Transductive Transfer Learning for Action Recognition in Tennis Games

Nazli FarajiDavar, Teófilo de Campos, Josef Kittler, Fei Yan
CVSSP, University of Surrey, Guildford GU2 7XH, UK
N.Farajidavar@surrey.ac.uk
http://www.ee.surrey.ac.uk/CVSSP/

Abstract

This paper investigates the application of transductive transfer learning methods for action classification. The application scenario is that of off-line video annotation for retrieval. We show that if a classification system can analyze the unlabeled test data in order to adapt its models, a significant performance improvement can be achieved. We applied it for action classification in tennis games for train and test videos of different nature. Actions are described using HOG3D features and for transfer we used a method based on feature re-weighting and a novel method based on feature translation and scaling.

1. Introduction

In action recognition, as in most classification problems, systems are trained with samples from one setup and often expected to be applied to another setup. A system is trained with feature vectors $\mathbf{X}^{train} = \{\mathbf{x}_i : i \in train\}$ defined in a space $\mathcal{X}^{train}$, obtained, for instance, from videos of a set of actions with labels $\mathbf{Y}^{train} = \{y_i : i \in train\}$ defined in problem of space $\mathcal{Y}^{train}$. These samples $\mathbf{X}^{train}$ are obtained from a set of people in a set of environments, illumination conditions, camera configurations/types and with certain types of background. Let $\mathcal{D} = \{\mathcal{X}, p(\mathbf{X})\}$ be domain defined by the space and the marginal probability distribution, and let $\mathcal{T} = \{\mathcal{Y}, P(\mathbf{Y}|\mathbf{X})\}$ be the classification task. A pattern recognition system is expected to perform well if the test samples $\mathbf{X}^{test}$ are obtained with the same conditions as $\mathbf{X}^{train}$, *e.g.* if they are new videos of the same people performing the same actions in the same environments as before.

However, in most application scenarios there is a change of scene, video quality, performing actors *etc.*, so although $\mathcal{X}^{train} = \mathcal{X}^{test}$ and $\mathcal{Y}^{train} = \mathcal{Y}^{test}$, the domains are different because $p(\mathbf{X}^{train}) \neq p(\mathbf{X}^{test})$. The usual approach is to assume $D^{train} \approx D^{test}$, and treat this as a classical generalization problem in machine learning by means of classifier regularization [6].

We consider the application scenario of an autonomous system that is capable of detecting when a change of domain happens. An example is that of Almajai *et al.* [3], who present a method that detects *anomalies* if a system trained to automatically annotate videos of tennis singles is presented with videos of tennis doubles. Their system uses an effective ball tracker [24] to detect sequences of events with the HMM-based method of [2]. It also uses player action classification cues, but does not use the number of detected people as a cue so the analysis is purely based on the sequence of events.

Once a change of domain is detected, a system for automatic video annotation can easily start to gather data from the new domain in order to adapt the models for it. If the application considered allows off-line processing, such as video annotation for after-match analysis or for data retrieval, such a scenario is possible and we show that it leads to better performance on action classification.

This problem characterizes a case of *transductive transfer learning*, as defined by Pan and Yang [18]: we aim to improve the learning of the target predictive function $P(\mathbf{Y}^{trg}|\mathbf{X}^{trg})$ in $\mathcal{D}^{trg}$ using the knowledge in $\mathcal{D}^{src}$ and $\mathcal{T}^{src}$, where $\mathcal{D}^{src} \neq \mathcal{D}^{trg}$. For that, we evaluate two methods that transform the features of $\mathbf{X}^{src}$ so that they become more similar to $\mathbf{X}^{trg}$ and the classifier is re-trained using the transformed samples.

In the experiments of this paper, we assume that labels of all the source domain samples are available, so $src = train$. We also assume that all the target samples are available but none of their labels are, and the challenge consists in classifying all the target samples, so $trg = test$. But it is relevant to point out that once the transfer has been learnt, the classifier should become apt for application on unseen samples in the target domain.

In the next section we give an overview of related work

This manuscript is a pre-print submitted to the 3rd International Workshop on Video Event Categorization, Tagging and Retrieval for Real-World Applications (**VECTaR2011**), in conjunction with **ICCV2011**, ©IEEE. This project was sponsored by the EPSRC/UK grant EP/F069421/1 (ACASVA).

and follow by reviewing the work of Arnold *et al*. [4] in Section 3, which we use as a base. Next, we propose a new method in Section 4. In Section 5 we describe the experimental setup. The results are then presented in Section 6. This paper concludes in Section 7 where the contributions are highlighted and future work is discussed.

