Transfer Learning for Human Action Recognition

Ana Paula B. Lopes*†, Elerson R. da S. Santos*, Eduardo A. do Valle Jr‡,
Jussara M. de Almeida* and Arnaldo A. de Araújo*
*Depart. of Computer Science - Universidade Federal de Minas Gerais (UFMG) Belo Horizonte (MG), Brazil
(paula,elerson,jussara,arnaldo)@dcc.ufmg.br
†Depart. of Exact and Tech. Sciences - Universidade Estadual de Santa Cruz (UESC) Ilhéus (BA), Brazil
‡Institute of Computation - Universidade Estadual de Campinas (UNICAMP) Campinas (SP), Brazil
mail@eduardovalle.com

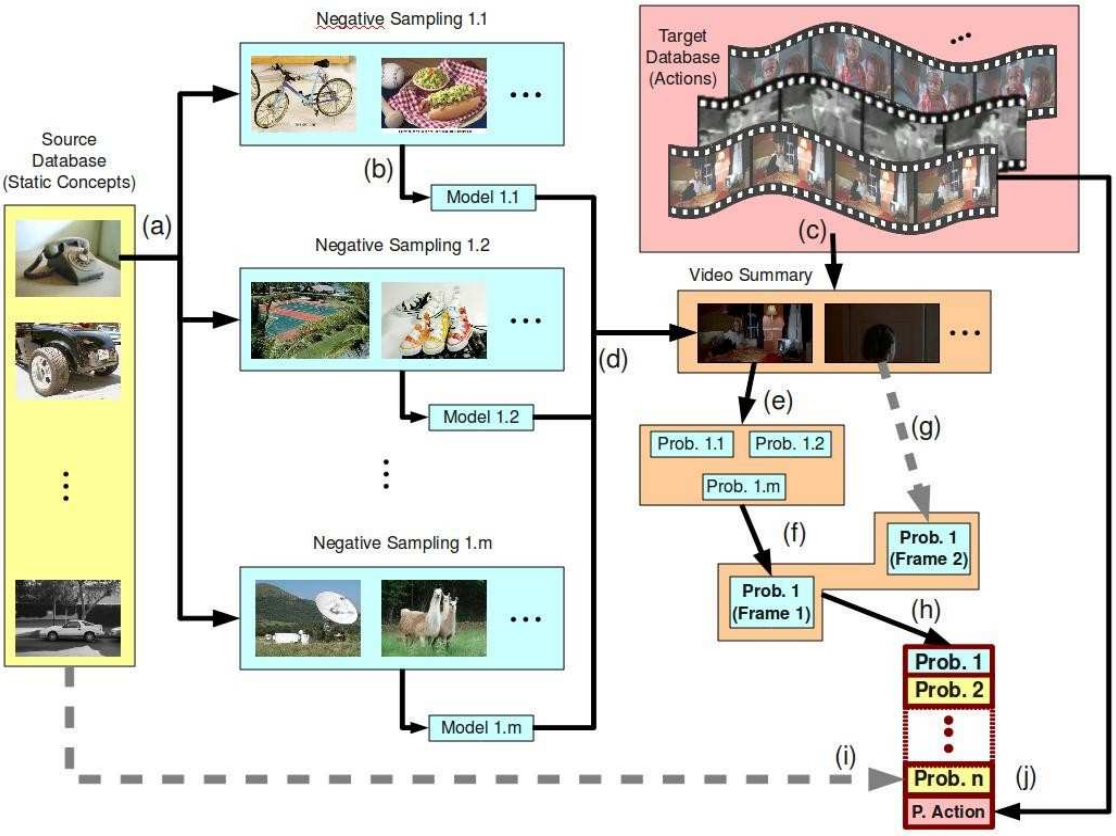


Fig. 1. Overview of the feature extraction process for the proposed transfer framework. Firstly, several models are created from an auxiliary (source) database (a–b), and applied to the target (action) database (C–d). The results of those models are combined in several ways (e, f, g) and then used as features for the final action classifier (h), together with the results of the baseline classifier (j). More details will be provided in Section IV

***Abstract*—To manually collect action samples from realistic videos is a time-consuming and error-prone task. This is a serious bottleneck to research related to video understanding, since the large intra-class variations of such videos demand training sets large enough to properly encompass those variations. Most authors dealing with this issue rely on (semi-) automated procedures to collect additional, generally noisy, examples. In this paper, we exploit a different approach, based on a Transfer Learning (TL) technique, to address the target task of action recognition. More specifically, we propose a framework that transfers the knowledge about concepts from a previously labeled still image database to the target action video database. It is assumed that, once identified in the target action database, these concepts provide some contextual clues to the action classifier. Our experiments with Caltech256 and Hollywood2 databases indicate: a) the feasibility of successfully using transfer learning techniques to detect concepts and, b) that it is indeed possible to enhance action recognition with the transferred knowledge of**

even a few concepts. In our case, only four concepts were enough to obtain statistically significant improvements for most actions.

