of the similarity between its two ending nodes. We have modeled the similarity using

$$\boldsymbol{W}_{ij}^{m} = exp\left(-\frac{\left\|\boldsymbol{x}_{im} - \boldsymbol{x}_{jm}\right\|^{2}}{\sigma^{2}}\right), m = 1,2,3 \qquad (6)$$

Where σ is a scalar parameter.

At first, each graph is pruned using the *k-nearest-neighbor* (KNN) method. In KNN, for each node only the k edges with highest similarity measures are kept. After pruning, manifold regularization is applied to each graph. The output of manifold regularization for the *m*-th graph form svector $\boldsymbol{f}^{m}$. The label of each node as positivity or negativity is decided by its corresponding element in $\boldsymbol{f}^{m}$; $\boldsymbol{f}^{m}$ is larger for the samples which are more likely to be positive.

Efficiency of different modalities of videos varies for distinct concepts (or classes). For instance,*mel-frequency cepstral coefficients*(MFCC) which is based on audible properties of videos is highly efficient in detecting a dog in a video but inefficient to detect butterflies [2]. Therefore, there is a need to fuse the results of the graphs over different modalities.

In this paper, we propose to use the co-training-like method for fusion.

As shown in**Error! Reference source not found.**, at the first step, three graphs are made over three modalities using the *manifold regularization* (MR) method.Then, a ranking method votes on the output of manifold regularization step to label the best samples. The metric of the ranking is

$$\boldsymbol{\beta}(v) = max(\boldsymbol{f}^{m}(v)), \qquad m = 1,2,3 \qquad (7)$$

thenn_{pos} samples with large $\boldsymbol{\beta}$formthe candidates to be labeled as positive. Also, n_{neg} samples with the lowest β are added to negative labels. Addition to negative samples requires no further process; because of the large number of negative data. But, adding samples to the positive set should be done with more precaution.

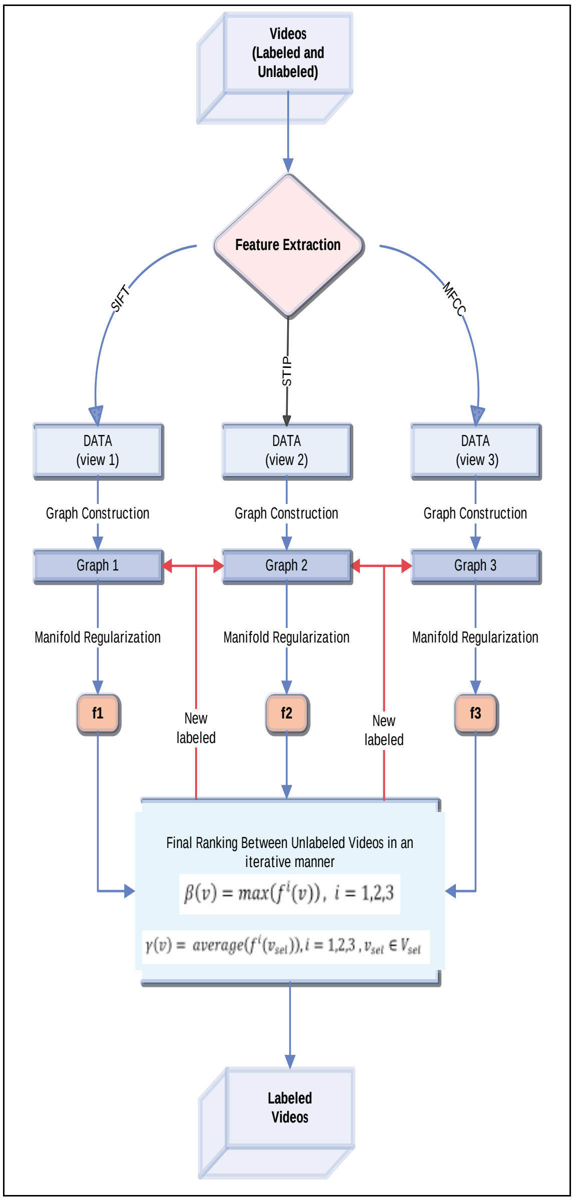

Fig.1. Block diagram of proposed SCGF method

Therefore, another step is required to label a sample

$$\gamma(v) = average(f^{m}(v_{sel})), m = 1,2,3 \,, v_{sel} \in V_{sel} \qquad (8)$$

as positive. Suppose V_{sel} is the set of n_{pos} selected candidates of the current step. By defining a sample with maximum γ is labeled as positive.

