Improved Action Recognition by Combining Multiple 2D Views in the Bag-of-Words Model

Gertjan Burghouts, Pieter Eendebak, Henri Bouma, Johan-Martijn ten Hove
TNO
Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands
gertjan.burghouts@tno.nl

Abstract

Action recognition is a hard problem due to the many degrees of freedom of the human body and the movement of its limbs. This is especially hard when only one camera viewpoint is available and when actions involve subtle movements. For instance, when looked from the side, checking one's watch may look very similar to crossing one's arms. In this paper, we investigate how much the recognition can be improved when multiple views are available. The novelty is that we explore various combination schemes within the robust and simple bag-of-words (BoW) framework, from early fusion of features to late fusion of multiple classifiers. In new experiments on the publicly available IXMAS dataset, we learn that action recognition can be improved significantly already by only adding one viewpoint. We demonstrate that the state-of-the-art on this dataset can be improved by 5% - achieving 96.4% accuracy - when multiple views are combined. Cross-view invariance of the BoW pipeline can be improved by 32% with intermediate-level fusion.

1. Introduction

Recognizing human actions is a critical aspect of many types of surveillance, ranging from the security of a business park, to monitoring of patients in a hospital, to surveillance in a public space such as a railway station. Human actions can be indicative of a wide range of unwanted situations such as aggression, vandalism, theft, somebody falling, and becoming unwell. Recognizing actions is an active field of research [1-4] and promising results have been shown on explicit actions such as aggression detection [5], simple kinematic actions such as walk, bend and jump [6,7] and sports [8]. More complex actions that involve subtle motions, for instance give and put down, are still not yet well recognized [9].

An obvious way to improve the action recognition accuracy is to increase the number of viewpoints on the action, such that more details about the action are visible. Two categories of multi-view recognition systems can be distinguished [10]: combination of multiple 2D views, e.g. [11-14], and, 3D, e.g. [15,16], where there is a common reference frame between the viewpoints. A hybrid approach is presented in [17] where the learning is in 3D and during recognition time the 2D view is sufficient. The use of multiple 2D views without full 3D modeling is to be preferred if the person of interest may be partly occluded, or when the camera setup is not known beforehand. For a different camera setup, new calibration and learning is required. In surveillance applications, it is commonly not possible to fully control the camera setup. Therefore, we consider multi-view action recognition by combining 2D views.

Various methods have combined 2D views at a single level. An early fusion strategy for combination of viewpoints is to combine the low-level features [12]. An intermediate level fusion strategy is to concatenate the representations from different viewpoints [11]. Another strategy is to train on all instances obtained from various viewpoints, and test on a new viewpoint [13]. Late fusion can be performed by combining single-view classifiers. In [14] the silhouettes from each view are classified first by a random forest, and subsequently fusion is performed on the classifier level. A systematic evaluation of the benefits of early to late fusion has not been performed. This paper compares the action recognition accuracy when multiple 2D views are combined at the feature, representation or classifier level.

The bag-of-words (BoW) model, based on low-level motion features, is a simple model that has achieved good performance on a range of action recognition tasks [4-7,9,11]. Due to its simplicity, performance and generality, we consider this model as our basic pipeline, where we use STIP features [18] because they proved to be robust features for describing actions [4,5,9]. In the BoW framework, fusion of multiple 2D views can be performed on the feature level, histogram level and classifier level. We evaluate the performance of these fusion schemes and compare them against the state-of-the-art on the multi-view IXMAS dataset of 12 human actions recorded by 5 cameras (4 side and 1 top view) [17].

Recently, view-invariance was tested on the IXMAS dataset [26-28]. The proposed methods are based on correlation subspaces [27], latent kernelized structural SVM [28] and temporal self-similarities [26]. These methods give better cross-camera results, but their

http://dx.doi.org/10.1109/AVSS.2013.6636648

performance on the 4 side views is less than ours. This is noteworthy, because in surveillance the camera viewpoint is often sideways.

Previous experiments on this dataset have shown the merit of adding additional cameras [10,11,14]. In these works, there was no particular rationale for selecting the subsets of cameras. For surveillance systems, a key design issue is to properly select the number of cameras and place them such that the distinctive details about humans and their actions are visible.

In our experiments, we assess systematically the merit of each camera by analyzing the performance of pairs of cameras under 45º, 90º and 135º. Our experiments lead to insights on appropriate camera setups for human action recognition. Furthermore, we do cross-view analysis to assess the camera invariance of the fusion methods.