***Keywords*-action recognition; transfer learning; bags-of-visual-features; video understanding;**

I. Introduction

An important bottleneck in research on action recognition in realistic videos resides in the lack of large enough training data bases for the Machine Learning (ML) algorithms involved. In this work we propose a framework for action recognition in videos which rely on external auxiliary databases, drawing on the effort already spent in labeling those databases and avoiding the cost to produce more training action samples. To do this, we rely on the ideas of transfer learning techniques [1].

Transfer learning allows transfer of knowledge between different databases by relaxing the classical assumption of learning algorithms that training and testing data might come from the same probability distribution. Our experiments indicate that knowledge transfer between quite different visual databases is feasible under the assumptions of transfer learning. Also, they show that even a challenging action database as Hollywood2 [2] can benefit from the knowledge of a few concepts brought from the Caltech256 [3] still-image database.

The assumptions underlying this work are: (a) transfer learning between a concept image database and a action video database is possible; b) by recognizing the presence/absence of a series of concepts along the video segment can bring relevant contextual clues to action videos; (c) the context in which an action occurs can help on recognizing the actions.

II. Related Work

A. Dealing with Scarce Training Data

In the pursuit of finding a scalable way for collecting training samples for different actions in movies, both [4] and [2] use textual information coming from movie scripts to automatically segment actions and scenes samples. Their procedure is further refined in [5], in which a clustering algorithm similar to k-means [6] is applied to temporal sub-windows of the initial video clip, in order to narrow it into a more precise action location in time.

In [7], the issue of collecting a large-enough amount of training data for high-level video retrieval is addressed by the usage of videos collected from Youtube[1] filtered by pre-defined categories and tags.

The resulting annotations achieved in those cases are quite noisy, although both the experiments reported in [4] and in [7] indicate that they are still able to produce classifiers above the chance level.

B. Transfer Learning for Images and Videos

In [8], it is proposed a solution to overcome the lack of training data for leaf image classification by using as an auxiliary external database images from an herbarium (the main database are images of a few individual fresh leaves, while in the herbarium – which contains much more examples – the leaves are dried and can be tied together in groups).

Another example of transfer learning in Computer Vision – yet in still images – is provided by [9], for face recognition. In that paper, they observe that examples of generic faces traits (e.g., lips, moustache, curly hair) are easier to collect than examples of every specific people in the database. So, instead of training a face classifier based on people's examples, they train 65 classifiers on selected traits and use the scores provided by such classifiers to compute the feature vectors.

The authors of [10] propose a series of pre-computed classifiers from several auxiliary data sources to concept detection on TRECVID2005 videos. They used three low level features (Grid Color Moments, Gabor Textures and Edge Histograms) extracted from keyframes for these classifiers. The final classifier using the auxiliary ones outputs is a modified version of least squares SVM (LS-SVM).

In [11], it is proposed a bird's eye view representation to transfer policy knowledge between keepaway robot soccer setups, with different number of objects in the scene.

To the best of our knowledge, no author has made use of transfer learning for action recognition in videos.

C. Contextual Information in Action Recognition

In [2], the usage of contextual information in aiding to recognize related actions is addressed. In that work, relevant scene types are learned from the texts of movies scripts. Scene and human-actions are then described as Bags-of-Visual-Features (BoVFs)[12] and *off-the-shelf* Support Vector Machines (SVM) classifiers are trained separately for both kinds of concepts (scenes and actions). Finally, the correlation between scenes and actions is explored to enhance the recognition results both for scenes and actions.

In [13], the context for events selected from the Large-Scale Concept Ontology for Multimedia (LSCOM) event/activity annotations [14] is captured by the trajectories of videos in the concept space, which is defined by a number of concepts selected from LSCOM-lite annotations ([15]). Concept detectors are trained on low level features like color, texture, motion and edges, and the trajectories themselves are analyzed by Hidden-Markov Models (HMM), whose scores are used to form a feature vector for the final SVM classifier.

The work of [16] explores context at lower levels. In that work, contextual information is encoded by different contextual descriptors computed on trajectories of Scale-Invariant Feature-Transform (SIFT) points [17]. Such descriptors are combined with different spatio-temporal grids in a multi-channel kernel scheme similar to that applied in [4].

