VISUAL EVENT RECOGNITION USING ADAPTIVE SUPPORT VECTOR MACHINE

R. Kavitha[1], D. Chitra[2]

ABSTRACT

Video has more information than the isolated images. Processing, analyzing and understanding of contents present in videos are becoming very important. Consumer videos are generally captured by amateurs using hand held cameras of events and it contains unstable camera, poor background and large difference in same type of events, making their visual volumes highly variable and less discriminant. So visual event recognition is extremely challenging task in computer vision. A visual event recognition framework for consumer videos are framed by leveraging a large amount of loosely labeled web videos. The videos are divided into training and testing sets manually. A simple method called Aligned Space Time Pyramid Matching method proposed to effectively measure the distances between two video clips from different domains. Each video is divided into space time volumes over multiple levels. A new transfer learning method is referred to as Adaptive Multiple Kernel Learning fuse the information from multiple pyramid levels, features and cope with the considerable variation in feature distributions between videos from two domains web video domain and consumer video domain. With the help of MATLAB simulink videos are divided and compared with web domain videos. The inputs are taken from Kodak data set and the results are given in the form of MATLAB simulation.

Keywords - adaptive svm, classifiers, kernel learning, pyramid matching, Support vector machine.

I. INTRODUCTION

In the past few years, computer vision researchers have witnessed a surge of interest in human action analysis through videos. With the rapid adoption of digital cameras and mobile cameras, event recognition in all videos produced by consumers has become an important research topic due to its usefulness in automatic video retrieval and indexing. Event recognition from visual cues is a challenging task because of complex motion, poor backgrounds, as well as geometric and photometric differences of objects. Previous work on event recognition can be roughly classified as either activity recognition or abnormal event recognition. First, a large corpus of training data is collected, in the concept labels are generally obtained through expensive annotation. Next, classifiers also called models or concept detectors are learned from the collected data. The classifiers are

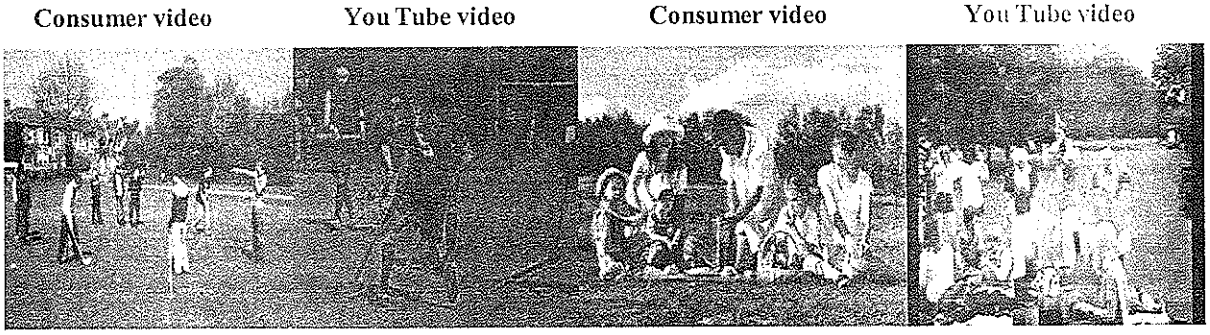
[1] Assistant Professor, Department of CSE, P.A. College of Engineering and Technology
E-mail : [1]rkavitha.pacet@gmail.com

[2]Professor and Head, Department of CSE, P.A. College of Engineering and Technology
E-mail : [2]chitrapacet@gmail.com

used to detect the presence of the concepts in any test data. Sufficient and strong labeled training samples are taken , all other event recognition methods have provide promising results. It is already-known that the learned classifiers from a limited number of labeled training samples are usually not robust and do not generalize well. This project proposes a new event recognition frame work for consumer videos by leveraging a large number of loosely labeled You Tube videos. A large amount of loosely labeled You Tube can be readily obtained by using keywords based search. You Tube videos are down sampled and compressed by web server, so the quality of You Tube videos is generally lower than consumer videos. YouTube videos may have been selected and edited to attractive attention, but consumer videos are in their natural captured state. In Figure 1.1 shows four frames from two events picnic and sports as examples to illustrate the considerable appearance differences between consumer videos and YouTube videos. The feature distributions of samples from the two domains web video domain and consumer video domain may change considerably in terms of the statistical properties, they are mean, intraclass variation and interclass variation.