The paper is organized as follows. In Section 2 we propose the seven fusion schemes to combine the multiple 2D views. Section 3 is about the same-view experiments, where we establish the performance of the combiners, compare to state-of-the-art, and assess the merit of multiple cameras and their viewing angles. Section 4 shows the cross-view experiments, and finally, Section 5 concludes with the main findings.

2. Multi-View Combination Schemes

In Section 2.1 we discuss the bag-of-words model for action detection and in Section 2.2 we propose its extensions to combine multiple 2D views.

2.1. Bag-of-Words Action Detectors

We used a bag-of-word (BoW) [19] pipeline consisting of STIP features to capture motion, a random forest to quantize the features into histograms, and an SVM classifier that serves as action detector. The STIP features are extracted with Laptev's code (v.1.0) and the default parameters [18]. The descriptor of 162 values is obtained by concatenation of all HOG and HOF features. STIPs appeared to be superior to local or bounding-box features [20,21]. The next step is to construct a representation of a video segment with a random forest, which is obtained with Breiman's code [22] with the randomness-parameter M set to all 162 features. Each forest contains 10 trees with 32 leafs, resulting in a 320-bin histogram that is normalized to unity. The final step is classification with an SVM. We use the libSVM code [23]. For each action, a separate SVM is obtained and a test sample is assigned to the action of which the SVM's output (i.e. the posterior probability) is maximal [24]. The SVM is trained using a χ^2 kernel (C=1), and the weight of the positive class (i.e. samples of a particular action) is set to (#pos+#neg)/#pos and the weight of the negative class (i.e. samples without this action) to (#pos+#neg)/#neg, where #pos and #neg are the amount of positive and negative samples [19].

2.2. Multiple 2D Views

In the BoW framework, the multiple 2D views can be combined in several ways. The STIP features can be collected from all cameras, and transformed into a single histogram, i.e. early fusion (feature level).

The features at each camera can be transformed into a view-specific BoW histogram, after which there are two alternatives for combination. The first histogram combiner is to collect the set of histograms and thereby obtaining more training samples. The second histogram combiner is to concatenate the histograms and thereby obtaining longer and richer training samples. We call both alternatives intermediate level fusion.

Late fusion can be performed on the classifier level, after each action-specific SVM has produced its posterior probability. Recall from Section 2.1 that we have one SVM per view v_i and per action a_j. This gives $I \cdot J$ SVMs in total for I views and J actions, each produces a posterior probability. We distinguish between four types of late fusion. The first type assigns to the current sample the action that maximizes the average posterior over the I SVMs. The second type is similar, but assigns to the action that has the maximum posterior from all $I \cdot J$ SVMs. The third type [25] is to learn on the train set the optimal view per action. During testing, the optimal views for all actions are known. For each action, one posterior is obtained by selecting its optimal view v_i. The sample is assigned to the action that has the maximum posterior probability. The fourth type is an adaptation from [21], where a second-stage SVM is introduced that takes all posteriors from all $I \cdot J$ SVMs as input values and trains a mapping from those values to action a_j. Its input values are all between 0 and 1, and they are no histograms. For these inputs, the Euclidean distance proved to work best [21]. Given the outputs of the second-stage SVMs for each action, the sample is assigned to the action that has the maximum posterior probability.

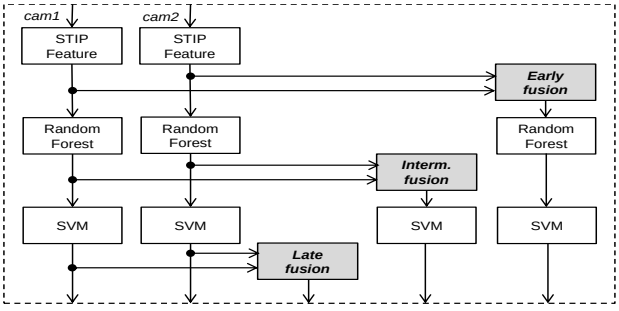

Figure 1: Our BoW framework (Section 2.1) and the proposed extensions for early, intermediate and late fusion, at resp. the feature (STIPs), representation (RF-histograms) and classifier (SVM) levels (Section 2.2).