Eq. (7) guarantees that we are selecting the best modality for each of the videos. The rationale behind this formula is simple and yet intuitive. If one of the modalities strongly tends to classify a video as positive, then the corresponding video is a candidate to be labeled as positive. Over the entire positive candidates, a video is labeled positive when all modalities meet Eq. (8). Furthermore, if a video strongly tends to be negative by all modalities it is labeled as negative.

Based on the above algorithm, in each step, one video is labeled as positive and n_{neg} videos are labeled as negative. Positive and negative labeled videos are juxtaposed in the manner shown in Fig.2 to form a ranking vector.

O = Ranked outputvector.
Size (O) = Number of unlabeled data

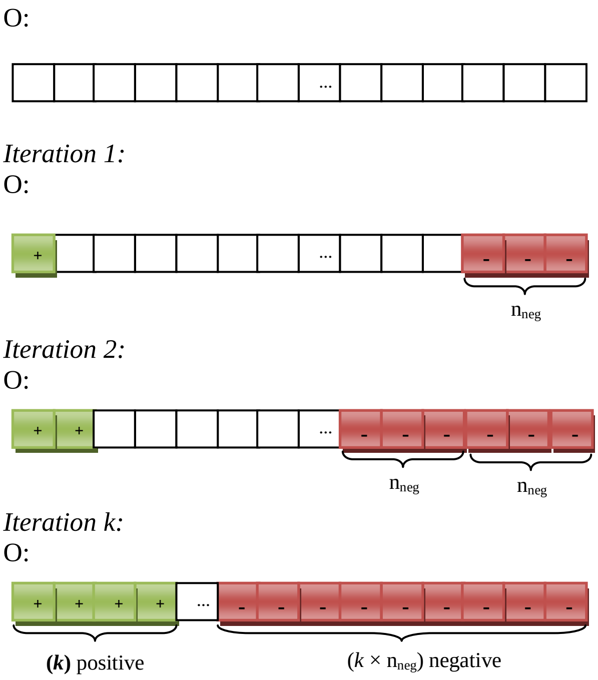

Fig.2. Forming ranked output vector.

In a ranking vector, shown in Fig.2.,the righter the sample in the vector, the more negative it tends to be and vice versa.

Fig. 3shows the proposed SCGF algorithm, step by step.

B. Probabilistic Mode

The drawback of introduced SCGF-D algorithm is that at each iteration, it only uses former iteration labeled videos. It may seem more rational to use all labeled videos of previous iterations. This strategy drastically reduces the performance of the algorithm due to error propagation problem. This problem shows itself more obviously in manifold regularization algorithm, that's why we don't apply all labeled data of previous iterations. Therefore, in this subsection, we modify our proposed method to use more labeled video in each of the iteration while avoiding error propagation. The new SCGF is shown in Fig. 4.

The above mentioned method increases the accuracy because:

1) More labeled videos are used in each iteration; making the manifold regularization output as a better representation of video classes.