Interestingly enough, the comparison among several 3D point detectors and descriptors performed by [18] concluded that, except for the KTH [2] database [19], regularly spaced dense samplings of points perform better at action recognition than interest points. This result reproduces similar ones obtained for scene recognition by [20], and can be considered

[1] http://www.youtube.com

[2] Acronym for Royal Institute of Technology in Swedish

as additional indirect indications of the importance of context in realistic videos, since denser samplings, by covering larger areas, should be better able to capture contextual information. The distinct result for KTH can be explained by the fact that it is a controlled database, whose backgrounds are almost completely uniform. Hence, there is virtually no context to be described in KTH.

III. Transfer Learning

Classical Machine Learning (ML) algorithms rely on the assumption that training and test data come from the same probability distribution. In fact, though, it is rare a case of practical application in which such an assumption is concretely guaranteed. Most applications of ML randomly splits the available data between training and test sets (or among validation, training and testing sets) and assume that future real data to which the trained algorithm will be applied will follow the same distribution.

Nevertheless, as it is argued in [1], there are plenty of real-world examples in which such assumption is not realistic. A typical example is the classification of web pages, in which data is constantly changing, thus letting the underlying trained algorithms outdated. They go on by citing sentiment classification of products in shopping websites. Such classification is based on user review data, which can present huge variations among different categories of products. In this case, a sentiment classification algorithm trained on a few products will probably fail on most other products, and even specialized classifiers will not be able to cope with the variations of perceptions of the product users along time. In some cases, as in the one tackled in this work, there is simply not enough available labeled data in the target database, and its acquisition is too expensive and error-prone.

Transfer Learning (TL) techniques come up to deal with those kinds of applications, by allowing distributions, tasks and domains of training and test data to vary. The notation, definitions and classification of TL algorithms of [1] are going to be applied in most of this review.

According to them, a *domain* $\mathcal{D}$ is composed of a *feature space* $\mathcal{X}$ and a *probability distribution* over that space $\mathcal{P}(\mathcal{X})$, and can be mathematically defined by:

$$\mathcal{D} = \{\mathcal{X}, \mathcal{P}(\mathcal{X})\} \tag{1}$$

A *task* $\mathcal{T}$ is defined by a *label space* $\mathcal{Y}$ and a *prediction function* $f(.)$ which is the equivalent to the conditional probability $\mathcal{P}(\mathcal{Y}|\mathcal{X})$:

$$\mathcal{T} = \{\mathcal{Y}, \mathcal{P}(\mathcal{Y}|\mathcal{X})\} \tag{2}$$

The target domain $\mathcal{D}_t$ is the domain of interest for the application, usually with few or no labeled data at all. A source domain $\mathcal{D}_s$ is an auxiliary source of information, generally – but not always – with a great amount of labeled data, which hopefully can be used to aid the ML algorithm to be adequately trained for the target task.

From those definitions, three main categories of TL algorithms can be identified:

Inductive Transfer Learning

Inductive Transfer Learning (ITL) concerns those cases in which the source and target tasks differ, no matter their domains. The most common case of ITL is when there is a lot of labeled data in the source domain, which makes this type of ITL similar to multitasking learning. For example, [21] proposes the TrAdaBoost extention of the AdaBoost algorithm to deal with new instances coming in an already trained system, potentially provoking a continuous distribution change.

A less common case of ITL occurs when there is a related source domain, but their instances are all unlabeled. An example of such a technique more known as self-taught learning – can be found in [22] in which a image representation is learned from unlabeled instances, to be applied in the target domain (whose task is also unsupervised).

Transductive Transfer Learning

Transductive Transfer Learning (TTL) involves cases in which the tasks are the same, but the domains vary. Looking at Equation 1, it is possible to see that different domains can vary in two aspects: they can have different feature spaces (for example, text classification for different languages) or they can share the same feature space but have varying probability distributions (for example, text classification on different specialized databases).

Unsupervised Transfer Learning

Unsupervised Transfer Learning (UTL) techniques are developed to the cases in which there is no labeled data on neither source or target domains.

In all cases of TL, the knowledge can be extracted from the source instances, from learned feature representations or from model parameters. In case of relational datasources, it is possible to occur the transfer of relational knowledge either.

IV. Transferring Concepts from Still Images to Videos

The transfer framework proposed in this work is depicted in Figure 1. It illustrates the complete feature extraction process, including information both from the source database – in our case Caltech256 concept images database – and from the target database – action videos Hollywood2 database – until the final representation for the video is achieved, that can be submitted to the final SVM classifier.