All early, intermediate and late fusion processing pipelines are indicated in Figure 1. Their naming – to which we will refer in the experiments (Section 3) – and their characteristics are summarized in Table 1.

Fusion	Combiner	Characteristic
Early	STIPs	more features in histogram
Intermediate	set of RF histograms	more histograms
Intermediate	concatenation of RF histograms	longer histograms
Late	average vote	across all views
Late	max. vote	across all views
Late	best view	learned best view per action
Late	2nd stage SVM	classifier on posteriors across views

Table 1: The proposed fusion methods to combine the 2D views (Section 2.2). We distinguish early (STIP features), intermediate (RF histograms) and late fusion (posterior probabilities).

3. Action Recognition Experiments

The experiments are performed on the IXMAS dataset [17] and we evaluate the performance of the combination schemes from Section 3 and – for the best combiner – we establish the added value of adding multiple camera viewpoints. The performance is compared to state-of-the-art.

3.1. IXMAS dataset

The IXMAS dataset [17] consists of 12 complete action classes with each action executed three times by 12 subjects and recorded by five cameras with the frame size of 390 × 291 pixels. These actions are: check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point and pick up. The body position and orientation are freely decided by different subjects.

In the experiments, we refer to the following camera viewpoints, numbered from 1-4, see Figure 2. We do not include camera 5 because its view is from above and for security camera networks this is not common.

Figure 2: The viewpoints from cameras 1-4 of the IXMAS dataset.

3.2. Measure of Performance

The standard setup on the IXMAS dataset is leave-one-actor-out cross validation, which we also used. For each experiment, a new random forest is generated. The classification accuracy and the confusion matrix are the performance measures. The best performance to date on this dataset is 98.8%, achieved by the Stieffel manifold kernel, using 3D information [16]. The best performing method that is based on four 2D side-views is latent kernelized SVM [28], achieving 91.7%.

3.3. Performance of the 2D View Combiners

We evaluate all fusion methods from Section 2.2against a single-view setup (Table 2). The best performing multi-view combination schemes are the posterior average voting and RF histogram concatenation. Surprisingly, these methods perform significantly better than the combination of all STIP features from all viewpoints. Apparently it is advantageous to represent STIP features from each view by a view-specific optimized bag-of-words model first, rather than making the model after collecting all the STIP features. We hypothesize that each view has a distinct influence on the way that the human and action are perceived and that it is easier to generalize feature representations at the view level than across the views. For the same reason, learning a single action recognizer at once from the representations from all views together does not work well. In summary, the methods that combine the multiple views after the representation work best.

Fusion	Combiner	Acc.
None	n/a (single view)	87.9
Early	STIPs	88.6
Interm.	set of RF histograms	81.4
Interm.	concatenation of RF histograms	95.3
Late	average vote	96.4
Late	max. vote	92.2
Late	best view	90.0
Late	2nd stage SVM	94.4

Table 2: Performance of the 2D view combiners compared to average result over the single-view setup per camera.

To understand why histogram concatenation achieves a significant improvement over the single-view setup, we analyze the confusion matrices. Confusions by the single-view setup are displayed in Figure 3 (left) so they can be compared to the confusions by the multi-view histogram concatenation (right). We mention confusions larger than 5% of the test samples. The hard cases in the single-view setup are: check watch (confused with cross arms), cross arms (confused with check watch and scratch head), scratch head (confused with cross arms), punch (confused with kick and wave).