2. Related Work

The problem described above relates to a number of other problems in Machine Learning, such as semi-supervised learning and domain adaptation [18]. Perhaps the main difference is that we assume that a set of samples from the new domain $\mathbf{X}^{trg}$ is given all at once and the problem switches to classifying elements of this new domain, *i.e*., we do not necessarily require the new model to be applicable to samples in $\mathbf{X}^{src}$. Transductive transfer learning seems to have been dealt with by a relatively small niche of researchers, despite its broad range of applications.

Dai *et al*. [9] try to translate $\mathcal{X}^{src}$ into $\mathcal{X}^{trg}$ so that learning can be done within a single feature space. Their aim is to link the two feature spaces with the construction of a feature translator $p(\mathbf{X}^{trg}|\mathbf{X}^{src})$. This approach is not directly related to our problem because we assume the typical values $p(\mathbf{X})$ change, but the feature space remains the same $\mathcal{X}^{src} = \mathcal{X}^{trg}$.

A more related approach is that of estimating a low dimensional feature space $\mathcal{X}^{new}$ to which both source $\mathcal{X}^{src}$ and target $\mathcal{X}^{trg}$ spaces are mapped [16, 17, 7]. If a good mapping is found, the classifier learnt on the source domain will also be effective on the target domain. The downside is that these approaches may lead to loss of information and they use assumptions that may not generalize to all types of feature spaces.

Methods for domain adaptation can be formulated to apply to our problem. In speech recognition and audio processing, the problem of adaptation to new acoustic environments relates to adaptation of $p(\mathbf{X})$ [23]. One can use the maximum likelihood linear regression (MLLR) to estimate a set of linear transformations for the Gaussian parameters of the HMMs [14]. In [12], a vector Taylor series approach for HMM adaptation was introduced for decoding noisy utterances at test time. An intermediate step of noise adaptive model training is used and results in Pseudo-clean model parameters. In [19], Rodriguez *et al*. proposed an adaptation scheme for semi-continuous HMMs for unsupervised writer style adaptation in handwritten word spotting. In image segmentation, Maximum a Posteriori (MAP) adaptation was used for play field segmentation in order to re-estimate the GMM parameters [5]. One issue with all the approaches discussed in this paragraph is that they work on the assumption that the observation data follows a probabilistic generative model.

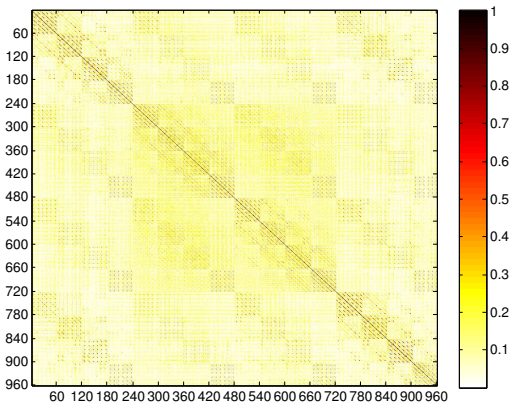

Figure 1. Absolute values of the correlation matrix of the HOG3D feature space (960 dimensions), obtained from the 1277 action samples of the tennis singles dataset. There appears to be some mild patterns at every 60 dimensions, probably because we used 3 temporal splits and icosahedrons for edge orientation quantization (see Section 5).

3. Transfer Learning by Feature Re-weighting

The feature extraction method used in this paper (see Section 5) gives a high dimensional space of features in the range $[0, 1]$ with relatively small correlation between each other (see Figure 1). Given the nature of this feature space, the method based on MaxEnt proposed by Arnold *et al*. [4] seemed to be very appropriate and it possibly is the least expensive in terms of computational cost. In this section we describe it and propose some revisions.