The process is performed as follows: in (a) m negative samplings, equaling the number of positive ones are prepared. In (b) static concept models are built for each negative sample, with the libsvm software[23], using a pyramid kernel with 2 levels [24]. Meanwhile, (c) shows that a video summary is created according to a simplified version of the algorithm presented in [25]. Most summaries for the video fragments ended up composed of 2-4 frames. In (d), the models created in (b) are applied to the summary frames, providing (e) m different probabilities for each concept, for each summary

frame. In (f), the replications for each concept are combined to provide a unique result for each frame. The arrow indicated by (g) indicates the repetition of the process described from (a) to (f) for every frame in the summary. In step (h) the probabilities for each frame are combined at a video level. The combination is done by max certainty[3], since we are interested to know whether the concept appears or not in the entire video. The arrow (i) indicates the repetition of the process described from (a) to (h) to every concept in the database. Finally, the video has a representation composed of n concept probabilities added by the probability of the baseline action classifier. That baseline probability is computed from the SIFT-based BoVFs [4].

V. Experiments

The instantiation of the main elements of the proposed framework for the experiments are described in Table I. In all the experiments, the number of replications is $m = 5$.

Two sets of experiments were conducted in order to evaluate the assumptions of this work (stated at the introduction). We address those assumptions by trying to specifically answer to the questions addressed in the next two sections.

TABLE I
Source and target databases used in the experiments

Generic Element	Instantiation
Source database	Caltech256 photos, described by HueSIFT binary BoVFs. To introduce some geometric information to the representation, a two-level pyramid kernel [24] was applied. The first experiments use rotary-phone examples only and the concepts used in the second row of experiments are: car-tire, car-side, rotary-phone, telephone-box.
Target database	Hollywood2 videos, described by STIP binary BoVFs, classified by a multilevel kernel as that of [4]. In the first set of experiments, only frames of AnswerPhone against all other frames are tested. In the second set of experiments, all actions are tested.

A. Is it possible to transfer knowledge from Caltech256 to Hollywood2?

Firstly, we needed to assess whether image-based classifiers would be able to identify the same concepts in videos frames from an unrelated collection. To establish this, Caltech256 rotary-phone images where used to train five ($m = 5$) classifiers with different negative samples. In addition, to verify the idea of representation transfer (Section III), the visual vocabulary was randomly selected from: (a) the source-only; (b) the target only and, (c) a mixed selection of visual words from both source and target samples.

Another evaluation was performed at a kernel level, and the following configurations were tested: linear, χ^2, multilevel [4] and pyramidal kernel [24], all using both one and two-level representations.

[3]maximun distance from probability=0.5

Such classifiers were applied in a phone baseline database built on frames of Hollywood2 actions database. Such baseline was built by manually collecting 90 positive examples of images presenting phones from the summaries of the Hollywood2 action class *AnswerPhone*. The same amount of negative examples were randomly chosen among frames coming from summaries of videos of all other action classes. The summarization algorithm is a simplified version of that presented in [25]. Hollywood2 original training and test sets separation - which come from different movies – was maintained.

Our aim in these first round of experiments is to verify the viability of transfer from Caltech256 to Hollywood2 in the concept level, showing how a transfer classifier would compare with a *classical* one – the baseline. It is interesting to notice that this baseline provides a *ceil* to the transfer result (instead of a *floor*), given that – from the point of view of traditional ML theory - the non-transfer setup, with training and testing coming from the same distribution is the *ideal* one. In other words, the transfer classifier is supposed to work, at best, slightly worse than the baseline. Thus, the main advantage of using a transfer-based classifier is that it opens the possibility of using any additional sources of information which would otherwise be inaccessible to a classical classifier. Hopefully, such additional information would compensate the bias introduced by the extraneous source dataset.

B. Does detecting Caltech256 concepts on Hollywood2 frames enhance action recognition?

In this second round of experiments, the feature vectors extracted by the procedure described in Figure 1 for every video are submitted to a new SVM classifier, this time with a kernel based on χ^2 distances (the same of [4], but using a unique channel).

Similarly to the case with the concepts, not all the training set were used at once for training the baseline. Instead, all the positive samples for each action were taken together with a random selection of negative samples of the same size. This more lightweight experimental setup (when compared to a typical full Hollywood2 classification setup as those in [4], [2]) was chosen to speed-up the verification of our main assumption: the ability of Caltech256 concept classifiers to enhance action classification on Hollywood2.