2) As in the first iterations the initially labeled data play a more pivotal role it's less probable to have false positive or negative sample in them. Thus, our algorithm tends to use sampled video of first iterations more.

```
1.  Separate the video database into labeled and
    unlabeled videos $D = \{V_{labeled}, V_{unlabeled}\}$ and
    denote each video by three modalities
        $\mathrm{v_i} = \{\mathrm{x_{i1}}, \mathrm{x_{i2}}, \mathrm{x_{i3}}\}$.
Set $V_{curr} = V_{labeled}$ .
Set $Out = \emptyset$.
2.  A KNN graph is constructed for each of the
    modalities by:
        $\boldsymbol{W_{ij}^m} = \exp\left(-\frac{\left\|\boldsymbol{x_{im}} - \boldsymbol{x_{jm}}\right\|^2}{\sigma^2}\right), m = 1,2,3.$
3.  While $V_{unlabeled} \neq \emptyset$
    3.1. Run the manifold regularization for each of
         the graphs and denote the output by $f^i$.
         [Note: $V_{curr}$ is considered as a labeled video
         set and all other videos are considered as
         unlabeled.]
    3.2. Form the ranking metric using Eq. (7). If
         $n_{neg}$ videos with the lowest $\beta$ are shown by
         the set $V_{neg}$ and $n_{pos}$ videos with the highest
         $\beta$ are shown by set $V_{sel}$, then
             $Out = Out + V_{neg}$ .
    3.3. Find a video $v_{pos}$ which has the maximum $\gamma$
         in $V_{sel}$, then
             $Out = Out + \{v_{pos}\}$.
    3.4. Set $V_{curr} = V_{labeled} + V_{neg} + \{v_{pos}\}$ and
           $V_{unlabeled} = V_{unlabeled} - V_{neg} - \{v_{pos}\}$.
    3.5. Form a ranking vector as shown in Fig.2.
```

Fig. 3. SCGF-D algorithm**.**

```
1. and 2. Same as SCGF-D.
    3. While $V_{unlabeled} \neq \emptyset$
      3.1.  Run the manifold regularization for
            each graph and denote the output by$f^i$.
            [Note: $V_{curr}$ is considered as a labeled video
            set and all other videos are considered as
            unlabeled in the current iteration.]
      3.2.  Form the ranking metric in Eq. (7). If
            $n_{neg}$ videos with the lowest $\beta$ are shown by set
            $V_{neg}$ and $n_{pos}$ videos with the highest $\beta$ are
            shown by set $V_{sel}$, then
                $Out = Out + V_{neg}$
      3.3.  Find a video $v_{pos}$ which has the
            maximum $\gamma$ in $V_{sel}$, then
                $Out = Out + \{v_{pos}\}$.
      3.4.  Randomly select labeled data from
            the first $r$ iterations, i.e., set
                $V_{curr} = V_{labeled} + Out(rand(r))$
                $r < current\ iteration\ number$
              $V_{unlabeled} = V_{unlabeled} - V_{neg} - \{v_{pos}\}$
            [Note: $rand(r)$ returns a random integer
            between 1 and $r$.]
      3.5.  Form a ranking vector as shown in
            Fig.2.
```

Fig. 4. SCGF-P algorithm.

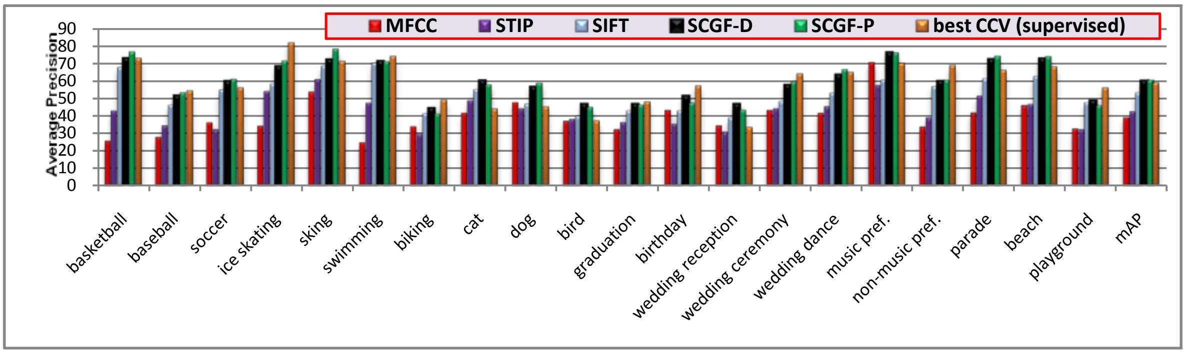

Fig.5. Comparison of our proposed SCGF method with different existing classification algorithms