For ease of notation, x_j^i is the feature j of sample $\mathbf{x}_i$ and $E^{src}[x_j, y]$ is used to represent $E^{\mathbf{X}^{src}}[x_j, y]$ which is the expectation of having feature x_j with label equal to $y, \forall \mathbf{x} \in \mathbf{X}^{src}$:

$$E^{src}[x_j, y] = \frac{1}{N_{train}^{src}} \sum_{i=1}^{N_{train}^{src}} x_j^i \mathbb{1}_{[y]}(y_i) \,, \qquad (1)$$

N_{train}^{src} is the number of labeled training data in the source domain and

$$\mathbb{1}_{[y]}(y_i) = \begin{cases} 1 & \text{if } y_i = y, \\ 0 & \text{otherwise} \end{cases}$$

The problem in transductive tasks is that the joint distribution of the features with labels differs between the source and target domains, so $E^{src}[x_j, y]$ is not the same as $E^{trg}[x_j, y], \forall j \in 1, \cdots, D$, where D is the dimensionality of $\mathcal{X}$. If the expectations in the train and test datasets are similar, then the model Λ learnt on the training data will generalize well to the test data. In Arnold *et al*.'s algorithm [4] a transformation $G(\cdot)$ of the feature space $\mathcal{X}$ can be learnt such that the joint distributions of the source and target features with their labels are aligned:

$$E^{trg}[G(x_j), y] = E^{src}[G(x_j), y], \forall x_j \in \mathcal{X} \,. \qquad (2)$$

It can be too challenging (if not impossible) to estimate a single transformation of the feature space that would generate a space where $E^{trg} = E^{src}$. A possibility would be to have one transformation for the source and another for the target samples. This condition can even further be relaxed by arguing that it is enough to transform only one of the domains, say the source data, so that data from both domains could be separated by a single hyper-plane. In maximum entropy phraseology, the relaxed transformation is:

$$E^{trg}[x_j, y] = E^{src}[G(x_j), y], \ \forall x_j \in \mathcal{X} \ . \tag{3}$$

The problem with this, of course, is that in the unsupervised transductive transfer case, we do not have $\mathbf{Y}^{trg}$ and therefore cannot estimate $E^{trg}[x_j, y]$. Hence an approximation $E^{trg}[x_j, y]$ using the joint estimates on the target unlabeled data from a model learnt using the source data is applied, as proposed in [4]:

$$E^{trg}[x_j, y] \approx E^{trg}_{\Lambda_{src}}[x_j, y] = \frac{1}{N^{trg}_{test}} \sum_{i=1}^{N^{trg}_{test}} x^i_j P_{\Lambda_{src}}(y|\mathbf{x}_i) \ , \tag{4}$$

where N^{trg}_{test} is the number of target domain (unlabeled) test examples. Note that in [4], the authors use the joint probability $P_{\Lambda_{src}}(y, x_i)$, instead of the posterior $P_{\Lambda_{src}}(y|x_i)$. This approximation of $E^{trg}[x_j, y]$ may not reflect the true target expectation, but it is the best that can be done in the unsupervised transductive setting.

We suggest that more accurate definitions of $E[x_j, y]$ would have denominators depending on class labels/predictions and propose the following modifications of Equations (1) and (4):

$$E_{train}[x_j, y] = \frac{\sum_{i=1}^{N^{train}_{src}} x^i_j \mathbb{1}_{[y]}(y_i)}{\sum_{i=1}^{N^{train}_{src}} \mathbb{1}_{[y]}(y_i)}, \tag{5}$$

and

$$E^{trg}[x_j, y] \approx E^{trg}_{\Lambda_{src}}[x_j, y] = \frac{\sum_{i=1}^{N^{trg}_{test}} x^i_j P_{\Lambda_{src}}(y|\mathbf{x}_i)}{\sum_{i=1}^{N^{trg}_{test}} P_{\Lambda_{src}}(y|\mathbf{x}_i)} \ . \tag{6}$$

Based on these expectations the source domain transformation $G(\cdot)$ is defined as:

$$\forall i = 1\colon N^{src}_{train}, \ G(x^i_j) = x^i_j \frac{E^{trg}_{\Lambda_{src}}[x_j, y_i]}{E^{src}[x_j, y_i]} \ , \tag{7}$$

The effect is to re-scale x_j, giving more weight to features that occur frequently in the target but rarely in the source (in a conditional sense), and down-weighting features that are common in the source but seldom seen in the target [4].