An event recognition framework extend the recent work on pyramid matching and present a new matching method called Aligned Space-Time Pyramid Matching to effectively measure the distance between two video clips that may be from different domains. Divide each video into space-time volumes over multiple levels and calculate the pair wise distances between any two volumes and further integrate the information from different volumes with integer flow Earth Movers Distance to explicitly align the volumes. The Earth Mover's Distance (EMD) is a method to evaluate dissimilarity between two multi-dimensional distributions in some feature space [3,11]. The EMD lifts this distance from individual features to full distributions.

A technique that uses local space-time features to classify six human actions like walk, jog, run, wave, clap, and box in challenging real-world video sequences. This technique achieves comparable performance in the presence of motion of camera, scale variation, time



(a) "sport sample" (b) "picnic sample"

Figure : 1 Four samples frames from consumer videos and youtube videos. The examples from two events illustrate the appearace differences between consumer videos and you tube videos.

variation and viewpoint changes. Hinder the use of 2D local descriptors for object detection in static images also impact spatiotemporal local descriptors. Cross- domain learning method, this method called as Adaptive Multiple Kernel Learning (A-MKL) [6,2,18], in order to cope with the considerable variation in feature distributions between videos from the web domain and consumer domain. Each pyramid level and each type of local features, train a set of Adaptive SVM classifiers. Based on a combined training set from two domains by using multiple base kernels of different kernel types and parameters, are further fused with equal weights to obtain an average classifier. A new objective function to learn an adapted classifier based on multiple base kernels and the prelearned average classifiers by minimizing both the structural risk functional and mismatch of data distributions from two domains.

II. Related Works

Event recognition methods can be roughly categorized into model-based methods and appearance- based techniques. Model-based approaches deal with various models including HMM, coupled HMM, and Dynamic Bayesian Network [20] to model the temporal evolution. Appearance-based approaches employed space-time features extracted from salient regions with significant local variations in both spatial and temporal dimensions [11, 19, 22].

Statistical learning methods including Support Vector Machine (SVM) [22], probabilistic Latent Semantic Analysis (pLSA) [19], and Boosting [8] were applied to the space-time features to obtain the final classification. Promising results [1, 19, 22] have been reported on video data sets under controlled settings, such as Weizman [1] and KTH [22] data sets. Classifier adaptation can be seen as an effort to solve the fundamental problem of mismatched distributions between the training and testing data. This problem occurs in concept detection in a video corpus such as TRECVID [13], contains data from different sources programs. In existing approaches [12, 14, 17], concept classifiers are built from and applied to data collected from all the programs without considering their difference on distribution. In this paper, a different scenario where classifiers trained from one or several programs are adapted to a different program are considered.

The proposed classifier adaptation method is related to the work on drifting concept detection in the data mining community and transfer learning and incremental learning in the machine learning community. Incremental learning methods, such as incremental SVMs [15, 1, 9], continuously update a model with new examples without re-training over all the examples. The training and test distribution is identical, A-SVMs can be treated as a generic incremental method can handle classifiers of any type. It is also more efficient than existing methods [15, 1, 9] whose training involves at least part of the previous examples support vectors.

III. Pyramid Matching

Spatial pyramid matching [8,12] and its space time extension [10] used fixed block-to-block matching and fixed volume-to-volume matching. In contrast, this aligned pyramid matching extends the methods of Spatially Aligned Pyramid

Matching (SAPM) [20,18] and Temporally Aligned Pyramid Matching (TAPM) [16] from either spatial domain or temporal domain to joint space time domain, the volumes across different space and time locations may be matched.

Similar to [10], divide each video clip into 8^l non overlapped space-time volumes over multiple levels, l=0,….,L-1 where the volume size is set as $1/2^l$ of the original video in width, height and temporal dimension. Following [10], extract the local space-time (ST) features including Histograms of Oriented Gradient (HoG) and Histograms of Optical Flow (HoF), are further concatenated together to form lengthy feature vectors. Sample each video clip to extract image frames and then extract static local SIFT features from them [17]. This method consists of two matching stages. In the first matching stage, calculate the pairwise distance Drc between each two space-time volumes Vi(r) and Vj(c), where r,c = 1,…..,R with R being the total number of volumes in a training and testing data. All features are vector-quantized into visual words and then each space-time volume is represented as a token-frequency feature. As suggested in [10], to measure the distance Drc using equation (3.1) Note that each space-time volume consists of a set of image blocks.