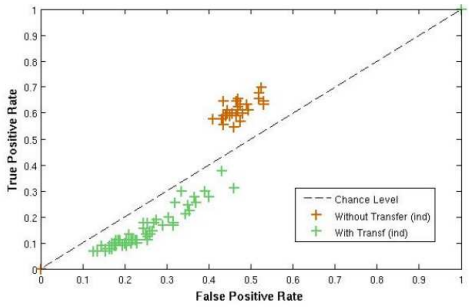
VI. Results and Discussion

A. Transferring from Caltech256 images to Hollywood2 frames

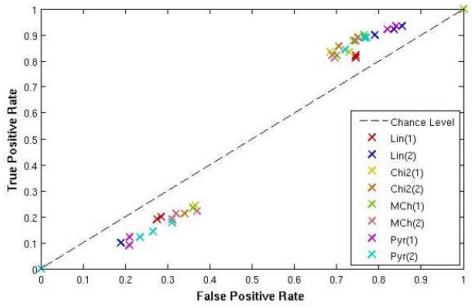
In this section we show that, indeed, knowledge transfer from Caltech256 to Hollywood2 frames is feasible.

Figure 2 (a) shows the individual results for Phone classification in Hollywood2 frames, with and without transfer. As suggested by [1], such 'brute force' transfer led to negative transfer for individual classifiers. Nevertheless, in Figure 2 (b) the individual results for transfer settings are combined, using five replications per kernel configuration, indicating that in some cases it is possible to achieve results above chance level, although skewed to the positive side. It is possible to observe

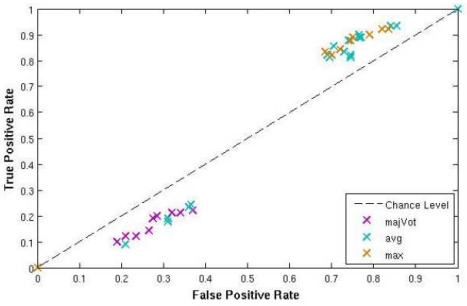
also that the less skewed results tend to be provided mostly by the more sophisticated kernels, namely those proposed in [4] and [24]. Figure 2 (c) shows the same data colored by combination scheme, showing that average and maximum distance from the mid-point probability are those responsible for the best transfer results.



(a) Individual Results, showing that initially, there is negative transfer.



(b) Results combined (colored by kernel), showing that model combination overcomes negative transfer.

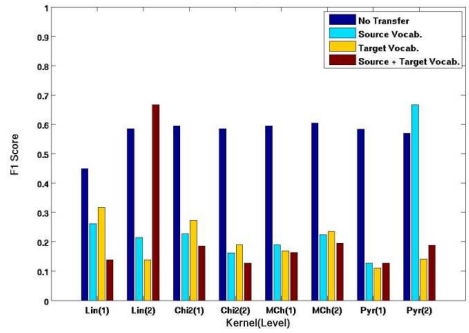


(c) Results Combined (colored by comb. strategy), showing that average and maximum distance from the p=0.5 work better than majority voting.

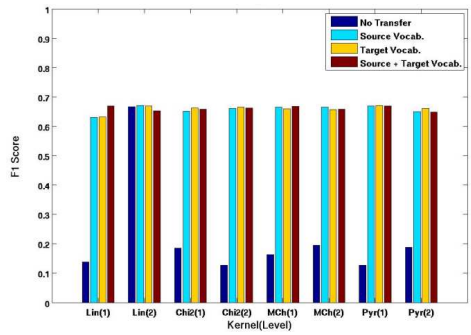
Fig. 2. Transfer of Knowledge about Phones (k=400, m=5)

Replications of these experiments with $k = 4000$, indicated no expressive enhancement over $k = 400$, while the results obtained with $k = 100$ individual tend to present stronger fluctuations, thus becoming less reliable. Their equivalent ROC graphs are not shown due to lack of space.

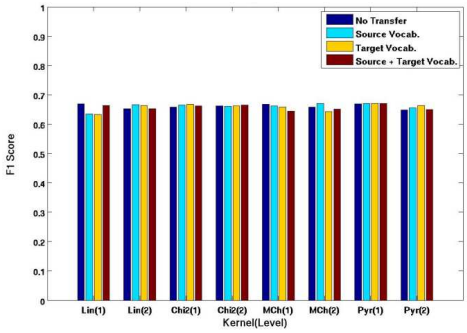
Figure 3 shows the F1 score between the baseline and the results with transfer from Caltech256, using three different sources for the visual vocabulary. Each graph illustrates a different combination scheme for the classifiers. In Figure 3 (a) individual classification results are combined by majority voting for each item; In Figure 3 (b), the probabilities of positive samples provided by the classifiers are averaged. In Figure 3 (c) the combined classifier relies on the individual classifier whose probability is farthest from the middle-point ($p = 0.5$), meaning that in this scheme the final classifier uses the result of the classifier which has the greatest certainty of its response for that item.



(a) Majority Voting is generally better for non-transfer setups.