EXPERIMENTAL RESULTS

As a benchmark for our algorithm, we run our method on CCV, a video database collected at Columbia university [2]. It contains 9317 videos in 20 classes. One-against-all method is utilized for classifying each of 20 classes. For each class, first a set of random samples are added to the labeled set. Our labeled set members are only 10% of all existing videos while the supervised algorithm in [2] uses 50% of videos in its labeled set. Then, using our method, the remaining video samples in the database are labeled. The precision-recall and average precision are utilized as the metric of comparison. Precision is the ratio of true positive estimated labels to all estimated positive ones. Recall is the ratio of true positive estimated labels to all the positive samples of database. *Precision-recall* (PR) is the plot of precision against recall. Average precision is the area below the PR curve encounters for both the precision and recall measures. Methods with higher average precision have better performance in classification.

C. Analysis and Discussion

The parameters in our algorithm are shown in Table.1. These values are obtained for best performance. For example, the experiments run on three classes of CCV dataset (Basketball, Soccer, and Wedding dance) for different values of '*k*' in the range [0 100], shows that the best value of '*k*' is approximately 10. So we have used it in all conducted experiments. **Error! Reference source not found.**, shows the plot of average precision versus different values of '*k*'.

TABLE.1. SCGF PARAMETERS.

n_{neg}	n_{pos}	σ	k (in KNN)
10	1	1.05	10

It is more intuitive to test the effect of '*k*' on k-nearest neighbor in a logarithm plot of the average precision. As can be seen in

Fig.7, value of 10 is an appropriate choice for *k*. improves. But, we noticed that for some classes like 'swimming', this improvement fails. The reason behind this observation is that the data of this class is scattered in the feature space, so that some positive labeled samples are very close to negative labeled samples. Thus, when at the beginning of the process, some labeled data are selected randomly and then their labels propagated according to our semi-supervised approach,the more number of labeled data selected, the more error propagation happened. The results shown in

Fig.8 proves this assumption. All experiments in this paper are run on 10% of initial labeled data.

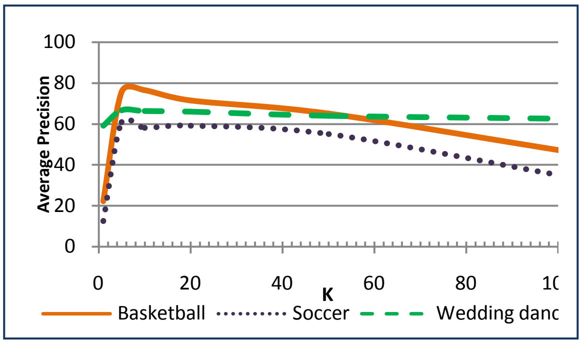

Fig.6.Effect of variable '*k*' on average precision.

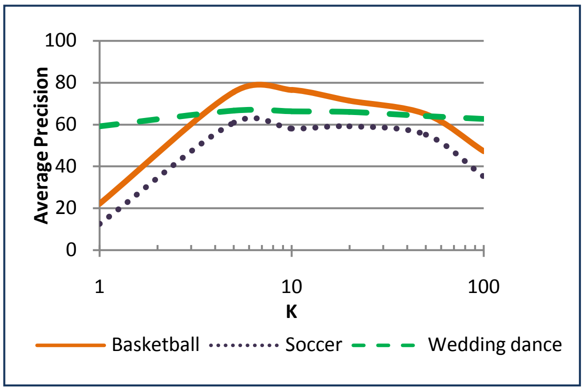

Fig.7. Logarithm plot of '*k*' against average precision.