Token-frequency (tf) features from each image blocks are extracted by vector-quantizing the

$$Drc = \frac{\sum_{u=1}^{H} \sum_{v=1}^{I} \hat{f}duv}{\sum_{u=1}^{H} \sum_{v=1}^{I} \hat{f}uv} \qquad \text{(Eq. 3.1)}$$

where H, I are the numbers of image blocks in Vi(r), Vj(c) respectively, duv is the distance between two image block Euclidean distance is used in this work, and $\hat{f}uv$ is the optimal flow that can be obtained by solving the linear programming problem as follows:

$$\hat{F}rc = \underset{Frc}{\operatorname{argmin}} \sum_{H=}^{R} \sum_{C=} FrcDrc \qquad \text{(Eq. 3.2)}$$

$$\text{s.t.} \sum_{C=1}^{R} Frc = 1, \forall r \quad \sum_{l=a}^{R} Frc = 4, \forall c \qquad \text{(Eq. 3.3)}$$

In the second stage, further integrate the information from different volumes with Integer-flow EMD to explicitly align the volumes. Try to solve a flow matrix $\hat{F}_{rc}$ containing binary elements that represent unique matches between volumes Vi(r) and Vj(c). As suggested in [21], such binary solution can be conveniently computed by using the standard Simplex method for linear programming.

IV. Adaptive Multiple Kerenel Learning

The proposed framework consists of three contributions:

An event recognition framework for testing videos with only a limited number of labeled consumer videos by leveraging a large amount of loosely labeled web videos.

Pyramid matching extended by presenting a new matching method called aligned space-time pyramid matching (ASTPM) to effectively measure the distances between two video clips.

A cross-domain learning method Adaptive Multiple Kernel Learning (A-MKL), used to cope with the considerable variation in feature distributions between videos from the web video domain and consumer video domain by minimizing both the structural risk functional and mismatch of data distributions from two domains.

Web video domain taken as the auxiliary domain $T T D^A$ source domain and the consumer video domain as

corresponding SIFT features into visual words. Based the target domain D^T. $D^T = D_l \cup D_u$, Where D^T_l and on the SIFT features, as suggested in [16], the pairwise distance Drc between two volumes Vi(r) and Vj(c) is calculated by using Earth Mover's Distance (EMD), D^T_u represent the labeled and unlabeled data in the target domain. Transfer learning domain adaptation or cross domain learning methods have been proposed for many applications. To take advantage of all labeled patterns from both auxiliary and target domains, in previous work proposed a Feature Replication (FR) [7, 18] by using augmented features for SVM training. In Adaptive SVM (ASVM) the target classifier $f^T(x)$ is adapted from an existing classifier $f^A(x)$ as auxiliary classifier trained based on the samples from the auxiliary domain. Figure 4.1 illustrate event recognition for consumer videos by leveraging a large number of loosely labeled You Tube videos.

Divide each video into 8^l non-overlapped space- time volumes over multiple levels, l=0,…,L-1. where the volume size is set as $1/2^l$ of the original video in width, height and temporal dimension. The partition for two videos Vi and Vj at level-1. The local space-time (ST) [8, 21, 20] features including Histograms of Oriented Gradient (HoG) and Histograms of Optical Flow (HoF), are extracted and further concatenated together to form lengthy feature vectors. Sample each video clip to extract image frames and then extract static local SIFT features from them.

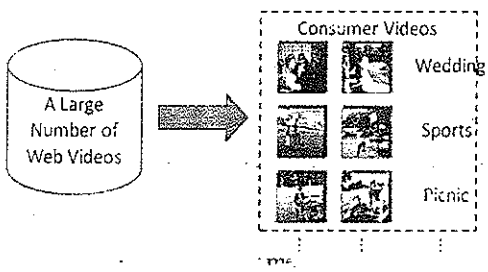


Figure : 1. Event recognition in consumer videos by leveraging large number of web videos.

The two matching stages are:

In the first matching stage, calculate the pairwise distance Drc between each two space-time volumes Vi(r) and Vj(c), where r,c=1,…..,R with R being the total number of volumes in a video.

In the second stage, further integrate the information from different volumes with Integer flow Earth Mover's Distance to explicitly align the volumes. Solve a flow matrix $\hat{F}_{rc}$ containing binary elements that represent unique matches between volumes Vi(r) and Vj(c) :

$$\hat{F}_{rc} = \underset{F_{rc}}{\operatorname{argmin}} \sum_{r=1}^{R} \sum_{c=1}^{R} F_{rc} D_{rc} \quad \text{(Eq. 4.1)}$$

$$\sum_{r=1}^{R} F_{rc} = f_n s \frac{2}{1}, \forall r \text{ s.t.} \sum_{c=1}^{R} F_{rc} = 1, \forall r \sum_{r=1}^{R} F_{rc} = 1, \forall c \quad \text{(Eq. 4.2)}$$