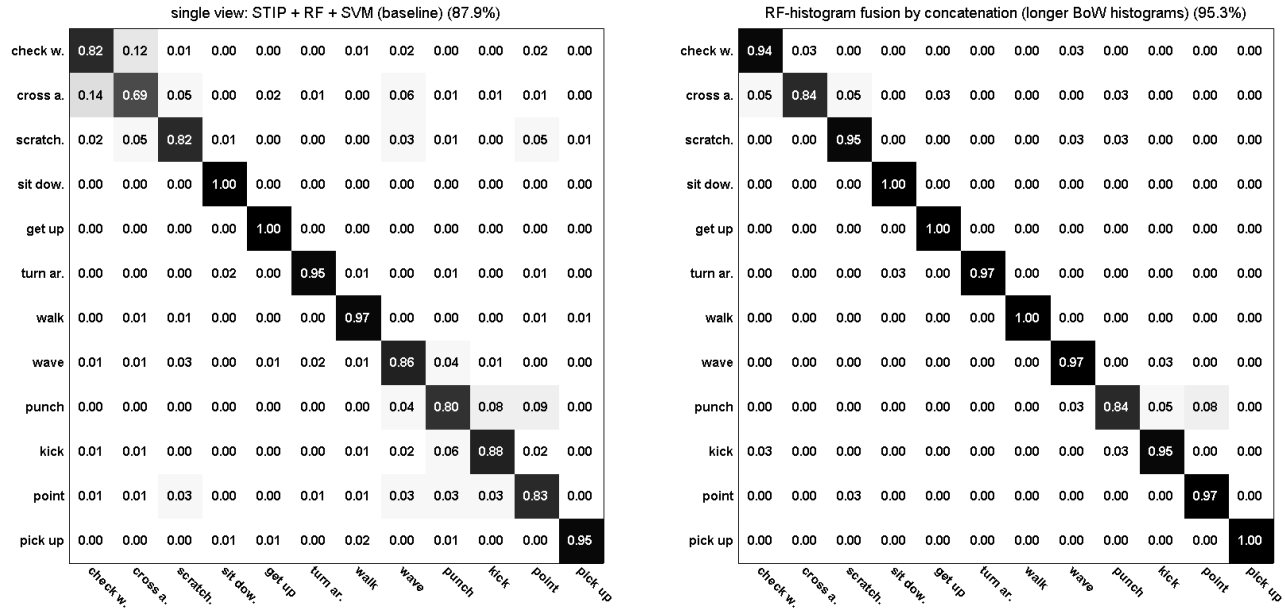

Figure 3: Confusions by our RF histogram concatenation method between the 12 human actions from IXMAS.

The confusions make sense: the actions are visually very similar, especially when the action is partly occluded (e.g. cross arms viewed from the back and from the side will show only one arm which may look like check watch). In the single-view case, there are 7 confusions larger than 5% of which 5 confusions are larger than 10%. For the multi-view case, there are only 4 significant confusions, which are a subset of the single-view confusions: check watch (cross arms), cross arms (scratch head and check watch) and kick (point). None of the actions have a confusion that is larger than 10%. We conclude that the multiple views on an action disambiguate it from visually similar but different actions.

3.4. Comparison to State-of-the-Art

Our best fusion methods are the RF histogram concatenation and posterior average voting. They outperform significantly the state-of-the-art method AFMKL [11] and latent kernelized SVM [28] on the four 2D side views of IXMAS, by a relative improvement of 8.2% and 4.7% resp. (see Table 3). Our method uses multiple 2D views and interestingly our performance approaches the best 3D method on this dataset, the Manifold Kernel [16] and the Latent kernel on 5 views [28], which are still resp. 2.4% and 0.7% better.

Method	Acc.
AFMKL, Wu et al. 2011 [11]	88.2
Latent kernel SVM, Wu e.a.. 2012 [28]	91.7
Our intermediate RF concatenation	95.3
Our late posterior average-vote	96.4

Table 3: Comparison of our best 2D view combiner to the best methods on IXMAS that use multiple 2D views or full 3D.

3.5. Best Combiners: From One to Four Views

We are interested in how the performance of our best fusion method improves when cameras are added. Table 4 shows the accuracies for incrementally adding cameras. Note that the average accuracy were reported over all combinations of 4 separate cameras, 6 camera pairs and 4 camera triplets. Adding the second camera improves the action recognition for RF histogram concatenation accuracy most, a relative improvement of 6.3%, where the third and fourth camera add on average only 1.7%. Note that the standard deviation over the different pairs and triples is much smaller for histogram concatenation than for average voting.

#Cams	Cameras	Acc: interm. RF concat.	Acc: late av. vote
1	4 x separate	87.9	87.9
2	6 x pairs	92.5 ± 0.5	91.5 ± 2.4
3	4 x triplets	94.4 ± 0.3	91.3 ± 3.0
4	1 x all	95.3	96.4

Table 4: Average accuracies of action recognition. The results show an improvement when multiple 2D views are used. RF histogram concatenation and late average voting were used for fusion.

To understand which camera setup is to be preferred if only two cameras are available, we have tested with pairs of cameras that are 45º, 90º and 135º apart. Table 5 summarizes the findings. The performance of all setups is similar, yet the combinations of cameras at 90º are slightly better.