In practice, since the target expectation $E^{trg}_{\Lambda_{src}}[x_j, y]$ is only approximate, the transformed features need to be smoothed with the original ones in each iteration as follows:

$$G'(x^i_j) = (1 - \theta)x^i_j + \theta G(x^i_j) \ , \tag{8}$$

where θ controls the degree to which we use the target conditional estimates to alter the source conditionals. Once the labeled samples have been transformed by $G'(\cdot)$, it is necessary to update the model Λ_{src} and retrain the classifier. If a kernel method is used, this also means the kernels have to be re-computed.

4. Translating and Scaling Features

The feature re-weighting scheme of Section 3 is probably ideal for binary feature spaces, but may be too simple for other types of features. We propose to translate and scale the samples of the training set based on the expected value and standard deviation of features for each class:

$$\forall i = 1\colon N^{src}_{train},$$
$$G(x^i_j) = \frac{x^i_j - E^{src}[x_j, y_i]}{\sigma^{src}_{j,y_i}} \sigma^{trg}_{j,y_i} + E^{trg}_{\Lambda_{src}}[x_j, y_i] \ , \tag{9}$$

where σ^{src}_{j,y_i} is the standard deviation of feature x_j of the source samples labeled as y_i and

$$\sigma^{trg}_{j,y_i} = \sqrt{\frac{\sum_{k=1}^{N^{trg}_{test}} (x^k_j - E^{trg}_{\Lambda_{src}}[x_j, y_i])^2 P_{\Lambda_{src}}(y_i|\mathbf{x}_k)}{\sum_{k=1}^{N^{trg}_{test}} P_{\Lambda_{src}}(y_i|\mathbf{x}_k)}} \ . \tag{10}$$

The smoothing function is then applied as before (Equation 8).

5. Experimental Setup

For the experiments in this paper, we used the tennis dataset described by de Campos *et al.* [11] and used the features obtained by the method described as space-time-shape (STS) in that paper. Videos of tennis games obtained from TV broadcast in standard resolution (SD) were processed and Yan *et al.*'s ball tracker [24] was used to detect relevant instances in time, i.e., when the velocity vector of the ball suddenly changes. The frame of those relevant instances were analyzed and using a method based on background subtraction, the players were detected.

For each detected player, a single HOG3D feature vector [13] was extracted. We used HOG3D [13] descriptors for our task because it was among the top performing methods evaluated in Wang *et al.*'s survey [22]. HOG3D is a three dimensional generalization of SIFT [15] or local histograms of oriented gradients (HOG) [10]. It uses polyhedral structures for the quantization of the 3-D spatio-temporal edge orientations to avoid the singularities in the use of polar coordinate systems (as performed in [20]). Another advantage of HOG3D [13] is its computational efficiency due to the use of three-dimensional integral images. The bounding box of each player is given, so a HOG3D feature vector is obtained by analyzing that bounding box in a buffer of 12 frames around each detected player.

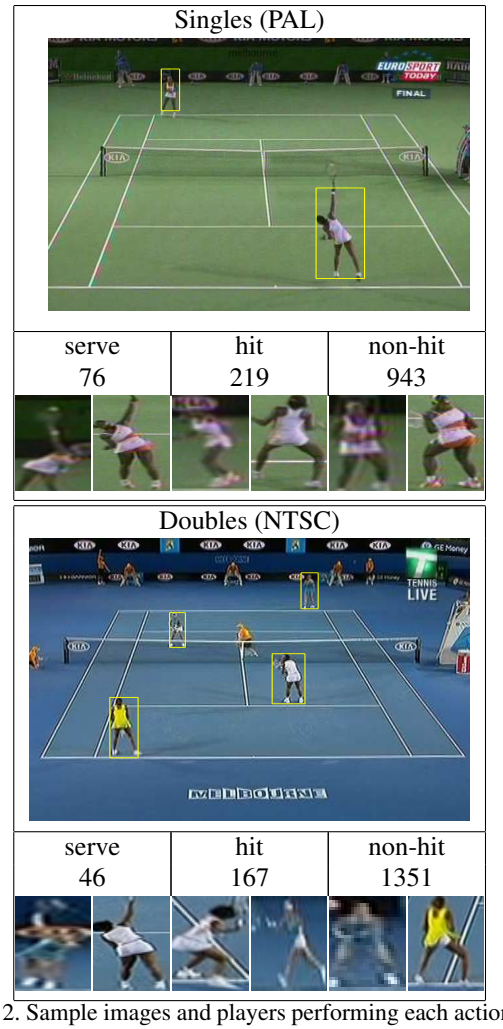

Figure 2. Sample images and players performing each action from de Campos *et al.*'s dataset, obtained from [11] ©IEEE 2011. The number of samples for each class is also shown under their label.