Then, the distance between two videos Vi and Vj can be directly calculated by

$$D_i(V_i, V_j) = \frac{\sum_{r=1}^{R} \sum_{c=1}^{R} F_{rc} D_{rc}}{\sum_{r=1}^{R} \sum_{c=1}^{R} F_{rc}} \quad \text{(Eq. 4.3)}$$

The matching results are obtained by using ASTPM method. Each pair of matched volumes from two videos is highlighted in the same color. Cross- domain learning methods have been proposed for many applications [4, 5, 15]. To take advantage of all labeled patterns from both auxiliary and target domains, Daum´e III [4] proposed Feature Replication (FR) by using augmented features for SVM training. In Adaptive SVM (A-SVM)], the target classifier fT (x) is adapted from an existing classifier fA(x) referred to as auxiliary classifier trained based on the samples from the auxiliary domain.

The target decision function is defined as While A-SVM can also employ multiple auxiliary classifiers, these auxiliary classifiers are equally fused to obtain fA(x). Moreover, the target classifier fT (x) is learned based on only one kernel.

Recently, Duan [5] proposed Domain Transfer SVM (DTSVM) to simultaneously reduce the mismatch in the distributions between two domains and learn a target decision function.

The prelearned classifiers are used as prior for learning a robust adapted target classifier. Train a set of independent classifiers for each pyramid level and each type of local features using the training data from two domains. The prelearned classifiers are used as prior for learning a robust adapted target classifier. Further equally fuse these classifiers to obtain average classifiers $f_l^{STT}(x)$ and $f_l^{ST}(x)$ l=0,....,L-1. These classifiers are then used as prelearned classifiers $f_p(x)|_{p=1}^{T}$. T

The k is kernel function ,it is linear combination of base kernels km's, $k = \sum_{m=1}^{M} d_m k_m$, where dm is the linear combination coefficient, and the kernel function km is induced from the nonlinear, feature mapping function $\varphi_m(.)$. In A-MKL, the first objective is to reduce the mismatch in data distributions between two domains.

$$DIST_K^2(D^A, D^T) = \Omega(d) = h^T d \quad \text{(Eq. 4.4)}$$

Where h = [tr(K1S,....,tr(KMS)] , and $[\varphi_m(x_i)^T \varphi_m(x_j)] \in R^{NxN}$ is the mth base kernel matrix defined on the samples from both auxiliary and target domains.

The second objective of A-MKL is to minimize the structural risk functional. MKL methods utilize the training data and the test data drawn from the same domain. They come from different distributions, MKL methods may fail to learn the optimal kernel. This would degrade the classification performance in the target domain. On the contrary, A-MKLcan better make use of the data from two domains to improve the classification performance.

The matching results are obtained by using ASTPM method. Each pair of matched volumes from two videos is highlighted in the same color. The mismatch was measured by Maximum Mean Discrepancy (MMD) [2] based on the distance between the means of samples from the auxiliary domain DA and the target domain DT in the Reproducing Kernel Hilbert Space (RKHS), namely:

$$DIST_K(D^A, D^T) = \left\| \frac{1}{n_A} \sum_{i=1}^{n_A} \varphi(x_i^A) - \frac{1}{n_T} \sum_{i=1}^{n_T} \varphi(x_i^T) \right\|_H \quad \text{(Eq. 4.5)}$$

Where x^A 's and $x^T{}_i$'s are the samples from the auxiliaryand target domains, respectively. A-SVM [18,8]also assumes that the target classifier f^T (x) is adapted from existing auxiliary classifiers. Event in consumer video is recognized using large number of loosely labeled web videos and limited number of labeled consumer videos. Aligned Space Time Pyramid matching is used to find out the similarity between videos. Cross-domain learning method Adaptive Multiple Kernel Learning handles the mismatch between the data distributions of the consumer video domain and the web video domain.

V. Conclusion

A new event recognition framework for consumer videos are framed by leveraging a large amount of loosely labeled YouTube videos. A new pyramid matching method called ASTPM and a novel transfer learning method, A-MKL to better fuse the information from multiple pyramid levels and different types of local features and to cope with the mismatch between the

feature distributions of consumer videos and web videos. A possible future research direction is to develop effective methods to select more useful videos from a large number of low-quality YouTube videos to construct the auxiliary domain.

The adaption between the web domain and consumer domain studied in this work and other examples that vision researchers have recently been working on include the adaptation of cross category knowledge to a new category domain, knowledge transfer by mining semantic relatedness and adaption between two domains with different feature representations. In the future, this method will be extend to A-MKL for internet vision applications.

REFERENCES

[1] M. Brand, N. Oliver, and A. Pentland, *"Coupled Hidden Markov Models for Complex Action Recognition,"* Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 994-999, 1997.