(b) Average Probability is generally better for transfer setups.



(c) Max Distance from p=0.5 works equally well for both transfer and non-transfer setups

Fig. 3. F1 Scores show that kernel and vocabulary sources do not present significant differences for transfer, while the combination scheme can greatly influence transfer results. (k=400, m=5)

Dark blue bars represent the baseline results (without transfer) while the other colors indicate different visual vocabulary sources. Each group of bars was computed using a different kernel.

In terms of combining schemes, these graphs show that: a) majority voting seems to be appropriate only for traditional classification setups, while being unstable in transfer setups; b) using the average probability for decision is better for transfer setups than to a traditional one and, c) the usage of the greater certainty classifier seems to be the most stable between transfer and traditional setups.

In addition, these graphs reinforce how insignificant are the differences among kernel configurations or vocabulary sources. Such insensibility to vocabulary sources is an important result for our transfer setup, since it means that there is no transfer of representation knowledge and positive results can be obtained simply by brute-force transfer followed by classifier combinations which take into account the probabilities provided by libsvm, instead of its binary (i.e., positive/negative) results.

B. Using transfer to improve action recognition

Based on the results of the previous experiments, it is now assumed that it is viable to transfer at least some knowledge obtained from the source database (Caltech256) to the frames of the target database (Hollywood2) at a per concept basis. The results of this new set of experiments are shown in Figure 4, which presents the differences between precision values of the transfer results in relation to the no-transfer baseline (STIP-only) showing an increase in the majority of precision values when transfer is applied.

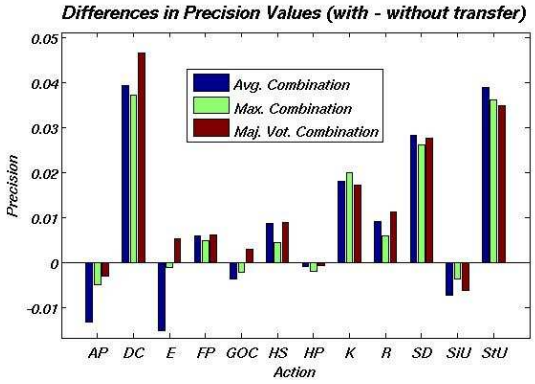


Fig. 4. Indicates that even with only ***4 (four)*** concepts used for transfer, ***most precision values increase***, no matter the classifier combination scheme applied for the concept classifiers. Actions: AP - AnswerPhone, DC - DriveCar, E - Eat, FP - FightPerson, GOC - GetOutCar, HS - HandShake, HP - HugPerson, K - Kiss, R - Run, SD - SitDown, SiU - SitUp, StU – StandUp. Transfer Concepts: car-tire, car-side, rotary-phone, telephone-box

Table II shows the differences for the average combination scheme and their statistical significance. From Figure 4 and Table II it is possible to see that there are 4 decreases in precision only, out of 12, and three of them are not significant. For the the precision increases (8 out of 12), only *FightPerson* and *SitUP* results are insignificant. In other words, 6 out of 8 increases in precision are statistically significant.

Observing Figure 4 for different actions, it is possible to see that the *DriveCar* action had the largest enhancement, what should be expected, since two transfer concepts in this experiment were related to cars (*car-tire* and *car-side*). However, unexpected negative results come up for *AnswerPhone* and *GetOutCar*, since the transfer concepts were semantically related to them. In case of *AnswerPhone*, the insignificant impact of the phone concept can be explained by the fact that telephones usually appear as small objects and play a small part in the context.

The *GetOutCar* action is notoriously tricky in Hollywoo2, since there are scenes in which a car does not even appears in the scene, or the action is extremely occluded, as can be seen in Figure 5.

Fig. 5. A sequence showing a difficult example of *GetOutCar* action. Observe that the car is presented in a frontal point-of-view (instead of *car-side* and the visual information related to the action of getting out of the car is very small (magenta arrow).

The positive results of the other classes (*HandShake*, *Kiss*, *Run*, *SitDown* and *StandUp*) could also be considered somewhat unexpected, given the apparent *unrelatedness* of them with the selected transfer concepts. Such results reinforces our main thesis that general concepts can have a positive impact on the overall action recognition performance, no matter the actions. In addition, it also reinforces our hypothesis that even apparently unrelated concepts are able to convey indirect contextual clues to the action classifier.

Figure 6 shows the the ROC points for the *DriveCar* action, in which the gains in precision were the largest ones.

Finally, the same graphs for *GetOutCar* action (Figure 7) show another interesting effect of using transfer information, which was found in other actions either (not shown due to lack of space). In this case, although the average precision gain is very small or negative (depending on the combination scheme), it is possible to see how the transfer-based classifiers tend to be less biased than the baseline classifier, without transfer.