The parameters used for HOG3D were those optimized for the KTH dataset in [13]. They generate $960D$ vectors $(4 \times 4 \times 3 \times 20)$ using a 4×4 grid in space, 3 splits in time and, for each sub-block, a histogram of edge orientations is quantized using a icosahedron (20 faces regular polyhedron).

In de Campos *et al.*'s dataset, two videos were processed using the above procedure, a video of singles (recorded in PAL) and a video of doubles tennis (recorded in NTSC) [11]. They have different background, illumination conditions and players. In the video of singles, the players' scale is in general larger than that of the doubles' video. Figure 2 presents a sample frame of each of the videos and some sample actions. Actions are labeled as *serve*, *hit* and *non-hit*, and the dataset is quite unbalanced.

For classification, we followed [11] and used KLDA (Kernelised Linear Discriminant Analysis [8]). A minor difference is that instead of using the χ^2 measure to build the RBF kernel functions, we used the ℓ_1 distance because χ^2 is not a metric and ℓ_1 seem to be just as good as χ^2 to compare histogram-based feature vectors. Given the training kernels, KLDA generates a $|\mathcal{Y}| - 1$ dimensional space where samples are projected ($|\mathcal{Y}|$ is the number of classes). Due to the nature of this dataset, KLDA over-fits the data, so all the samples in the training set belonging to the same class are projected to the same point, making it impossible to estimate the covariance matrix for each class. This problem was not approached in [11] because the authors used KLDA as a discriminative classifier, *i.e.*, without using a generative data model. In this paper, in order to apply the transfer algorithms of Sections 3 and 4, it is necessary to estimate Λ_{src} of $P_{\Lambda_{src}}(\mathbf{Y}|\mathbf{X})$. For that, we used a five-fold cross validation in the training set to obtain estimates of the mean and covariance matrix for each class.

6. Results

As discussed in Section 5, the games of singles and doubles in the dataset are quite different from each other. In [11], the authors simply took the game of singles for training and the doubles for testing. In this paper, we evaluate the use of transductive transfer learning to improve the classifier performance.

First of all we reproduced the experiment of [11]. They evaluated the results in terms of mean Area Under the ROC Curve (mAUC) and obtained 90.3%. In the same experiments (training on singles, testing on doubles, without using transfer learning), we obtained an mAUC of 91.2%. Our baseline is slightly higher than that of [11] probably because of the use of a full generative model in $P_{\Lambda_{src}}(\mathbf{Y}|\mathbf{X})$ and because we use ℓ_1 instead of χ^2 for the kernel radial basis function.

In the analysis presented in the rest of this paper, we do not evaluate results using area under ROC curves because we used a generative MAP-based three-class classifier instead of using three discriminative binary classifiers trained in a one-vs-others fashion (used in [11]). A more appropriated performance measure was the mean accuracy (mAcc) for the 3 classes. The mAcc is measured by averaging out the correct classification rate for each class, giving equal importance to all classes, so it is not affected by data skew. On the same experiment as above, we obtained an mAcc of 58.72%, which seems quite low, but this is a challenging dataset.

Figure 3 presents the mAcc as a function of the transfer rate. It present results with Arnold *et al.*'s method [4] and shows that the rectifications of Equations 5 and 6 lead to much better results (see *reweight* curve), presenting a steady climb in mAcc as the transfer rate grows. The method of Section 4 (*trans+scale*) reaches a higher peak of classification performance, presenting an improvement of nearly 20%. The confusion matrix obtained with $\theta = 0.4$ is shown