Average precision of four methods is compared in Fig.5. The first algorithm is a single-modality manifold regularization, the second one is a SCGF-D, and the third one is SCGF-P. Also, the results of the best supervised [2]method are plugged in the figure for convenience of comparison between our semi-supervised methods and the supervised one. *Mean average precision* (MAP) of all these methods are shown in the last column of Fig.5. Although, our proposed methods are semi-supervised, their performances are quite like (or even better than) supervised ones that use many labeled data. As seen, the results of our fusion methods are very promising as predicted.

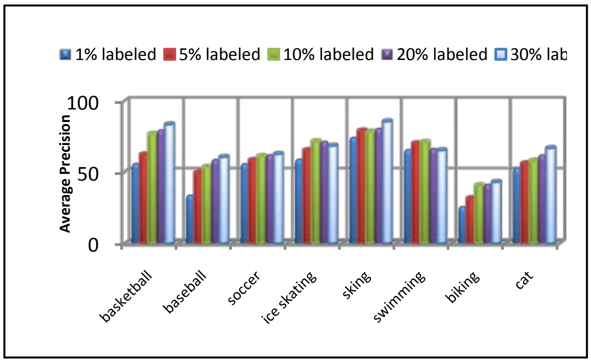

Fig.8. Effect of initial number of labeled data on average precision

In Fig.9, Precision-Recall of our method is compared with that of other existing pioneer methods. As it is seen in Fig.9, our multimodal method surpasses all single modality methods. For instance , consider MFCC feature. It is a very weak discriminating factor in all classes except for the ones for which videos contain music or special audios. Our fusion method, automatically weighs MFCC high in music videos as a discriminating factor. Also, SCGF-P which mostly uses the labeled samples of initial iterations yields the best results in the most classes.

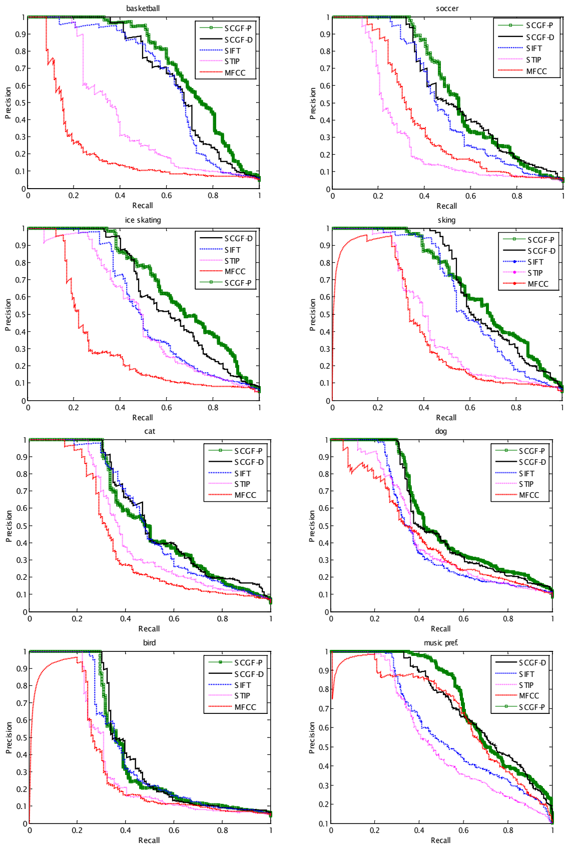

Fig.9.Precision-recall plots for some CCV classes.

Error! Reference source not found., shows the five top true-positive and the first top false-positive results for three classes of CCV database.

V. CONCLUSION

In this paper, a semi-supervised classification method was proposed for video databases which use iterative manifold regularization for classification. We have used a semi-supervised engine (manifold regularization) in our semi-supervised classification method. Co-training was also utilized to exploit multimodal characteristic of videos. To enhance the results of co-training, it was adapted in a way that unlabeled data also play their role for assigning labels to videos. Our algorithm called SCGF was designed to work in two modes, namely deterministic mode (SCGF-D) and probabilistic mode (SCGF-P). It was shown that SCGF algorithm outperforms the supervised algorithms. Our proposed ranking fusion over the graphs and the graph structure are still active research areas. Because of the multimodal nature of the method, many works can be done on the feature extraction and feature selection parts of the method to improve the results.