TABLE II
DIFFERENCES IN PRECISION VALUES WITH THEIR STATISTICAL SIGNIFICANCE AT A 95% LEVEL

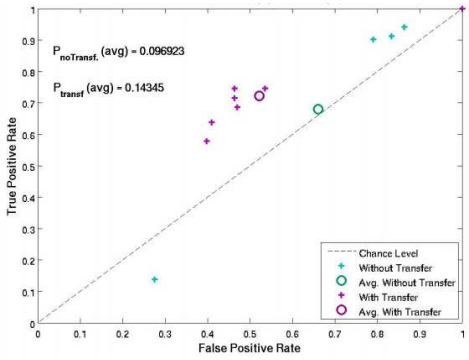
Action	Difference	Significant?
AnswerPhone	-0.0132	no
DriveCar	0.0395	**YES**
Eat	-0.0152	**YES**
FightPerson	0.0060	no
GetOutCar	-0.0036	no
HandShake	0.0087	**YES**
HugPerson	-0.0009	no
Kiss	0.0181	**YES**
Run	0.0090	**YES**
SitDown	0.0283	**YES**
SitUp	-0.0073	no
StandUp	0.0390	**YES**

VII. CONCLUSION

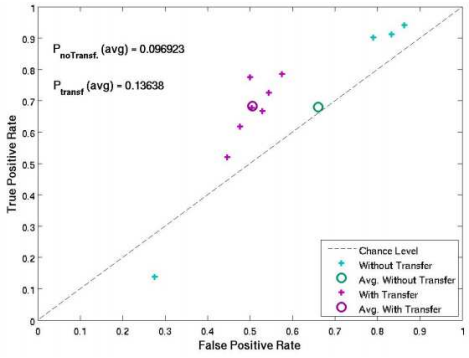
This paper aimed at assessing the hypothesis that Transfer Learning makes it possible to use the information contained in already annotated databases to improve action recognition results in challenging, realistic videos. In a first step, an image-to-frame transfer evaluation was performed using the Caltech256 as the source concept database and Hollywood2 as the action target database, indicating that the transfer of knowledge about concepts was possible at frame level, despite the different distributions between source and target databases.

The second round of experiments was built on the insights obtained in that first round, and assuming that a) concepts carry contextual information of video, and b) contextual clues can improve action recognition in realistic videos. For these proof-of-concept experiments, we selected four source concepts from Caltech256 and devised a scheme to apply that knowledge directly for action recognition in the Hollywood2 target database. The scheme depicted in Figure 1 was applied and most differences in action recognition with transfer were positive and statistically significant, indicating that transfer learning assumptions can be successfully applied to action recognition in videos.

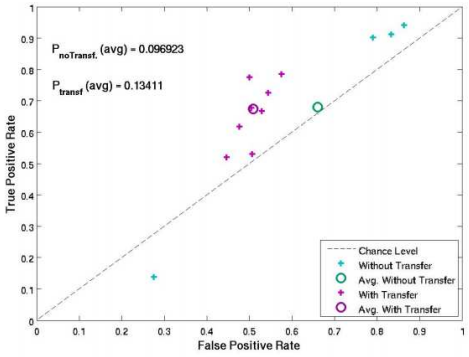
The simplified experimental setup selected prevents a direct comparison to other recognition rates for Hollywood2 found in the literature. Their most important outcomes are the strong indications that such transfer-learning-based framework can effectively enhance action recognition rates. Those outcomes provides us with clear directions for future work: to test our transfer framework on a full action recognition scheme and compare our results with state-of-the art. This is meant to be done not only with Caltech256/Hollywood2 databases, but other also with other source/target database pairs. Some of the unexpected results also provide us with a few additional interesting lines of future investigation, such as how the relationship between the concept classification task and the action classification task is established. Finally, some parameters which were fixed in our experiments deserve some more attention, such as the number of replications, the number of concepts, the criteria for concept selection, and even the number of source databases, given that there is a wealth of concept databases out there that could also be added to Caltech256 in our framework without the need of further modifications.



(a) Majority Voting



(b) Average Probability



(c) Max Distance from p=0.5

Fig. 6. *DriveCar* Transfer Results (k=400, m=10). Observe how the information introduced by transfer move the average precision away from the chance level line.