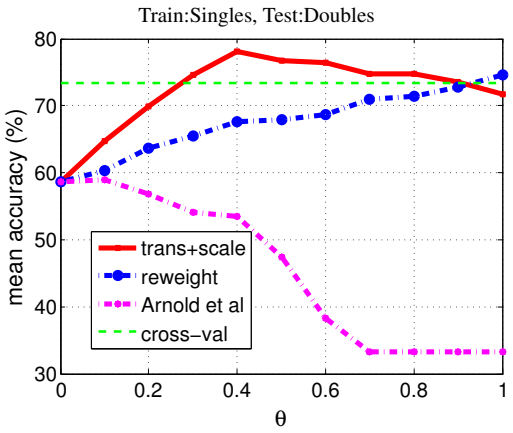

Figure 3. Mean accuracy obtained as a function of the transfer rate θ of (8), using $\mathbf{X}^{src}_{train}$ = singles and $\mathbf{X}^{trg}_{test}$ = doubles. The curve *Arnold etal* refers to the method of [4], *reweight* refers to its modifications proposed in Section 3 and *trans+scale* refers to Section 4. The best mAcc of those methods are 78.14% for *trans+scale* and 74.64% for *reweight*. The baseline (no transfer) gives 58.72% and the cross-validation on the test set gives 73.34%.

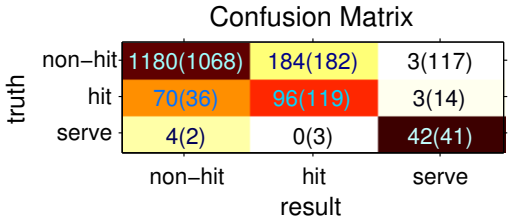

Figure 4. Best confusion matrix obtained with the *trans+scale* method ($\theta = 0.4$) on the tennis dataset, training with singles and testing with doubles. The numbers in brackets are results from [11].

in Figure 4. In brackets, the same figure also shows results presented by de Campos *et al*. in [11], who manually selected thresholds on the classifier output ("*thresholds selected so that the true positive rate is 77.62% and the false positive rate is 22.38%*" [11]). Note that our results are better at detecting *serves* and *non-hits*, but *hits* get confused with non-hits more often, whereas in [11] *non-hits* are often confused with serves.

After $\theta = 0.4$, the performance of *trans+scale* starts to drop possibly because its transformation is more complex and may lead to over-fitting to unlabeled data that was not necessarily classified correctly.

Figure 5 shows the results obtained by swapping the source/train and target/test samples. Note that with conservative transfer rate both methods increase the performance, but for $\theta \geq 0.4$, *trans+scale* leads to negative transfer.

For an additional analysis, we also performed a five-fold cross validation experiment on each domain to estimate what would be the best expected result of using transductive transfer. The resulting mAcc was: Singles: 92.29% and

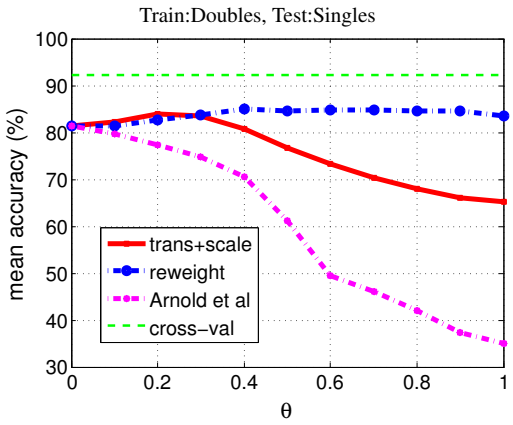

Figure 5. Same as Figure 3 but using $\mathbf{X}^{src}_{train}$ = doubles and $\mathbf{X}^{trg}_{test}$ = singles. The peak mAcc are: 83.89% (*trans+scale*), 84.99% (*scale*), the baseline is 81.30% and cross-validation on the test set gives 92.29%.

Doubles: 73.34%. These results are shown as *cross-val* in Figures 3 and 5.

Note in Figure 3 that both methods lead to a better performance than the cross-validation result on the target domain. In other words, both were able to achieve full benefit of having the test samples as unlabeled data for transfer. The same is not true in Figure 5. We suggest that this is because the doubles dataset is more unbalanced and noisier than the singles dataset, thus the cross-validation results on doubles are not so good but on singles they were very good. This also means that when the singles game is used as the test set, the baseline mAcc is much higher and the improvement of 3.69% in mAcc (obtained with *reweight* and $\theta = 0.4$) actually means a reduction of 5.07% in mean error rate, which is quite significant.

7. Conclusion

In this paper, we investigated a novel application of transductive transfer learning for video annotation. More specifically, we assume that a system that is able to detect a change of context is available and when such a change is detected, the system can start to gather new unlabeled samples. Once a set of samples is gathered, the system can apply methods of transductive transfer learning in order to adapt the models and to improve classification in this new domain.