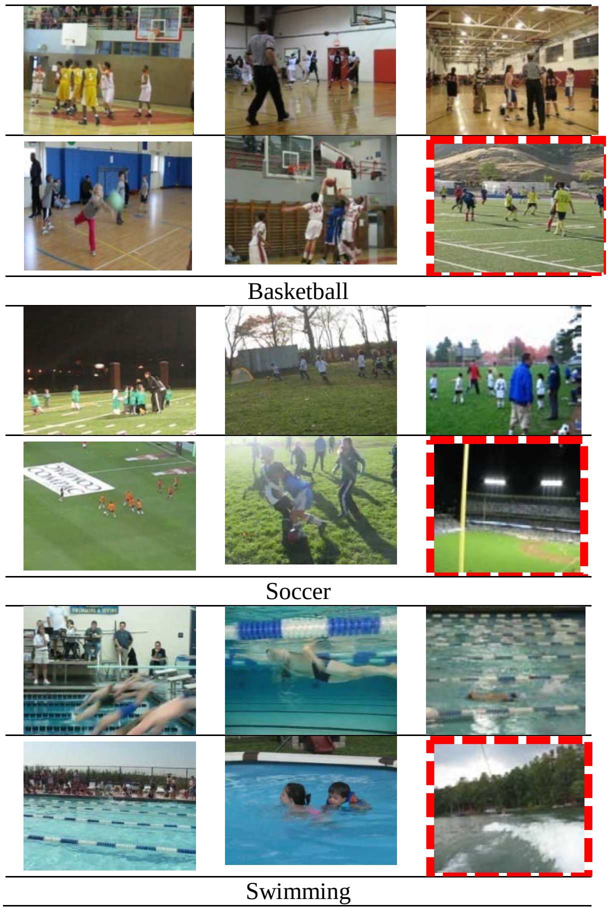

Fig.10.Five top true-positive and first top false-positive results for some classes of CCV database[Dashed red line (— — —) shows the top false-positive result.]

REFERENCES

[1] E. Ackerman and E. Guizzo, “5 technologies that will shape the web”, *Spectrum, IEEE*, vol.48, no.6, pp.40-45, June 2011.

[2] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, “Consumer video understanding: A benchmark database and an evaluation of human and machine performance”, *in Proceedings of ACM International Conference on Multimedia Retrieval (ICMR)*, 2011.

[3] X. Zhu, “Semi supervised learning literature survey”, *Computer Sciences Technical Report*, 2007.

[4] Jiang, Wei, Cotton Courtenay, Loui, Alexander C., "Automatic consumer video summarization by audio and visual analysis", *Multimedia and Expo (ICME), IEEE International Conference on*, pp.1-6, 2011.

[5] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training", *Proc. 11th Annul. Conf. Computational Learning Theory*, pp.92 - 100, 1998.

[6] Y. Rong, M. Naphade, "Co-training non-robust classifiers for video semantic concept detection", *Image Processing, IEEE International Conference on ICIP*, 2005.

[7] V. Sindhwani, P. Niyogi, and M. Belkin, “A co-regularization approach to semisupervised learning with multiple modalities”, *In ICML*, 2005.

[8] Y-h. Han, J. Shao, F. Wu, and B-G. W, “Multiple hypergraph ranking for video concept detection”, *Journal of Zhejiang University - Science C.* pp.525-537, 2010.

[9] W. Meng, X-S Hue, X. Yuan, Y. Song, and L-R Dai, "Multi-Graph SemiSupervised Learning for Video Semantic Feature Extraction", *Multimedia and Expo, IEEE International Conference on*, pp.1978-1981, 2007.

[10] M. Culp and G. Michailidis, “A co-training algorithm for multimodal data with applications in data fusion”, *Journal of chemometrics*, pp.294–303, 2009.