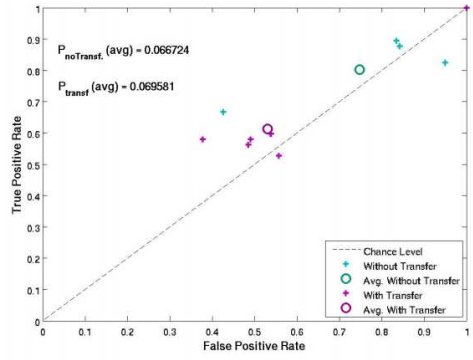
ACKNOWLEDGMENT

The authors are grateful to FAPEMIG, FAPESP, CNPq and CAPES, Brazilian research funding agencies, for the financial support to this work.

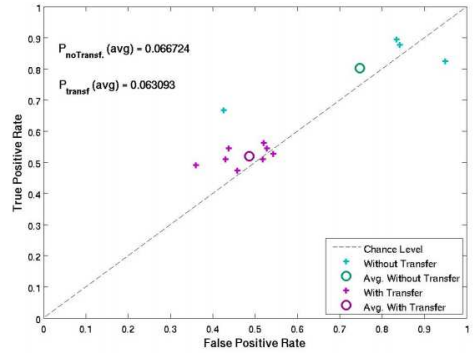
REFERENCES

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *Transactions on Knowledge and Data Engineering (pre-print)*, 2009.

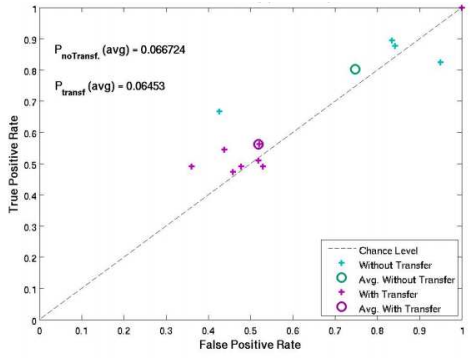
[2] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR '09*, June 2009, pp. 2929–2936.



(a) Majority Voting



(b) Average Probability



(c) Max Distance from p=0.5

Fig. 7. *GetOutCar* Transfer Results (k=400, m=10)

[3] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007. [Online]. Available: http://authors.library.caltech.edu/7694

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR '08*, June 2008, pp. 1–8.

[5] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic annotation of human actions in video," in *ICCV '09*, 2009.

[6] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[7] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel, "Learning automatic concept detectors from online video," *Computer Vision and Image Understanding*, vol. In Press, Corrected Proof, pp. –, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6WCX-4X1J787-3/2/944190566d7103b11000f88dcc2eb526

[8] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, August 2004.

[9] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV '09*, 2009.

[10] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua, "Domain adaptation from multiple sources via auxiliary classifiers," in *Proceedings of the 26th International Conference on Machine Learning*, L. Bottou and M. Littman, Eds. Montreal: Omnipress, June 2009, pp. 289–296.

[11] P. Verbancsics and K. O. Stanley, "Evolving static representations for task transfer," *J. Mach. Learn. Res.*, vol. 11, pp. 1737–1769, August 2010. [Online]. Available: http://portal.acm.org/citation.cfm?id=1756006.1859909

[12] A. P. B. Lopes, R. S. Oliveira, J. M. de Almeida, and A. de Albuquerque Araújo, "Comparing alternatives for capturing dynamic information in bag of visual features approaches applied to human actions recognition," in *Proceedings of MMSP '09*, 2009.

[13] S. Ebadollahi, L. Xie, S. fu Chang, and J. Smith, "Visual event detection using multi-dimensional concept dynamics," in *ICME '06*, vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, 2006, pp. 881–884.

[14] L. Kennedy, "Revision of LSCOM Event/Activity Annotations, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia," Columbia University, Tech. Rep., December 2006.

[15] L. Kennedy and A. Hauptmann, "LSCOM Lexicon Definitions and Annotations(Version 1.0)," Columbia University, Tech. Rep., March 2006.

[16] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 2004–2011.

[17] D. Lowe, "Object recognition from local scale-invariant features," *ICCV '99*, vol. 2, pp. 1150–1157, 1999.

[18] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC '09*, 2009, pp. 1–5.

[19] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *ICPR '04*, vol. III, 2004, pp. 32–36.

[20] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," *CVPR*, pp. 524–531, 2005.

[21] W. Dai, Q. Yang, G. Xue, and Y. Yu, "Boosting for transfer learning," pp. 193–200, 2007.

[22] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng, "Self-taught learning: transfer learning from unlabeled data," pp. 759–766, 2007.

[23] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[24] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR06*, vol. II, 2006, pp. 2169–2178.

[25] S. E. F. de Avila, A. P. B. a. Lopes, A. da Luz, Jr., and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recogn. Lett.*, vol. 32, pp. 56–68, January 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2010.08.004