We based our work on the method introduced in [4], where the source domain features are re-weighted based on the ratio of the joint expectation of features and class labels in target and source domains. We then proposed a modification for a more complex transformation, based on translation and scaling of each feature, for each class label.

We presented experiments in a dataset of action recognition in tennis games and showed that, in one scenario, the

proposed method can lead to an increase of nearly 20% in mean accuracy, giving results that are better than a 5-fold cross-validation on the test data set.

7.1. Future Work

An obvious next step is to perform experiments with more datasets for a broader evaluation of the application of transductive transfer learning.

The methods evaluated in this paper are based on transfer learning via feature space transformation. We plan to evaluate other modalities of transductive transfer via samples analysis, such as that of Acharya *et al*. [1], which combines ensembles of classifiers and clusterers to generate a more consolidated classifier. Another approach that should be investigated is that of transfer via hyperplane adaptation, such as that of transductive SVM [21].

References

[1] A.Acharya, J. E.R. Hruschka, and S.Acharyya. C3e: A framework for combing ensembles of classifiers and clusterers. In *10th Intenational Workshop On Multiple Classifier Systems*, pages 269–278, Naples, Italy, June, 2011. MCS.

[2] I. Almajai, J. Kittler, T. DeCampos, W. Christmas, F. Yan, D. Windridge, and A. Khan. Ball event recognition using hmm for automatic tennis annotation. In *Proceedings of Intl. Conf. on Image Proc.*, 2010.

[3] I. Almajai, F. Yan, T. de Campos, A. Khan, W. Christmas, D. Windridge, and J. Kittler. Anomaly detection and knowledge transfer in automatic sports video annotation. In *Proceedings of DIRAC Workshop, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010)*, 2010.

[4] A. Arnold, R. Nallapati, and W. W. Cohen. A comparative study of methods for transductive transfer learning. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, ICDMW '07, pages 77–82, Washington, DC, USA, 2007.

[5] M. Barnard, Odobez, and Jean-Marc. Robust playfield segmentation using map adaptation. In *17th International Conference on (ICPR'04) Volume 3*, pages 610–613, Washington, DC, USA, 2004. IEEE Computer Society.

[6] C. Bishop. *Pattern Recognition and Machne Learning*. Springer, 2006.

[7] J. Blitzer, R. Mcdonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *In EMNLP*, 2006.

[8] D. Cai, X. He, and J. Han. Efficient kernel discriminant analysis via spectral regression. In *International Conference on Data Mining*, 2007.

[9] W. Dai, Y. Chen, G. rong Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, pages 353–360, 2008.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 2005.

[11] T. de Campos, M. Barnard, K. Mikolajczyk, J. Kittler, F. Yan, W. Christmas, and D. Windridge. An evaluation of bags-of-words and spatio-temporal shapes for action recognition. In *IEEE Workshop on Applications of Computer Vision (WACV)*, Kona, Hawaii, January 2011.

[12] O. Kalinli, M. L. Seltzer, and A. Acero. Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:3825–3828, 2009.

[13] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *British Machine Vision Conference*, pages 995–1004, sep 2008.

[14] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9:171–185, 1995.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int Journal of Computer Vision*, January 2004.

[16] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, pages 677–682. AAAI Press, 2008.

[17] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[18] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

[19] J. A. Rodriguez, F. Perronnin, G. Sanchez, and J. Llados. Unsupervised writer style adaptation for handwritten word spotting. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008.

[20] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc of the ACM Multimedia*, Augsburg, Germany, September 23–28 2007.

[21] G. Teng, Y. Liu, J. Ma, F. Wang, and H. Yao. Improved algorithm for text classification based on tsvm. In *Proceedings of the First International Conference on Innovative Computing, Information and Control - Volume 2*, ICICIC '06, pages 55–58, Washington, DC, USA, 2006. IEEE Computer Society.

[22] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc 20th British Machine Vision Conf, London, Sept 7-10*, 2009.

[23] P. Woodland, M. Gales, and D. Pye. Improving environmental robustness in large vocabulary speech recognition. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:65–68, 1996.

[24] F. Yan, W. Christmas, and J. Kittler. Layered data association using graph-theoretic formulation with application to tennis ball tracking in monocular sequences. *Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1814–1830, 2008.