Orient.	Cameras	Acc. interm. RF concat.	Acc. late av. vote
45º pair	[1,2], [2,3], [3,4]	92.4 ± 0.2	90.4 ± 3.1
90º pair	[1,3], [2,4]	92.8 ± 1.0	93.0 ± 0.8
135º pair	[1,4]	92.1	91.7

Table 5: Comparison of setups with 2 cameras under various relative orientations. RF histogram concatenation and late average voting were used.

4. Camera invariance experiments

The results from the previous sections have been obtained by training and evaluating on the same camera. In applications it is not always feasible to train on a given camera. For example in a large camera surveillance network we might want to train on a couple of representative cameras, and then use the trained classifiers on other cameras. For fair comparison with existing state-of-the-art, we show the camera invariance of our current pipeline.

4.1. Single-view camera invariance

We have tested single view camera invariance by training on a single camera, and evaluating on another camera. The cross-view analysis in Table 6 shows that the single-view method is not camera invariant, since the average performance decreases from 87.9 on the same camera to 50.9 on the other cameras (off-diagonal mean in Table 6). Note that the 2 cameras which are similar (e.g. camera 1 and 2) have a good score, and the cameras which are less similar have less good scores.

	Eval cam1	Eval cam2	Eval cam3	Eval cam4
Train cam1	87.0	71.8	46.4	44.0
Train cam2	67.4	87.8	41.1	53.6
Train cam3	46.6	36.2	90.3	56.9
Train cam4	47.4	50.2	48.8	86.6

Table 6: Single-view cross-camera accuracies.

4.2. Camera invariance for multi-view combiners

We evaluate all fusion methods for camera invariance, where the method was trained on two cameras and evaluated on two other cameras. The concatenation of RF histograms fusion method is most camera invariant, with a relative improvement of 32% compared to the single-view pipeline (Table 7). However, the recently proposed Latent kernelized SVM [28] appears to be superior in the cross-camera experiment (Table 8).

Fusion	Combiner	Acc.
None	n/a	50.9
Early	STIPs	66.3
Interm.	set of RF histograms	60.5
Interm.	concatenation of RF histograms	67.3
Late	average vote	64.9
Late	max. vote	61.8

Table 7: Multi-view cross-camera accuracies, with training on two cameras and evaluating on two other cameras (6 pairs).

Method	Acc.
Self-similarities, Junejo e.a. [26]	71.6
Correlated subspace, Huang e.a. 2012 [27]	61.5
Latent kernel SVM, Wu e.a. 2012 [28]	83.3
Our interm. concatenation of RF hist.	67.3

Table 8: Cross-camera accuracies for the four side-view cameras.

4.3. Best combiners: multi-view camera invariance

The number of cameras used for training and evaluation can be varied. We extend the early STIP fusion and average voting to multiple cameras. Table 9 and 10 show that multiple cameras for training and evaluation result in a much better camera invariance than a single-view setup. The case with two training and two evaluation cameras appears to give the best performance.

	1 other eval cam	2 other eval cams	3 other eval cams
1 train cam	50.9	55.9	57.8
2 train cams	57.5	66.3	n/a
3 train cams	57.7	n/a	n/a

Table 9: Average accuracies using early-STIP fusion for different number of training and evaluation cameras. We used all permutations of cameras to compute the average accuracy

	1 other eval cam	2 other eval cams	3 other eval cams
1 train cam	50.9	57.8	60.7
2 train cams	57.5	64.9	n/a
3 train cams	61.3	n/a	n/a

Table 10: Average accuracies using posterior average voting for different number of training and evaluation cameras.

Both tables show that the camera invariance improves if we add more cameras. Adding more training cameras prevents overtraining and allows for better generalization. Adding more evaluation cameras gives more information, and therefore yields better results. Our best mixed camera method (the RF histogram concatenation) requires training and evaluation with the same number of cameras. Therefore Table 9 for this method would only make sense for the 2x2 combination as for the 3x3 combination not enough cameras are available.

5. Conclusions

We have investigated how much action recognition can be improved when multiple views are available. Within the robust and simple bag-of-words (BoW) framework, we have considered various schemes to combine camera viewpoints. Early fusion involves the combination of the STIP features. Intermediate-level fusion combines the views after BoW histograms. Late fusion involves the combination of multiple classifiers, one for each viewpoint. In new experiments on the IXMAS dataset, we have learned that action recognition can be improved significantly already by only adding one viewpoint. Furthermore, we have demonstrated that the state-of-the-art on the four side-views of this dataset can be improved by 5% – achieving 96.4% accuracy – when multiple views are combined by intermediate-level fusion of the BoW representations.