[11] X. Zhu, Z. Ghahramani, and J. Lafferty, “Semisupervised learning using gaussian fields and harmonic functions”, in *Proc. 20th Int. Conf. Machine Learning (ICML’03)*, 2003.

[12] D. Zhou and O. Bousquet, “Learning with local and global consistency”, *In Proceeding of IEEE international conference on neural information processing systems*, pp.321–328, 2003.

[13] Zhigang Ma; FeipingNie; Yi Yang; Uijlings, J.R.R.; Sebe, N.; Hauptmann, A.G., "Discriminating Joint Feature Analysis for Multimedia Data Understanding", Multimedia, IEEE Transactions on, vol.14, no.6, pp.1662-1672, Dec. 2012.

[14] Tianzhu Zhang; ChangshengXu; Guangyu Zhu; Si Liu; Hanqing Lu, "A Generic Framework for Video Annotation via SemiSupervised Learning", Multimedia, IEEE Transactions on, vol.14, no.4, pp.1206-1219, Aug. 2012.

[15] Meng Wang; Xian-Sheng Hua; Richang Hong; Jinhui Tang; Guo-Jun Qi; Yan Song, "Unified Video Annotation via Multigraph Learning", Circuits and Systems for Video Technology, IEEE Transactions on, vol.19, no.5, pp.733-746, May 2009.

[16] Hongzhong Tang; Huixian Huang; Songhao Zhu, "Video concept detection based on spatio-temporal correlation", Computer Application and System Modeling (ICCASM), 2010 International Conference on, vol.8, no., pp.V8-638-V8-642, 22-24 Oct. 2010.

[17] Songhao Zhu; Yuncai Liu, "A novel semantic model for video concept detection", Image Processing (ICIP), 2009 16th IEEE International Conference on, vol., no., pp.1837-1840, 7-10 Nov. 2009.

[18] Cordeiro Junior, Z.; Pappa, G.L.; , "A PSO algorithm for improving multi-view classification," Evolutionary Computation (CEC), 2011 IEEE Congress on , vol., no., pp.925-932, 5-8 June 2011

[19] Wei Jiang; Loui, A.C.; Lei, P.; , "A consumer video search system by audio-visual concept classification", Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, vol., no., pp.39-44, 16-21 June 2012.

[20] Lee, K.; Ellis, D.P.W., "Audio-Based Semantic Concept Classification for Consumer Video", Audio, Speech, and Language Processing, IEEE Transactions on, vol.18, no.6, pp.1406-1416, Aug. 2010.

Mahmood Karimian received his B.Sc.degree in Computer Engineering from the Department of Engineering, Ferdowsi University of Mashhad, Iran, in 2009, and his M.Sc. degree from the department of Computer Engineering, Sharif University of Technology, Iran, in 2011. He joined Image Processing Laboratory (IPL) in 2009. His research interests include machine learning, pattern recognition, and computer vision.

Mostafa Tavassolipour received his B.Sc. degree with honor from the department of Computer and Electrical Engineering, Shahed University, Iran, in 2009, and his M.Sc. degree from the department of Computer Engineering, Sharif University of Technology, Iran, in 2011. He is a member of Image Processing Laboratory (IPL) since 2009. His research interests include image processing, content-based video analysis, machine learning, and statistical pattern recognition.

Shohreh Kasaei (M'05-SM'07) received her B.Sc. degree from the Department of Electronics, Faculty of Electrical and Computer Engineering, Isfahan University of Technology, Iran, in 1986, her M.Sc. degree from the Graduate School of Engineering, Department of Electrical and Electronic Engineering, University of the Ryukyus, Japan, in 1994, and the Ph.D. degree from Signal Processing Research Centre, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, Australia, in 1998. She joined Sharif University of Technology since 1999, where she is currently a full professor and the director of Image Processing Laboratory (IPL). Her research interests include multi-resolution texture analysis, 3D computer vision, 3D object tracking, scalable video coding, image retrieval, video indexing, face recognition, hyperspectral change detection, video restoration, and fingerprint authentication.