[2] C.-C. Chang and C.-J. Lin, *"LIBSVM: A Library for Support Vector Machines,"* 2001.

[3] I. Laptev and T. Lindeberg, *"Space-Time Interest Points."* Proc. IEEE Int'l Conf. Computer Vision, pp. 432-439, 2003.

[4] P.A. Jensen and J.F. Bard, Operations Research Models and Methods. John Wiley and Sons, 2003.

[5] J.T. Kwok and I.W. Tsang, *"Learning with Idealized Kernels,"* Proc. Int'l Conf. Machine Learning, pp. 400-407, 2003.

[6] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan, *"Learning the Kernel Matrix with Semidefinite Programming,"* J. Machine Learning Research, vol. 5, pp. 27-72, 2004.

[7] P. Dolla´r, V. Rabaud, G. Cottrell, and S. Belongie, *"Behavior Recognition via Sparse Spatio-Temporal Features,"* Proc. IEEE Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65-72, 2005.

[8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, *"Actions as Space-Time Shapes,"* Proc. 10th IEEE Int'l Conf. Computer Vision, pp. 1395-1402, 2005.

[9] K. Grauman and T. Darrell, *"The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features,"* Proc.10th IEEE Int'l Conf. Computer Vision, pp. 1458-1465, 2005.

[10] Y. Ke, R. Sukthankar, and M. Hebert, *"Efficient Visual Event Detection Using Volumetric Features,"* Proc. 10th IEEE Int'l Conf. Computer Vision, pp. 166-173, 2005.

[11] S. Lazebnik, C. Schmid, and J. Ponce, *"Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories,"* Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.2169-2178, 2006.

[12] J. Blitzer, R. McDonald, and F. Pereira, *"Domain Adaptation with Structural Correspondence Learning,"* Proc. Conf. Empirical Methods in Natural Language, pp. 120-128, 2006.

[13] K.M. Borgwardt, A. Gretton, M.J. Rasch, H.-P. Kriegel, B.Scho¨lkopf, and A.J. Smola, *"Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy,"* Bioinformatics, vol. 22, no. 4, pp. e49- e57, 2006.

[14] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A.C. Loui, and J. Luo, *"Large-Scale Multimodal Semantic Concept Detection for Consumer Video,"* Proc. ACM Int'l Workshop Multimedia Information Retrieval, pp. 255-264, 2007.

[15] H. Daume' III, *"Frustratingly Easy Domain Adaptation,"* Proc. Ann. Meeting Assoc. for Computational Linguistics, pp. 256-263, 2007.

[16] A.C. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K.Lee, and A. Yanagawa, *"Kodak's Consumer Video Benchmark Data Set: Concept Definition and Annotation,"* Proc. Int'l Workshop Multimedia Information Retrieval, pp. 245-254, 2007.

[17] J. Hays and A.A. Efros, *"Scene Completion Using Millions of Photographs,"* ACM Trans. Graphics, vol.26, no. 3, article 4, 2007.

[18] I. Laptev, M. Marsza³ek, C. Schmid, and B. Rozenfeld, *"Learning Realistic Human Actions from Movies,"* Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[19] L. Duan, I.W. Tsang, D. Xu, and S.J. Maybank, *"Domain Transfer SVM for Video Concept Detection,"* Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2009.

[20] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang, *"Action Detection in Complex Scenes with Spatial and Temporal Ambiguities,"* Proc. 12th IEEE Int'l Conf. Computer Vision,pp. 128-135, 2009.

[21] N. Ikizler-Cinbis, R.G. Cinbis, and S. Sclaroff, *"Learning Actions from the Web,"* Proc. 12th IEEE Int'l Conf. Computer Vision, pp. 995-1002, 2009.

[22] L. Duan, D. Xu, I.W. Tsang, and J. Luo, *"Visual Event Recognition in Videos by Learning from Web Data,"* Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2010..

Author's Biography

Dr. D. Chitra is a Professor in the Department of Computer Science and Engineering at P. A. College of Engineering and Technology, Pollachi, Coimbatore. She received her M.E. Degree in CSE from Anna University, Chennai and PhD degree in CSE from Anna University of Technology, Coimbatore. Her resource interests include image analysis, Pattern recognition and Computer Vision. She is a life member of ISTE.

Mrs. R. Kavitha is a Assistant Professor in the Department of Computer Science and Engineering at P. A. College of Engineering and Technology, Pollachi, Coimbatore. She received her B.E. Degree in CSE from Madurai Kamaraj University, Madurai and M.E. degree in CSE from Anna University Chennai. Her resource interests include image analysis.