The cross-view scores are significantly lower than scores on the same view. However, fusion methods such as BoW concatenation allows a relative improvement of 32%. For many real-world applications the camera invariance of the method is an important topic to further study. Future work may include the combination of temporal self-similarities with our multi-view fusion.

Acknowledgement

The work for this paper was supported by the Netherlands top sector 'High Tech Systems and Materials', roadmap Security, in the project 'Passive sensors'.

References

[1] L. Gorelick, M. Blank, E. Shechtmanm, M. Irani, R. Basri. Actions as space-time shapes, PAMI, 2007.
[2] A. Gupta, P. Srinivasan, J. Shi, L. S. Davis. Understanding videos, constructing plots: learning a visually grounded storyline model from annotated videos, CVPR, 2009.
[3] N. Ikizler-Cinbis, S. Sclaroff. Object, scene and actions: combining multiple features for human action recognition, ECCV, 2010.
[4] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld. Learning realistic human actions from movies, CVPR, 2008.
[5] I. Lefter, L.J.M. Rothkrantz, G.J. Burghouts. A comparative study on automatic audio-visual fusion for aggression detection using meta-information, PRL, 2013.
[6] T. Guha, R.K. Ward. Learning sparse representations for human action recognition, PAMI, 2012.
[7] C. Schuldt, I. Laptev, B. Caputo. Recognizing human actions: a local SVM approach, ICPR, 2004.
[8] S. Sadanand, J. J. Corso. Action bank: a high-level representation of activity in video, CVPR, 2012.
[9] G.J. Burghouts, K. Schutte. Spatio-temporal layout of human actions for improved bag-of-words action detection, PRL, 2013.
[10] M.B. Holte, T.B. Moeslund, C. Tran, M.M. Trivedi. Human activity recognition using multiple views: a comparative perspective on recent developments, HGBU, 2011.
[11] X. Wu, D. Xu, L. Duan, J. Luo. Action recognition using context and appearance distribution features, CVPR, 2011.
[12] Y. Song, L-P. Morency, R. Davis. Multi-view latent variable discriminative models for action recognition, CVPR, 2012.
[13] J. Liu, M. Shah. Learning human actions via information maximization, CVPR, 2008.
[14] F. Zhu, L. Shao, M. Lin. Multi-view action recognition using local similarity random forests and sensor fusion, PRL, 2013.
[15] P. Natarajan, R. Nevatia. View and scale invariant action recognition using multiview shape-flow models, CVPR, 2008.
[16] P. Turaga, A. Veeraraghavan, R. Chellappa. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision, CVPR, 2008.
[17] D. Weinland, E. Boyer, R. Ronfard. Action recognition from Arbitrary Views using 3D Exemplars, ICCV, 2007.
[18] I. Laptev. On space-time interest points, IJCV, 64 (2/3), 2005.
[19] G.J. Burghouts, K. Schutte, R. den Hollander, H. Bouma. Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos, Machine Vision and Applications, 2013.
[20] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, C. Schmid. Evaluation of local spatio-temporal features for action recognition, BMCV, 2009.
[21] H. Bouma, G.J. Burghouts, L. de Penning, et al. Recognition and localization of relevant human behavior in video, SPIE 8711, 2013.
[22] L. Breiman. Random forests, Machine Learning, 2001.
[23] R. Fan, P. Chen, C. Lin. Working set selection using second order information for training SVM. Journal of Machine Learning Research 6, 1889-1918, 2005.
[24] G.J. Burghouts, K. Schutte. Correlations between 48 human actions improve their performance, ICPR, 2012.
[25] A.K. Jain, P.W. Duin, J. Mao. Statistical pattern recognition: a review, PAMI, 2001.
[26] I. Junejo, E. Dexter, I. Laptev, P. Perez. View-independent action recognition from temporal self-similarities, PAMI, 2011.
[27] C. Huang, Y. Yeh, Y. Wang. Recognizing actions across cameras by exploring the correlated subspace, ECCV, 2012.
[28] X. Wu, Y. Jia. View-invariant action recognition using latent kernelized structural SVM, ECCV, 2012.